Subcellular Dynamics of RNAs and microRNAs in Mouse Embryonic and Trophoblast Stem Cells

by

Brian Park

A thesis submitted in conformity with the requirements
for the degree of Master of Science

Department of Physiology
University of Toronto

Subcellular Dynamics of RNAs and microRNAs in Mouse Embryonic and Trophoblast Stem Cells

Brian Park

Master of Science

Department of Physiology
University of Toronto

2020

Abstract

Cell fractionation coupled to high-throughput RNA-sequencing allows for the identification of cytoplasmic and nuclear over-represented RNA populations. In mouse embryonic and trophoblast stem cells, an asymmetric distribution of protein coding RNAs were observed with respect to gene function. Cell lineage specific differences were found between population of nuclear over-represented messenger RNAs, such that embryonic stem cells showed an abundance of transcripts related to cell division, cell cycling, and DNA repair, whereas trophoblast stem cells showed an abundance of transcripts related to cell-cell adhesion, cell junction formation, and cell migration. Lineage specific processes were also found over-represented between nuclear enriched transcripts containing exon-intron junctions, suggesting intron retention may play a role in maintenance of cell identity in development. Small RNA-sequencing data showed microRNAs are related to embryonic and trophoblast lineages with association with signaling pathways. Presence of mature microRNAs in the nucleus was also identified.

Table of Contents

List of Tables

List of Figures

Chapter 1

1 Introduction

1.1 RNAs are key mediators of cellular function and biological identity

1.1.1 Protein-coding RNAs serve as a blueprint for cell function

The transcriptome is defined as the complete collection of protein-coding messenger RNAs (mRNAs) and hence represent the product of the genome[1]. The transcription of the eukaryotic genome by RNA polymerase II (Pol II) is a conserved mechanism at the core of gene expression. Understanding this process is essential in understanding the functional role of the genome on the given system's development, function, and identity. Transcription functions to convert the genetic code of the organism into readable formulae for production of proteins, and thus links the genome and the proteome. Therefore, the population of poly-adenylated mRNAs in a biological system serves as a blueprint for cellular function and development as well as a proxy for interpreting the functional elements of the genome.

The role of RNAs, however, extends beyond their ability to transcribe genetic information into functional proteins. This is in part due to the fact that gene expression is a highly regulated process which modulates functional protein output in accordance with factors such as developmental timing, external chemical stimuli, or exposure to stress. The study of the transcriptome, coined *transcriptomics*, then must aim to profile not just the protein-coding mRNAs, but also RNAs that do not code for proteins at all. In fact, the coding exons of protein-coding RNAs have been shown to represent only 1.5% of the human genome[2], whereas up to 80% of the genome has been documented to be able to transcribe non-coding RNAs (ncRNAs)[3,4]. NcRNAs are heterogeneous in both function and physical size; efforts to profile ncRNAs in literature has shown their involvement in regulation of developmental processes and diseased states, as well as in cancer[5,6]. NcRNAs can also serve as *trans*-regulatory factors to regulate the behavior and fate of the mRNAs, thereby regulating gene expression[7].

Furthermore, literature in transcriptomics have catalogued the variability in both processing and subcellular localization of mRNAs as well – thus the progression of the newly synthesized, pre-mRNA from the chromatin, to the nucleoplasm for post-transcriptional processing, and finally to the cytoplasm for translation is a highly regulated process[8,9]. Consequently, Pol II transcripts are pervasive across distinct compartments within a given cell in varying degrees of maturation and processing, with heterogenous function. Due to this complexity in the life cycle and fate of RNAs, the quantification of the transcriptome as well as subcellular mapping of the RNA population is necessary to understand the cellular blueprint.

1.1.2 Understanding the blueprint gives insight into cell identity

Understanding the transcriptome *dynamics* – henceforth defined as the change in RNA expression levels as well as change in subcellular localization - of organs, tissues, and its constituent cells allows for valuable insight into the pertinent processes underlying gene expression and gene regulation. By identifying up-regulated and down-regulated population of transcripts and functional gene sets, one can infer individual gene function and role in the resulting phenotype. Furthermore, by comparing the transcriptome between conditions such as developmental timing or diseased states, it allows for inference on how changes in the external environment or senescence affects functional gene expression. This importance of RNA role as the complex intermediate between the genome and the proteome has led to demand and advancements in high-capacity RNA assays in order to identify and quantify gene expression[10]. In next-generation RNA-sequencing (RNA-seq), the ability to directly sequence the transcriptome at a single-nucleotide resolution level allows for a wide variety of applications[11-13]. RNA-seq allows for massively parallel analysis of the transcriptome across multiple experimental conditions, which allows user to qualitatively and quantitatively compare transcriptomic signatures of different physiological, chronological, or genotypic conditions.

1.2 Complex regulatory processes underly fate of RNA

1.2.1 Subcellular localization of RNAs is a conserved process

Subcellular trafficking of RNAs as a means of gene regulation is a conserved process in eukaryotic cells[14]. This localization dynamics of RNAs have been shown to be regulated by *cis*-acting

elements within the transcript sequences, which act as binding motifs for *trans*-acting factors. A well-documented example of such interaction is the case of RNA-binding proteins (RBPs) interacting with mRNAs in neuronal cells in development of postsynaptic dendritic spines[15]. Dendritic mRNA localization via RBPs is mediated by the interaction of the *trans*-acting factors with 3' untranslated regions (UTRs) of localized mRNAs; after assembly, the complex is transported along the cytoskeleton into dendrites, where synaptic proteins are subsequently translated upon transduction signal[16]. The interaction of RBPs and *cis*-regulatory elements of mRNAs have been documented to exhibit functional roles at multiple cellular levels, including the regulation of RNA splicing, nuclear export, cytoplasmic localization, and mRNA stability[17,18]. Furthermore, the binding motif on the target mRNA has been shown to be frequently located within 3'UTRs, while often bearing repeated sequences[19].

1.2.2 Non-coding RNAs are key regulators of gene expression

Regulatory RNAs can also act as *trans*-acting factors to regulate RNA behavior. Such ncRNAs can be divided into subclasses of small ncRNAs such as micro-RNAs (miRNAs), transfer RNAs (tRNAs), small nucleolar RNAs (snRNAs), and short-interfering RNAs (siRNAs), as well as longer RNAs such as ribosomal RNAs (rRNAs) and long non-coding RNAs (lncRNAs). LncRNAs in particular are a subclass of regulatory RNAs that play a role in gene regulation via mechanisms such as chromatin remodeling and miRNA sequestration[20,21]. Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) is an example of a highly conserved RNA whose dysregulation leads to an increase in invasion and metastasis of multiple cancer cells[22]. Knockdown studies of MALAT1 has shown, in return, a promotion of miRNA-140 expression and subsequent suppression of cancer cell migration and invasion[23]. This suggests RNA-RNA interactions have profound effects on cellular fate.

The miRNA-sponge role of lncRNAs is of particular interest, as miRNA themselves can act as regulators of gene expression at the post-transcriptional level[24,25]. As a subclass of small regulatory RNAs, miRNA biogenesis and maturation pathway follows one of two pathways: the canonical miRNA pathway and the mirtron pathway, in which as the name suggests, precursor miRNAs are formed from excised introns of a given transcript[26]. In the canonical biogenesis pathway, the primary miRNA transcripts are processed in the nucleus by the RNase Drosha[27]. Processed

transcripts are then exported into the cytoplasm to be cleaved by the endoribonuclease complex Dicer to form a mature duplex[26,27]. This duplex is then unwound by helicases to form two miRNA strands, one of which subsequently associates with the RNA-induced silencing complex (RISC)[28]. This functional single-stranded miRNA is of 18 to 22 nucleotides in length and primarily mediate post-transcriptional gene silencing by base-pairing of its 5′ seed sequence to the 3′ UTR of target mRNA – an interaction associated with the recruitment of RISC[29]. This mode of gene silencing has shown to exist in the form of translation repression or target RNA destabilization and degradation, with varying degrees of Watson-Crick base complementarity between the miRNA and the target RNAs[30].

An example of canonical miRNA-mRNA interaction is miR-140 in cancer cells. This miRNA has shown to directly target genes such as SOX9 and ALDH1, which are activated stem-cell factors in ductal carcinoma[31]. The interaction between miR-140 with SOX9 and ALDH1 transcripts was indeed validated as the base-pairing of miR-140 with the 3' UTRs of target mRNA[32]. As such, studies in miRNA profiling in breast cancer have shown that miR-140 has a significant role in regulating stem cell signaling in ductal carcinoma and expression of this miRNA is downregulated in cancer stem-like cells compared to normal stem cells. This base-pair-interaction type model of miRNA-mRNA regulation allows for profiling molecular mechanisms of disease and design of novel nucleotide therapeutics.

However, miRNA-mRNA interaction have also shown non-canonical behavior; numerous studies support miRNA binding ability to not only the 3' UTR of target RNA, but also the 5' UTRs and within the open reading frame, with varying degrees of base complementarity[33]. Furthermore, the post-transcriptional role of miRNA has shown not only suppressive but also promotive modulatory effects[34].

1.2.3 miRNAs can be localized to the nucleus

As with the subcellular trafficking of long RNAs across the cytoplasmic-nuclear boundary, the export of the precursor miRNA to the cytoplasm is not only an essential step in maturation, but a tightly controlled mechanism in gene regulation. As such, the expression of miRNA themselves is under post-transcriptional regulation at various stages of maturation[35]. In a study of a brain-specific

miRNA miR-138, Leuschner *et al* found that the maturation of miR-138 transcript is stalled in the cytoplasm prior to Dicer processing; this regulation of miRNA processing cascade and accumulation of pre-miRNAs showed differential behavior in cell lines, as only cells of the hippocampus, the neo-cortex, the cerebellum, and the fetal liver complete the processing of pre-miRNA into mature, single-stranded miRNA[36]. This suggests that the miRNA maturation process harbors physiological check-points that regulate levels of active miRNA expression. Indeed, primary miRNA transcripts of the Let-7 family were also shown to be halted in processing, but this time at the level of Drosha in the nucleus[37].

On the contrary, there is evidence in literature that mature, fully-processed miRNAs can be re-imported into the nucleus. In a study of the miR-29 family, Hwang *et al* found that miR-29b can be re-directed to the nucleus, whereas miR-29a cannot, with the distinctive and decisive feature being the presence of a hexanucleotide sequence element AGUGUU[38]. An insertion of this element on a modified siRNA led to the enrichment of this species in the nucleus. Further, Importin-8 (IPO-8) knockout cells showed a decrease in the nuclear enrichment of known nuclear miRNAs, without altering the total cellular levels[39]. Another study showed molecular interaction between IPO8 and Argonaute 2 protein (AGO2) – an active component of the RISC – such that IPO-8 knockout led to a decrease in nuclear enrichment of AGO2[40]. This nuclear enrichment of mature miRNAs and active RISC components suggests that miRNAs may exert regulatory function at the level of transcription. Indeed, nuclear miRNA studies show interactions with promoter sequences in target genes as a means of miRNA-directed transcriptional gene silencing[41].

1.3 Variability in RNA processing leads to protein diversity and gene regulation

1.3.1 Alternative splicing is a pervasive mechanism in eukaryotes

Processing of long RNAs is also not as straightforward as originally thought; advancements in technology such as RNA-seq, as well as the completion of the human genome has revealed that the transmission of genetic information from DNAs to RNAs to proteins is unbalanced in stoichiometry. The encyclopedia of DNA elements (ENCODE) suggests that over 90% of human genes are alternatively spliced, leading to the production of multiple functional proteins from a single transcript[42-44]. Previously thought of as "junk DNA", introns – defined as elements flanking

exons – has been shown to not only harbor classes of ncRNAs, but also play a role in regulation of gene expression.

Alternative splicing events (ASEs) can occur via exon skipping, alternative use of 5' and 3' splice sites, and intron retention (IR)[44]. IR in particular is of particular interest, as processed intron-containing RNAs were previously thought to be results of mis-splicing events and thus lead to pathology. However, recent research in IR has revealed that the fate of intron retaining transcripts is not simply degradation, but in fact can lead to production of protein isoforms or even be actively retained in the nucleus[45].

1.3.2 Intron retention is a major mechanism of gene expression regulation

IR is defined as a phenomenon in which a mature processed mRNA retains its introns. IR transcripts can be found in both the cytoplasm and the nucleus and has been shown to lead to either degradation or stabilization of the transcript.

Boutz *et al* have shown that, in poly-adenylated mRNA in mouse embryonic stem cells, intron retaining transcripts retained in the nucleus may not be immediately subject to nuclear degradation, and instead await for a signal to be processed and translated rapidly when required[46]. In another study in mouse neurons, Mauger *et al* showed that post-transcriptional processing of such *sentinel RNAs* occurs in response to $GABA_A$ receptor activation, such that a rapid increase of spliced mRNAs was observed following $GABA_A$ activation coupled to a pre-treatment with transcription inhibitors[47]. Such observations suggest that IR plays an important role in temporal regulation of gene expression, as well as mRNA export.

Alternatively, retained intronic sequences in the transcript can introduce premature termination codons (PTCs), which leads to subsequent nonsense mediated decay (NMD)[48]. IR-NMD pathway introduces a gene expression control in IR events; in the example of granulocyte formation, Lmnb1 gene – coding for the nuclear lamina – exhibits IR, leading to reduced expression. Wong *et al* showed that with an expression of "intron-less" Lmnb1 – such that IR-NMD was impossible – granulocyte population decreased in number and showed altered nuclear shape and volume[49]. Subsequent consequence was lower detectable amount of granulocytes in peripheral blood of mice

and thus disruption of normal granulopoiesis. Interestingly, it has also been noted that IR is implicated in diseased states; systematic transcriptomic profiling studies revealed numerous cases of IR in diseased tissue, such as in breast cancer, lung carcinomas, as well as in cancers of the bladder, colon, endometrium, kidney, and liver[50-52]. The consequence of new protein isoforms arising due to IR, as well as altered regulation in gene expression in relation to diseased tissues remain uncertain. The effect of IR may not be necessarily conducive to disease, as in the example of calcineurin gene, in which IR leads to a new isoform which improves cardiac function[53,54].

Effects of IR in biological processes has previously been shown to be conserved in mouse and human cells. In an effort to elucidate the levels of conservation of IR in vertebrates, Schmitz *et al* performed a phylogenetic analysis of IR quantification in granulocytes of species spanning 430 million years and found that IR provides a conserved mechanism of post-transcriptional regulation control[55]. Furthermore, a relatively larger number of miRNA binding elements were found on IR harboring genes indicating that IR-mediated and miRNA-mediated gene regulation control may be complementary[50].

1.3.3 Intron retention can be profiled using high-throughput sequencing data

The availability of fully annotated genomes in model organisms (e.g., from UCSC, GENCODE, ENSEMBL) allows for high-throughput transcriptomic profiling with RNA-seq at the level of transcripts and exon-intron boundaries. Quantifying levels of RNA-seq reads that span splice junction boundaries yield insight into processes governing differential splicing mechanisms. In order to quantify levels of IR, however, levels of reads spanning exon-intron boundaries as well as exon-exon boundaries per given gene must be quantified. Braunschweig *et al* suggest the calculation of percent intron retention (PIR) as an estimate of the extent of intron retention per given gene[56]. PIR per gene can be calculated as the ratio of unspliced (i.e., exon-intron) junction reads and the sum of unspliced and spliced (i.e., exon-exon) junction reads (**Figure 1**).

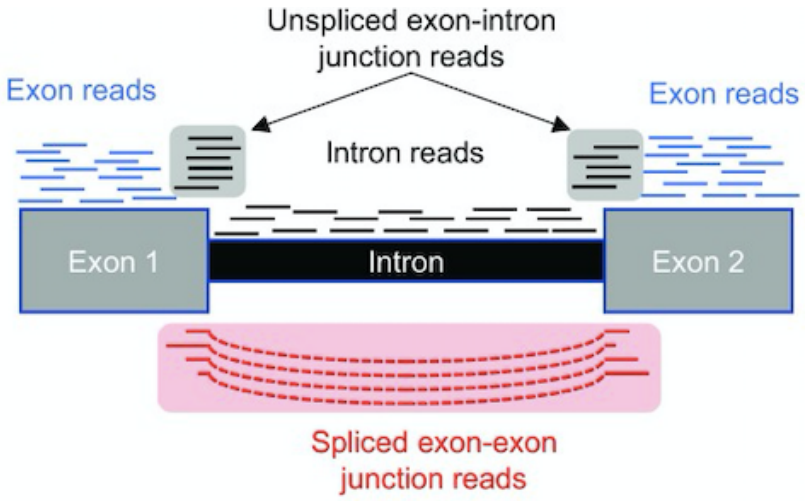


Figure 1: **Illustration of PIR calculation using RNA-seq data (Braunschweig *et al*, 2017)**

Intron retention can be estimated using calculation of percent intron retention (PIR), which can be quantified using a ratio of RNA-seq reads spanning unspliced exon-intron junctions to the sum of unspliced and spliced exon-exon junction reads.

Using PIR, Braunschweig and colleagues found that the transcriptome during cell differentiation is under regulatory control via IR, such that during differentiation into neural tissue in mouse embryonic stem cells, genes related to pluripotency and cell division showed high PIR (i.e., higher extent of IR per gene) whereas genes related to neural development showed low PIR[56]. This observation showed that IR facilitates the down-regulation of genes not required for cell lineage fate commitment via IR triggered NMD. Furthermore, using RNA-seq and PIR the authors were able to show that IR is a pervasive mechanism in mammalian cells, affecting the majority of multiexonic genes and their transcripts. This result showed that a) IR events can be profiled using high-throughput RNA-seq, b) gene expression modulation via IR is a conserved mechanism in various cell types, and c) IR is involved in cell fate specification and differentiation via modulation of gene expression.

The subcellular trafficking of both protein-coding and regulatory RNAs, as well as the regulatory processes found in both the cytoplasm and the nucleus, reiterates the need to characterize not only the overall gene expression but the subcellular localization of pertinent transcripts as well.

1.4 Cell fractionation and RNA-seq gives insight into subcellular transcriptome dynamics

1.4.1 Cell fractionation reveals asymmetrically expressed RNAs

Cell fractionation coupled to high-throughput RNA-seq has been used in literature as a means of characterizing asymmetrically distributed RNAs across cell compartments[57]. Isolation of RNA content in a compartmentalized manner allows for identification of RNAs enriched in subcellular compartments including the cytoplasm and the nucleus. RNA-seq and downstream differential expression analysis then allows for an assessment of the up-regulated RNA population in the cytoplasmic and nuclear fractions and associated biological function. Therefore, using cytoplasmic and nuclear RNAs – rather than whole cell RNA content - for differential expression and identification of IR events provides a spatial context of the transcriptome. This provides insights into the mechanism of regulatory dynamics in both subcellular localization of pertinent transcripts as well as intron retaining behaviour[58,59].

1.4.2 Intron retaining transcripts are present in both the cytoplasm and the nucleus

The subcellular localization of intron retaining transcripts is also of interest due to the presence of such transcripts in both the cytoplasm and the nucleus found in literature. The fate of intron retaining transcripts at the site of translation machinery often leads to degradation via NMD but can also lead to alteration in translation efficiency as well as production of protein isoforms[56] (**Figure 2**). Intron retaining transcripts in the nucleus can lead to both degradation and nuclear retention, as evident in literature outlining the phenomenon of *detained introns*[46]. Therefore, the accumulation of intron retaining transcripts in either subcellular compartment leads to a modulation of overall gene expression levels[60]. Profiling the population of intron retaining transcripts in cytoplasmic and nuclear fractions then yields insight into the cellular processes governed by location-specific regulatory processes.

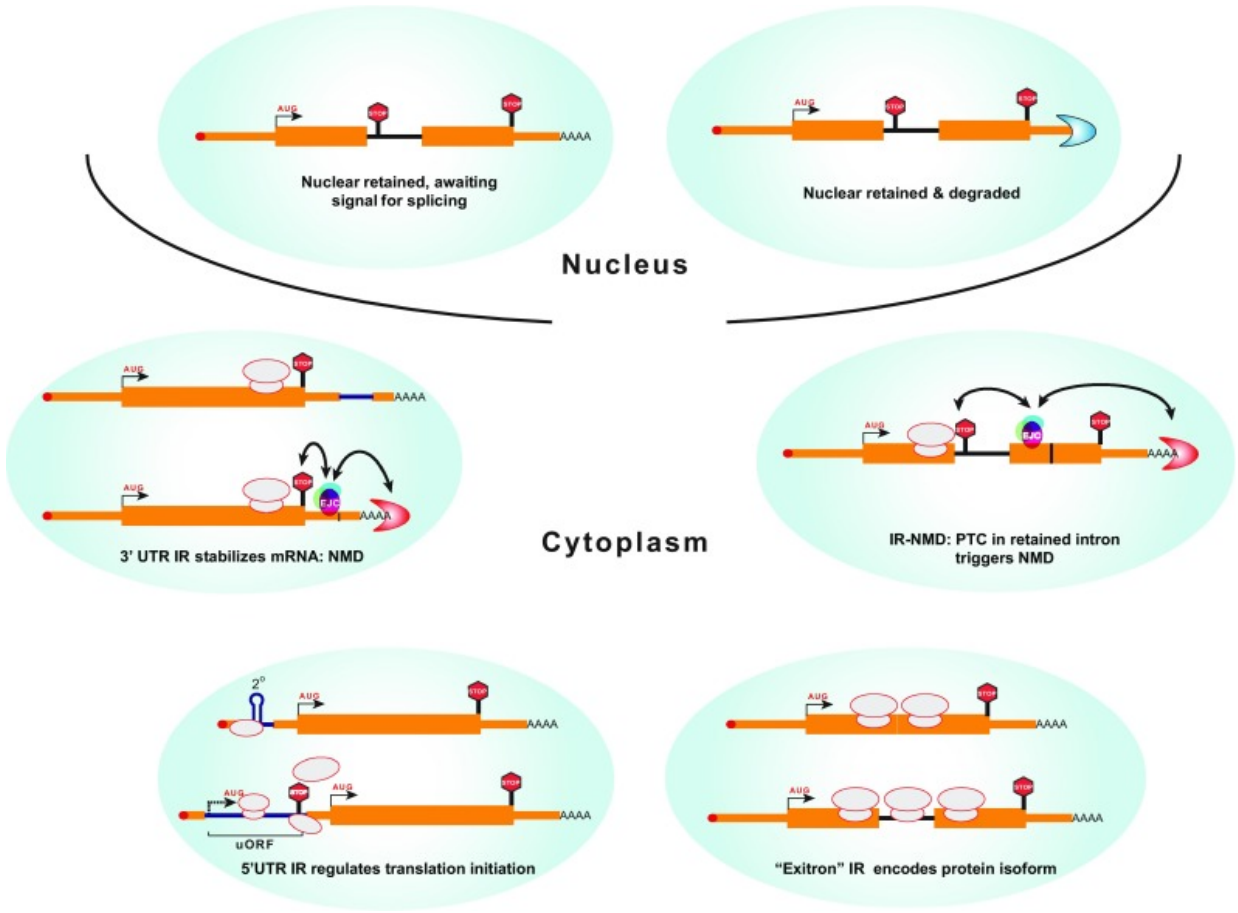


Figure 2: **Fate of intron retaining transcripts in the cytoplasm and the nucleus (Braunschweig *et al*, 2017)**

Intron retaining transcripts can be exported into the cytoplasm or retained in the nucleus. In the cytoplasm, the retained intron may lead to nonsense mediated decay due to an insertion of a premature stop codon. In some cases, the transcript may also lead to production of protein isoforms. In the nucleus, the intron retaining transcript may be degraded by nuclear exosome pathway or actively retained for rapid export and translation upon cellular signal.

The pervasiveness of IR and other alternative splicing events, the conserved subcellular trafficking of mRNAs, and the complex regulatory role of non-coding RNAs all suggest a dynamic, highly regulated transcriptome. Indeed, the controlled expression of pertinent genes as well as gene silencing is necessary in maintenance of cells' survival, function, and identity[61]. Determination of cell fate and differentiation is driven by the both the controlled up-regulation and down-regulation of select genes. The importance of cell fate decisions has been widely documented in literature of developmental biology, wherein pertinent mRNAs and proteins involved in embryogenesis have shown a high level of evolutionary conservation in model multicellular organisms such as *Drosophila melanogaster* and *Mus musculus*[62]. Therefore, it is no surprise that stringent gene

expression regulation and coordinated molecular signaling can be observed in pluripotent stem cells at crossroads in system development[63].

1.5 RNAs mediate early cell lineage fate decisions in mouse

1.5.1 RNA-RNA interactions play a key role in early development

A well-documented example of a tightly regulated cellular process is cell differentiation in early development. The mammalian embryonic development consists of a series of cellular fate decisions that require stringent reciprocal molecular signaling in order to restrict developmental potential. The first cell fate decision occurs with the formation of the blastocyst around 3 days post-fertilization in mice[64]. The totipotent stem cell population then differentiates into two distinct cell lineages: the extraembryonic trophoectoderm (TE) – which develops into the fetal portion of the placenta – and the inner cell mass (ICM) – which forms the embryo proper, as well as the amnion, the yolk sac, and the allantois[65,66]. This highly controlled process marks the first asymmetric restriction of developmental potential and cell fate decision[67].

TE and ICM specification require the expression and interactions between transcription factors that govern each cell lineage. The ICM fate is promoted via the expression of transcription factors Oct4 and Sox2, and subsequently, the expression of Nanog[68-70]. The expression of Cdx2, meanwhile, is induced by the suppression of Oct4 and the overexpression of Cdx2 alone was found to be sufficient in generating TE cells[64]. The presence of Cdx2 expression in TE cells is pertinent in subsequent suppression of ICM factors Oct4 and Nanog and therefore is crucial in maintenance of TE identity[71]. This reciprocal signaling between the TE and the ICM leads to a spatial-specific distribution of Sox2, Oct4, and Nanog proteins in the ICM and Cdx2 in the TE within the pre-implantation blastocyst[71-74] (**Figure 3**).

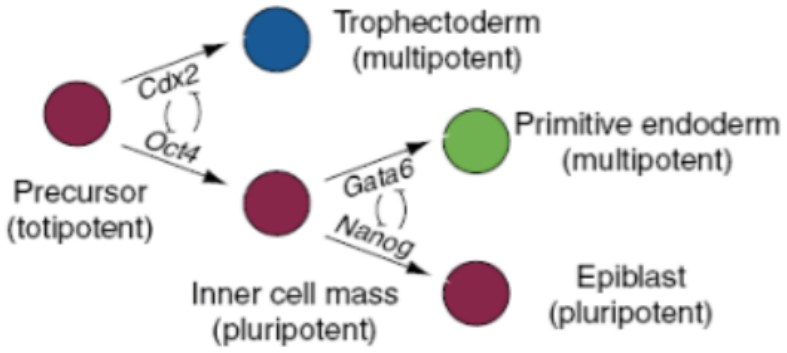


Figure 3: **Reciprocal molecular signaling governing ICM and TE fate**

The totipotent stem cell population in early development differentiates into the inner cell mass (ICM) and the extraembryonic trophoectoderm (TE) to form the early pre-implantation blastocyst. The ICM lineage is characterized by the expression of transcription factors Oct4 and Nanog, which maintains the pluripotent state in the cell population. Cdx2 expression in the TE suppresses the expression of ICM factors and promotes TE fate. This reciprocal signaling of transcription factors is essential in the formation of two distinct cell lineages within the blastocyst.

The expression of such transcription factors within the blastocyst is also associated with the up-regulation of many regulatory miRNAs and the corresponding down-regulation of pluripotency genes. It was shown that a deletion of DGCR8 – an RNA binding protein required in miRNA biogenesis – in differentiating embryonic stem cells led to the inability to down-regulate all pluripotent markers including Oct4[75]. Furthermore, levels of miR-21 have been shown to increase during cell differentiation and target the 3' UTR of mRNAs that code for Sox2 and Nanog, suggesting the interplay of transcription factors and miRNAs in gene regulation during cell differentiation[76,77]. As well, miRNAs such a miR-134, miR-296, and miR-470 have been shown to target regions outside of the 3' UTR in mRNAs of Nanog, Sox2, and Oct4[78]. Finally, the induction of miRNAs miR-15b, miR-322, and miR-467g in mouse embryonic stem cells was sufficient in promoting trophoblast morphology and an up-regulation of trophoblast markers Cdx2 and Gata3[79]. Thus it is apparent that the transcription factor network governing cell fate determination is intertwined with regulatory miRNA processes.

The maintenance of ICM and TE lineages via key transcription factors is also related to cellular processes such as regulation of metabolism and chromatin modifications. For example, binding sites for Sox2, Oct4, and Nanog has been found in enhancer sites of GLUT1, which leads to increased GLUT1 expression and glycolytic flux[80]. This finding suggests maintenance of pluripotency is related to the cells' ability to generate metabolites and meet biosynthetic demands[81].

Furthermore, Cdx2 has been shown to be involved in a positive feedback system between Elf5 and Eomes, which is established by the hypomethylation of the Elf5 promoter in trophoblast stem cells[82]. The feedback system between Cdx2, Elf5, and Eomes has been associated with maintenance of TE lineage[83]. Evidently, the expression of genes required for the maintenance of cell proliferation and self-renewal is related to a complex molecular network that prevents differentiation and promotes cell division via epigenetic and metabolic processes. Gene expression analysis of the ICM and TE lineages then, can provide insight into the pertinent cellular processes required for cell maintenance and regulatory processes that govern them.

1.5.2 Embryonic and trophoblast stem cells represent the ICM and the TE

The stem cell representatives of the ICM and the TE – embryonic stem cells (ESCs) and trophoblast stem cells (TSCs), respectively – can be maintained in cell culture using supplement enriched media (**Figure 4**). The *in vitro* self-renewal of mouse-derived ESCs can be achieved using leukemia inhibitory factor (LIF) supplemented in serum-containing cell medium[84,85]. Multipotent mouse-derived TSCs, on the other hand, can be maintained in cell culture with supplemental fibroblast growth factor 4 (FGF4)[86,87]. FGF4 and its receptor FGFR2 has been identified in literature as signals required for maintenance of multipotent state in undifferentiated TSCs[88].

The ability to maintain and harvest ESCs and TSCs in cell culture allows for isolation of cell lineage specific protein-coding RNAs and regulatory RNAs specific to each cell identity and function. A high-throughput, parallel analysis of the ICM and TE transcriptome then allows for a comparative study of mouse pre-implantation processes in each cell lineage.

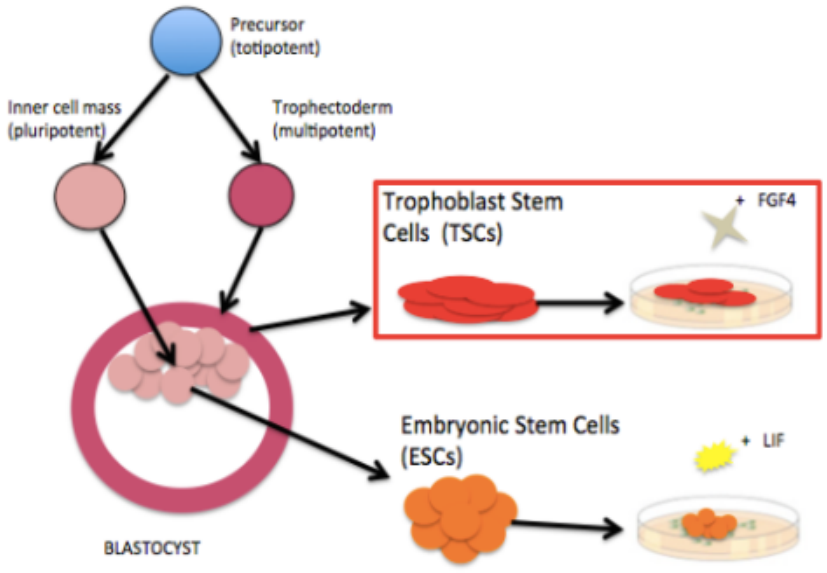


Figure 4: **Embryonic and trophoblast stem cells as representatives of the ICM and the TE**

Stem cell representatives of the ICM and the TE – embryonic and trophoblast stem cells, respectively – can be grown and maintained in cell culture for *in vitro* gene expression assays. Characterizing the RNA population from the cell lines allows for the analysis of genes pertinent for ICM and TE maintenance and self-renewal.

1.6 Objectives and hypotheses

The objective of the current study is to investigate the subcellular localization of RNAs and associated cellular function in ESC and TSCs. By performing cell fractionation coupled to high-throughput RNA-seq, gene expression data is contextualized by subcellular location. Differential expression and gene set enrichment analysis will be employed in order to identify cellular processes regulated by cytoplasmic-nuclear localization. Identification of ICM- and TE-related functional gene sets in the cytoplasmic and nuclear fractions allows inferences on how cells modulate the transcriptome to carry out key functions.

ESCs and TSCs are used as *in vitro* models in order to investigate how splicing regulation and subcellular localization are involved in cell fate commitment. As cell fate commitment is achieved by cell lineage specific transcriptional programs[89], identifying how the transcriptome is expressed and regulated in the representatives of the first cell fate decision in embryogenesis – the ICM and the TE – gives insight into how this decision is established and maintained.

The advantage of cell fractionation and RNA-seq is that alternative transcript processing events such as intron retention can also be estimated in context of subcellular location. The objective of this portion of the analysis is to investigate how ESCs and TSCs employ intron retention in order to maintain self-renewal and which intron retaining gene sets are subject to nuclear retention or cytoplasmic localization. This gives another mechanistic insight into gene expression regulation in ESCs and TSCs.

RNA-seq profiling of small RNAs will also be used to profile miRNAs and their regulatory role in ESCs and TSCs. The objective is to identify potential candidate miRNAs involved in the maintenance of ICM and TE lineage. Differential expression between the two subcellular fractions will also yield insight into whether functional miRNAs can be identified in the nucleus.

Following hypotheses are made:

1) The differential expression profile between two subcellular fractions will reveal an asymmetric distribution of up-regulated mRNAs with respect to function, such that:

    a. Up-regulated mRNAs in the cytoplasmic fraction in both ESCs and TSCs will be related to cellular processes associated with high rate of translation and cytoplasmic turnover

    b. Up-regulated mRNAs in the nuclear fraction in both ESCs and TSCs will be related to cellular processes associated with developmental timing

2) mRNAs with intron-retaining behaviour will show ESC and TSC differences, such that cell-specific processes will be influenced by intron retention.

An overarching theme is that ESCs and TSCs will show both similarities and dissimilarities in the mRNA differential expression profile. As both cells are associated with their ability to self-renew and maintain a level of developmental potential, regulatory processes such as subcellular localization and intron retention may govern common processes such as epigenetic mechanisms,

cell division, and energy metabolism. Evidence supporting this will suggest a level of conservation in gene expression control in the two lineages. However, due to the difference in phenotype and reciprocal signaling to form the TE and the ICM, the differential expression profile will also give insight into how lineage-specific processes in the pre-implantation blastocyst are regulated within the cell.

Chapter 2

2 Cell Fractionation and Classification of Subclasses of RNA-Seq Data

2.1 Methods

2.1.1 Cell culture

Frozen batches of mouse derived wild-type embryonic stem cells (ESCs, a male R1 cell line[90]) and trophoblast stem cells (TSCs, female and derived from E3.5 blastocysts obtained from ICR x ICR mating as previously described[91]) were thawed onto DMEM and RPMI culture media (Thermo Fisher), respectively. Both stocks of cell media were enriched with fetal bovine serum and as well, ESC media were supplemented with leukemia inhibitory factor (LIF) and TSC media with fibroblast growth factor 4 (FGF4) and heparin. ESC culture were split every 2 days at a passage ratio of 1:8 and had its media changed on a daily basis. TSC culture were split every 4 days at a passage ratio of 1:12 and had its media changed once every two days. Both cell lines were maintained on plates containing mouse embryonic fibroblasts (MEFs). ESC culture were split onto MEF-free gelatinized plates and TSC culture onto MEF-conditioned media two passages prior to harvest.

2.1.2 Cell fractionation

Compositions of buffers used in cell fractionation protocol are summarized in **Table 1.**

4 x 10cm plates (i.e., approximately 40 million cells) of ESCs on gelatine and TSCs on MEF-conditioned media were washed with PBS and lifted with trypsin treatment. This harvest for cell fractionation was done no longer than 14 days after initial thaw. All harvested cell culture were at between passage number of 18 to 26 and at approximately 80% plate confluency. Trypsinized cells were singularized, counted, and suspended in Suspension Buffer at a ratio of 40 million cells per 1mL of buffer. This suspension was centrifuged for 4 minutes at 1300 *g* and resulting supernatant was collected, re-suspended in lysis buffer to be incubated on ice for 5 minutes and centrifuged

again for 4 minutes at 1300 *g*. The resulting supernatant represented the cytoplasmic fraction and the remaining pellet the nuclear fraction. The supernatant set aside, the nuclear pellet was suspended and washed in suspension buffer to be re-suspended again. Both cytoplasmic and nuclear fraction samples were divided into 1.5 million cell equivalents per aliquot and stored at -70ºC for storage.

Table 1: **Buffers used in cell fractionation protocol**

Suspension Buffer		
Item	**Stock concentration**	**Final concentration**
HEPES buffer (pH 7.5)	1 M	20 mM
Potassium chloride	1 M	10 mM
Magnesium chloride	1 M	1.5 mM
Sucrose		0.34 M
Glycerol		10%
Deionized water		Top off
Lysis buffer		
Suspension buffer + Triton X-100 (Sigma)		0.2% Triton X-100

2.1.3 RNA isolation and cell fractionation quality control

RNAs from fractionated cytoplasmic and nuclear samples were collected using Norgen-Biotek Total RNA Purification Kit. All isolated RNA samples underwent DNase treatment. Collected RNA samples (i.e., cytRNA and nucRNA) were used to synthesize cDNA libraries for quantitative reverse transcription PCR (qRT-PCR). cDNA synthesis was done using High Capacity RNA-cDNA kit for RNAs (Thermo Fisher) and MicroRNA Reverse Transcription Kit (Thermo Fisher) for microRNAs and snoRNAs, respectively. qRT-PCR was then performed with TaqMan probes for cell-specific cytosolic markers Nanog, Oct4, and Cdx2, universal cytosolic markers Actb and Gapdh, and snoRNA markers sno136, sno142, and sno202.

A western blot of cytosolic marker p38 was also used for verification of cytosolic-nuclear separation. Gel electrophoresis was performed with 4x Laemmli Sample Buffer (Sigma) spiked with β-mercaptoethanol. Precision Plus Protein Dual Color Standards (Bio-Rad) were used as the pre-stained ladder. Resulting gel was transferred to a PVDF membrane to be blocked with skim milk powder solution (1 mg skim milk powder, 10 µL Tween, 1 ml 10x TBS, and 9 mL water). Primary antibody for p38 marker was used with anti-mouse secondary antibody. DAB staining kit (Abcam) was used for western blot visualization.

2.1.4 RNA concentration measurement and RNA-seq

1µL of cytRNA and nucRNA was diluted in 9µL distilled water to be loaded onto Qubit Fluorometer for RNA concentration measurement. All measurements were obtained at equal volumes. Each measurement for a replicate sample was taken in four technical replicates and the mean concentration was reported. 1µg of each RNA sample were sent for RNA and small RNA sequencing in three replicates. RNA-seq preparation was done by the sequencing facility with NEB Ultra RNA Library Preparation Kit with ribosomal RNA depletion, multiplexing, and ERCC spike-in quality controls. Sequencing was performed on Illumina NextSeq500 to generate 150 base-pair paired-end reads. Small RNA-seq preparation was done with Illumina TruSeq Small RNA Preparation Kit with ERCC spike-in quality controls. Single-end 150 base-pair reads were generated on Illumina NextSeq500.

2.1.5 Pre-analysis bioinformatics processing of sequencing data

FastQC, cutadapt, HISAT2, Integrative Genomics Viewer, RSeQC, and featureCounts were used in the command line interface running Ubuntu 18.04.

Obtained RNA-seq and small RNA-seq read files underwent quality control with FastQC and adapter trimming with cutadapt. Trimming with cutadapt was performed with minimum length of trimmed reads of 15 bases (`-m 15`). FastQC parameters used as quality assurance were base quality over read length, adapter content, and per-sequence quality scores. Alignment to the mm10 genome was done with the splice-aware alignment software HISAT2 on pair-end read mode and `--rna-strandedness RF` option. HISAT indices for the mm10 genome was downloaded from HISAT2's official database. Alignment files were visualized before feature quantification on

the Integrative Genomics Viewer as a means of alignment quality control. As well, RSeQC software was used for further quality control of alignment, particularly the `bam_stat.py`, `read_distribution.py`, and `genebody_coverage.py` modules.

Exon count data from alignment files were generated using featureCounts module under the Subread package using default options and mm10 gene annotation (GTF) file obtained from UCSC database. Custom intron annotation file was created using an R script (provided in **Chapter 6**) and the exon annotations from the previous GTF file. This custom intron GTF file was then used to generate intron count data with featureCounts, as before.

2.2 Results

2.2.1 *RNA concentration measurements across two fractions*

Accurate quantification of RNA concentration in the cell fraction samples is necessary for downstream mass balance correction. The Qubit Fluorometer was used to measure RNA concentration due to its ability to measure low amounts of RNA (i.e., lower limit of detection). This ability to accurately measure very low amount of RNA is important for the purpose of measuring differences in RNA content across fractionated cell lysates (i.e., cytoplasmic and nuclear compartments) as well as in measuring miRNA content in fractionated cell lysates (**Section 4.1.1**). Furthermore, high sensitivity in measurements is necessary as obtained RNA concentrations are used directly to calculate correction factors.

The Qubit Fluorometer measurements for RNA content is shown in **Table 2** and predictably indicate an imbalance of total RNA concentration across the cytoplasmic and nuclear fractions. Interestingly, the imbalance is reversed in ESC samples versus TSC samples, such that the mean concentration across the replicates is higher in cytoplasmic samples compared to nuclear samples in ESCs, but not in TSCs.

Each measurement was taken at four technical replicates and the mean concentration per sample replicate was used to calculate correction factors. Applying the concentration-based correction

factors to raw count data in downstream analysis ensures cytoplasmic-nuclear expression comparisons can be made without violating the assumption that total RNA content across samples is equal. This method of discrete count correction prior to bioinformatic processing corrects for the imbalance in cell equivalence at the RNA-seq step of the experiment; since even though an equal mass of RNA was amplified and sequenced across the fractions, they represent unequal amount of cell equivalence – as evident by RNA concentration measurements (**Table 2**). Taking the corrected counts to edgeR for filtering and voom/limma for differential expression analysis then allows for cytoplasmic-nuclear pairwise comparisons in gene expression without biases caused by compartmental mass imbalance.

Table 2: **Measured RNA concentrations in ESC and TSC fractions**

	Replicate sample	ESC sample (ng/μL)	TSC sample (ng/μL)
Cytoplasmic fraction	1	574	165
	2	906	152
	3	650	136
Nuclear fraction	1	324	222
	2	210	196
	3	264	180

2.2.2 Quality control for proper cell fractionation

A qRT-PCR survey was performed in order to validate the separation of cytoplasmic and nuclear compartments after the fractionation protocol. The analysis of subcellular fractions does not allow for normalization with housekeeping genes due to the uneven distribution of RNA species in the samples. Therefore it was ensured that equal amount of RNAs were used across the samples in preparation of PCR. TaqMan probes for ESC and TSC protein-coding RNAs were used to test for cytoplasmic enrichment and snoRNA probes were used for nuclear enrichment. Cell-specific

genes as well as ubiquitous cytoplasmic genes such as Actb and Gapdh were used in the panel to ensure signal detection.

PCR results show the enrichment of cytoplasmic mRNAs and cell specific mRNAs in the cytoplasmic fractions versus the nuclear fractions (**Figure 5, A** and **B**). The asymmetrical distribution across the fractions is conserved in all technical replicates and both cell samples. Inversely, snoRNA detection was significantly higher in the nuclear fractions versus cytoplasmic fractions in both cells and all replicates (**Figure 5, C** and **D**).

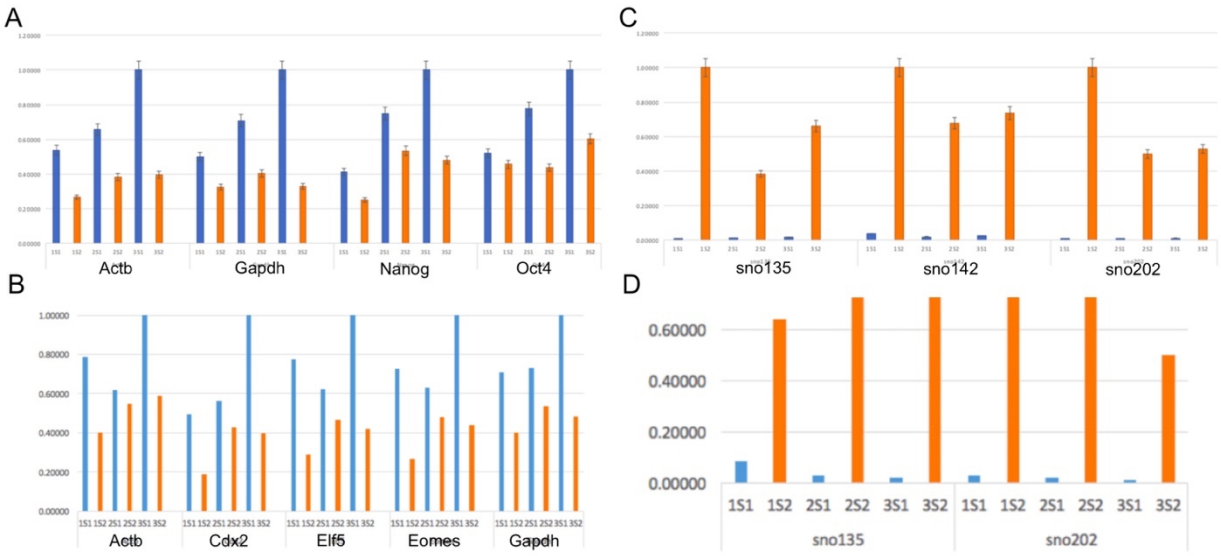


Figure 5: **qPCR panel for cell fractionation validation**

A) qPCR panel for Actb, Gapdh, Nanog, and Oct4 show higher level of expression in ESC cytoplasmic fraction across three replicate samples (blue) versus nuclear fraction (orange); **B)** qPCR panel in TSC samples show higher cytoplasmic enrichment of Actb, Gapdh, Cdx2, Elf5, and Eomes compared to nuclear fractions; **C)** Panel of snoRNAs show significantly higher expression in nuclear fractions in ESCs versus cytoplasmic fractions; **D)** sno135 and sno202 show higher expression in nuclear fraction in TSCs versus cytoplasmic fraction.

A western blot for the cytoplasmic marker p38 further validates the cytoplasmic-nuclear separation (**Figure 6**). Strong bands at around 37 kDa suggest p38 enrichment in the cytoplasmic samples while no such bands appear for the nuclear samples. This result suggests that, as with the PCR panel, the cell fractionation protocol was effective in separating cytoplasmic and nuclear content.

This ability to compartmentalize cellular content allows for identification of spatial specific transcriptomic processes in downstream computational analysis.

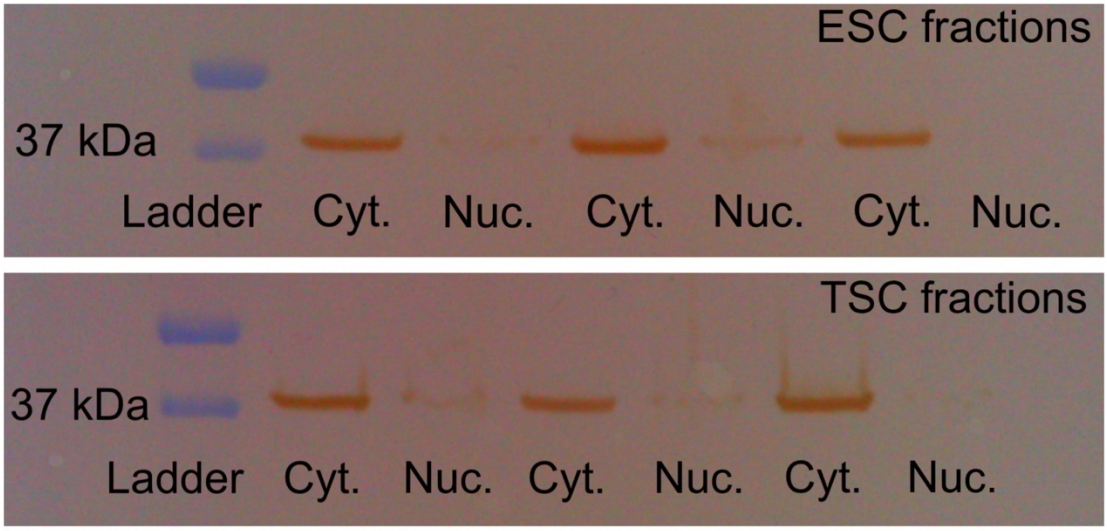


Figure 6: **Western blot validation for p38 enrichment in cytoplasmic fractions**

Western blot for p38 antibody show strong bands in the cytoplasmic fraction samples in ESCs (**top panel**) and in TSCs (**bottom panel**).

2.2.3 Visualization of gene alignment

After RNA-seq, the sequencing files underwent quality control for base sequence quality and adapter content, which was addressed to by sequencing primer trimming and filtering out extremely small reads (**Section 2.1.5)**. Subsequent alignment to the genome resulted in representations of fragmented reads mapping onto genomic coordinates. The discrete distribution of mapped reads is harnessed in downstream analysis to infer levels of gene expression, and by extension, differential gene expression when comparing across samples.

Gene alignment was visualized with Integrative Genomics Viewer to highlight the asymmetric distribution of read counts across the cytoplasmic and nuclear samples. Predictably, the density of reads mapped to the Cdx2 gene in TSC samples is significantly larger in the cytoplasmic fraction than in the nuclear fraction (**Figure 7**). This observation is reversed for reads mapped to the nuclear localized lncRNA Xist gene (**Figure 8**).

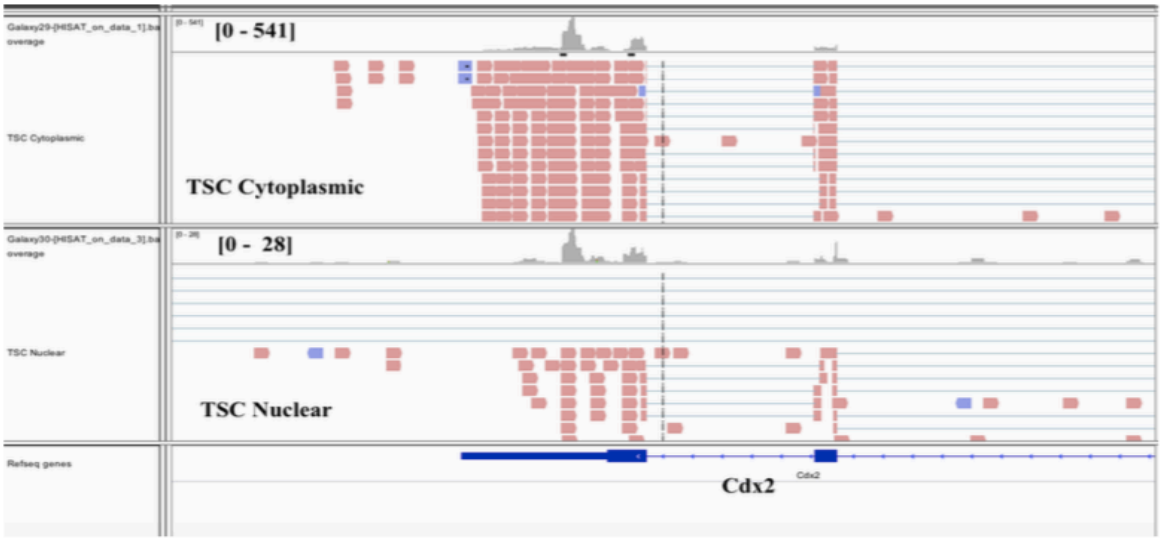


Figure 7: **IGV illustration of sequencing reads mapped to Cdx2 gene in TSC fractions**

IGV result show significantly higher population of mapped sequencing reads onto Cdx2 gene in the cytoplasmic fraction (**top panel**) compared to the nuclear fraction (**bottom panel**).

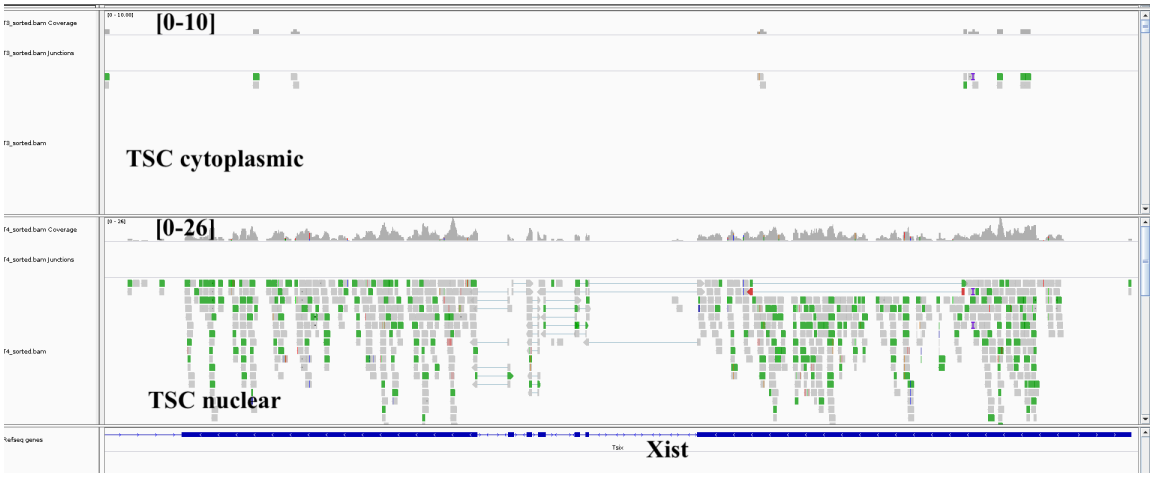


Figure 8: **IGV illustration of sequencing reads mapped to Xist gene in TSC fractions**

IGV result show significantly lower population of mapped sequencing reads onto Xist gene in the cytoplasmic fraction (**top panel**) compared to the nuclear fraction (**bottom panel**).

A splice-aware aligner such as HISAT2 also captures reads that span over intronic regions (i.e., split reads). Instances of split reads denote exon-exon boundaries in a spliced transcript and as such, will be used in downstream analysis to quantify splicing events.

Sashimi plots can be generated using the Integrative Genomics Viewer to visualize splice junctions and split reads. Such plots also show the differential distribution of reads not only spanning exon-exon boundaries, but also reads that fall within the intronic regions within the gene (**Figure 9**). In downstream analysis, reads that span not only the spliced boundaries, but as well the reads that span unspliced (i.e., retained) exon-intron boundaries will be quantified to infer differences in transcript processing in a cell fraction specific context.

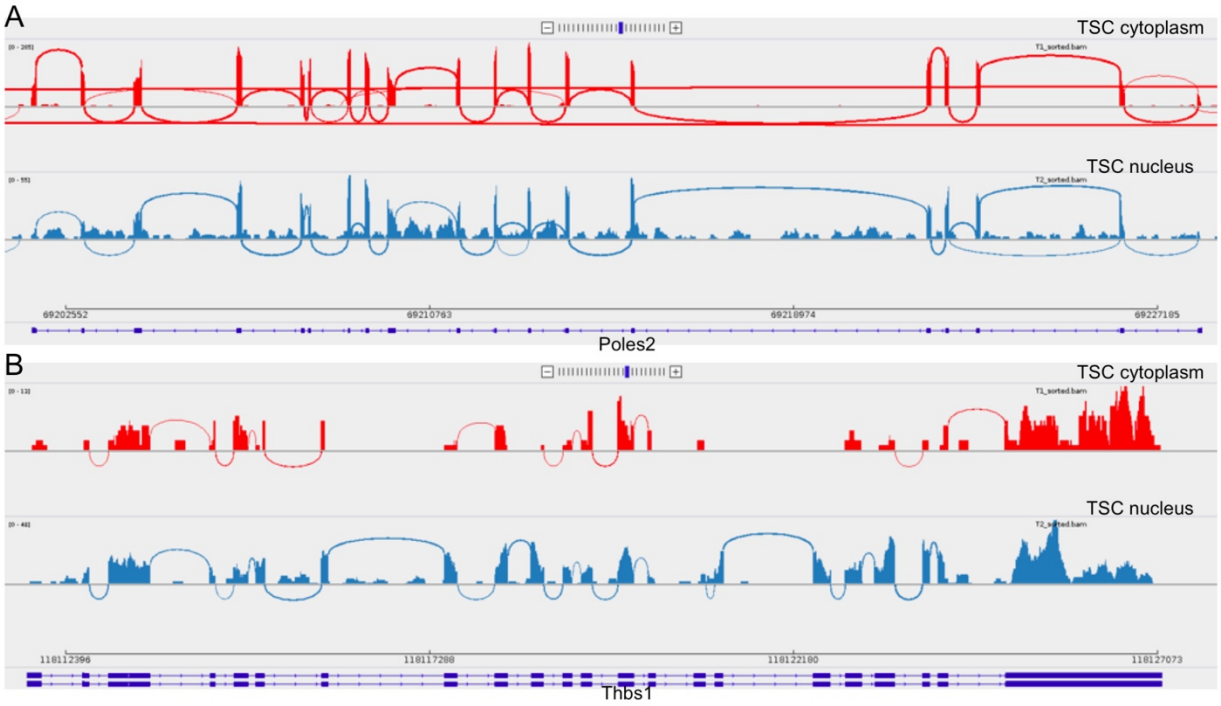


Figure 9: **Sashimi plot showing exon-exon junction boundaries in gene alignment**

Using a splice-aware aligner such as HISAT2 for gene alignment allows user to infer splice junctions from split reads; gaps in split reads can be visualized using sashimi plots.

2.3 Discussion

2.3.1 qRT-PCR and western blot results support subcellular fractionation

Asymmetric enrichment of RNA and protein species across the cytoplasmic-nuclear boundary allows for cell fractionation validation using detection assays. A complication in a qRT-PCR experiment of cell fractions, however, is that the analysis of subcellular fractions does not allow

for housekeeping gene normalization. This issue is addressed in literature with the emphasis on the importance of ensuring an equal amount of RNAs is used across PCR samples[92]. An equal amount of RNA for each subcellular fraction sample allows for representative measurement of probe RNA per unit RNA of each fraction. Therefore, from **Figure 5** it can be inferred that per equal unit of RNA, the cytoplasmic fraction samples in ESCs and TSCs show higher levels of mRNAs relative to the nuclear fraction (**Figure 5, A** and **B**). Reciprocal results are seen in snoRNA PCR panels, where snoRNA expression shows higher levels in nuclear fraction per unit RNA relative to cytoplasmic fraction in both ESCs and TSCs (**Figure 5, C** and **D**). Evidently, the equality of cytoplasmic and nuclear RNA used in PCR allows for such inferences to be made. The asymmetric expression normalized for RNA quantity supports fractionation of cytoplasmic and nuclear components.

PCR result for mRNAs is interesting as consistent nonzero expression is shown in nuclear fractions (**Figure 5**, **A** and **B**). This is in stark contrast to the PCR panel for snoRNAs where the cytoplasmic fractions show significantly lower expression levels relative to the nuclear fractions (**Figure 5**, **C** and **D**). This behaviour suggests transcripts for protein-coding RNAs exist in detectable amounts in the nucleus. Indeed, in a literature review by Ben-Yishay *et al*, the authors suggest the subcellular dynamics of mRNAs within the nucleus itself play a vital role in gene expression regulation machinery[93]. The authors suggest the life cycle of mRNAs in the nucleus consist of transcription, maturation, nucleoplasmic transport, and nuclear export – in which the steady-state population of mRNAs is most significantly affected by the rate of transcription. Additionally, Efroni *et al* claim that pluripotent stem cells such as ESCs exhibit a hyperactive transcriptional activity, due to their plasticity requiring silencing of tissue-specific genes[94]. The authors make this claim by showing that a larger portion of the ESC genome is active compared to differentiating cells – suggesting that pluripotent cells' genomes are globally hyperactive and express large portions of the genome. Finally, in a kinetics study of mRNA dynamics in *Drosphila* Kc167 cells, Chen *et al* found that the kinetic rate constant for transcription accounted for 89% of variance in steady state transcript abundance, whereas the rate of cytoplasmic decay and nuclear export accounted for just 10% and 0.5%, respectively[95]. Such findings suggest that the subcellular dynamics of mRNAs is heavily influenced by transcription. Factors that influence the rate of transcription then, should in turn influence the levels of cytoplasmic-nuclear differential

expression of mRNAs. It is also plausible then – especially with hyperactive transcriptional activity – there exist detectable amounts of RNA species retained in the nucleus. This nuclear enrichment of mRNA transcripts can be attributed to the nuclear life cycle of mRNAs as discussed above, or an active detainment of mRNAs – which will be discussed later.

As an additional layer of quality control, protein markers were assessed in a western blot in cytoplasmic and nuclear fractions. In particular, the enrichment of p38 marker in the cytoplasmic fractions across all replicate samples support the separation of two fractions (**Figure 6**). In a survey of best practices in subcellular fractionations, Mayer *et al* recommends western blots as a means of quality control using compartment specific protein markers[57]. In a detailed review of cell fractionation protocols, Gauthier *et al* also suggested western blotting as the standard for fractionation optimization[96]. Indeed, the low limit of detection as well as the high specificity of western blotting ensured minimal detection of p38 in nuclear fraction as well as strong detectable bands in the cytoplasmic fraction. The results from PCR and western blots together support the efficacy of fractionation prior to RNA-seq.

2.3.2 IGV reveals quantifiable genomic features in cell fraction data

Gene alignment of obtained RNA-seq reads using HISAT2 allows for read quantification necessary for differential expression analysis. In RNA-seq, gene expression data is quantified by the number of sequencing reads mapping to each genomic feature of interest[97]. This 'mapping' behaviour is visualized using IGV (**Figures 7** and **8**). IGV allows for a genome-wide exploration of mapped sequenced reads, and using its built-in mm10 genome annotations, it provides an additional means of fractionation quality check. As seen in **Figure 7**, the cytoplasmic fraction (top panel) show a significantly higher number of mapped reads onto Cdx2 gene compared to the nuclear fraction (bottom panel). For a nuclear localized gene such as Xist, this trend is reversed (**Figure 8**); this result suggest RNA-seq is able to recapitulate the asymmetric distribution of gene expression as shown in PCR results. As the expression level of a gene will be estimated by – following processing and normalization – mapped read counts, visualization with IGV also provides an early insight into the differential expression profile between the two cell fractions.

The advantage of RNA-seq over traditional gene expression assays such as microarrays is the ability to profile the genome at a single-base resolution. This allows for quantification of specific genomic features within a gene itself – such as exons, introns, and exon-intron junction boundaries – using mapped reads. Indeed, sashimi plots in **Figure 9** show splicing events captured by mapped gene alignment, as well as gene coverage in intronic regions in between adjacent exons.

Chapter 3

3 Differential Expression Profile of mRNAs and ncRNAs in ESC and TSC fractions

3.1 Methods

R packages NOISeq, edgeR, voom, limma, topGO, and clusterProfiler were used under R version 3.6.3 running on macOS Mojave. Command line tools BEDOPS and Samtools were run in a Linux environment running Ubuntu 18.04. Web interface tool REVIGO was accessed and used on Firefox Browser 76.0.1 on macOS Mojave. Python package CirGO was run under Python version 2.7.5 running on the command line interface.

3.1.1 Normalization of cytoplasmic-nuclear mass imbalance

Generated cytoplasmic and nuclear count data were normalized for cytoplasmic-nuclear RNA imbalance by scalar multiplication. Scalar normalization factors were derived from the original RNA concentration measurements prior to RNA-seq. Normalizing relative to the RNA concentration in the cytoplasmic fraction (i.e., setting $\alpha = 0$), values for ß and γ are calculated using quotients of average cytRNA, nucRNA, and total RNA concentrations. As such, the following correction factors were used on raw count data before processing with edgeR. Significant figures for the correction factors were carried over from the original fluorometer measurements for RNA concentration.

$$\alpha x + \beta y = \gamma z$$

$$x + 3y = 1.6z$$

(ESCs)

$$x + 0.57y = 0.75z$$

(TSCs)

3.1.2 Quality control of count data using NOISeq

R package NOISeq was used to generate quality control reports prior to differential expression analysis. Compartment-normalized read counts (from **Section 2.1.5**) are used as input for PCA plots and boxplots to detect any sample group biases. Cell type (i.e., ESC or TSC sample) and compartment (i.e., cytoplasmic or nuclear) information for each sample are passed as factors for `QCreport()` function with default parameters (i.e., `samples = NULL, norm = FALSE`)

3.1.3 Differential expression analysis with edgeR-voom-limma

Normalized cytRNA and nucRNA exon and intron counts were processed using edgeR. Log-counts-per-million (lcpm) values were calculated and lowly expressed counts filtered out using edgeR's built-in `filterByExpr()` function. voom was used to apply mean-variance weights to count data for Bayesian modeling with limma. Linear modeling in limma was carried out using `limfit()` and `contrast.fit()`. Design and contrast matrices in limma were constructed such that both cytoplasmic-nuclear fraction differential expression (i.e., `ESCcyt vs. ESCnuc` and `TSCcyt vs. TSCnuc`) as well as cell-to-cell comparisons (i.e., `ESCcyt vs. TSCcyt` and `ESCnuc vs. TSCnuc`) can be made. Further, read counts were segregated by the REFSEQ gene identifier prefix `NM_` and `NR_`, in order to separate counts belonging to mRNAs and ncRNAs. Differential expression results for each pairwise comparison, for each subclass of sequencing data, was generated using `decideTests()` with options `adjust.method = "fdr"` and `p.value = 0.05`.

3.1.4 Gene ontology enrichment analysis using TopGO

List of differentially expressed genes for each comparison with corresponding adjusted p-values were used as input for TopGO for gene ontology analysis. Gene Ontology database category for biological processes (GO: BP) was used to produce enriched terms for each differential expression profile. All annotations for ontological terms were loaded from the Org.mm.eg.db package. TopGO's built in function for Kolmogorov-Smirnov (K-S) test for overrepresentation on gene ontologies (`run.test(algorithm = "classic", statistic = "ks")`) was used to produce a list of top enriched ontological terms with corresponding K-S value.

3.1.5 Visualization of gene ontologies using REVIGO and CirGO

Lists of top enriched GO:BP terms for each differential expression profile with corresponding K-S values are used to generate tree-map representations on REVIGO's web interface (http://revigo.irb.hr/). REVIGO's clustering result is visualized using CirGO with default parameters (`numCat = 40`) to produce hierarchical representation of top enriched ontological terms per differential expression profile.

3.1.6 Correlation analysis between transcript features and differential expression

Using exon and intron annotations from **Section 2.1.5** were used to extract total intron lengths and the number of exons per given gene. Lengths for individual introns per gene were summed to yield the total intron length. Genes whose total intron length which were deemed outliers (i.e., outside 1.5 times the interquartile range above the upper quartile and below the lower quartile) were excluded from the following analysis. The differential expression analysis result from limma (**Section 3.1.2**) was used to generate Spearman rank correlation coefficient between log-fold-change of differentially expressed genes in the cytoplasmic-nuclear comparisons and intron length. This was repeated for correlation between log-fold-change and the number of exons per gene.

3.1.7 Calculation of exon-intron proportion quotients

Cytoplasmic-nuclear normalized exon and intron count data were collated by cell type. Lowly expressed count data were filtered out by requiring that at least 25% of samples have counts greater than 25. Filtered counts were then used to calculate the following ratio to estimate the extent of exon-intron proportions per individual gene:

$$Q_i = \frac{C_{exon_i}}{C_{exon_i} + C_{intron_i}}$$

where: C_{exon} = read count of exons per gene

C_{intron} = read count of introns per gene

Q_i = exon-intron proportion quotient per gene

3.1.8 Quantification of exon-intron junctions using BEDOPS

GTF file containing all genomic feature elements and base coordinates of the mm10 genome was used. All exonic features were extracted to create an `exons.bed` file usable for BEDOPS. A series of `awk` scripts (available in **Chapter 6**) were used to first convert exon annotations to include both exon and intron annotations, then to annotate for padded boundary junctions between the exon and intron coordinates. A base-pair pad of 5 bases (i.e., +/- 5 base-pairs around each exon-intron junction point) was used in current methodology to produce an exon-intron junction annotation file. Using BEDOPS command `bedmap` with options `--echo` and `--count`, gene alignment files for fractionated ESC and TSC data were mapped to the junction annotations, producing count tables. Raw count tables were corrected for cytoplasmic-nuclear mass balance, filtered for lowly expressed reads, and used as input for edgeR-voom-limma workflow as before to generate a list of differentially expressed junctions.

3.1.9 Quantification of split alignment reads

In order to infer population of alignment reads situated at exon-to-exon boundary junctions, `samtools` and `awk` was used to extract, from the original alignment files, all reads that contained the character 'N' in the CIGAR string. Resulting alignment file containing only the split reads were then mapped to the `exons.bed` annotation file with `bedmap` as before, to generate count tables. Raw count tables were processed as before (**Section 3.1.8**) and underwent differential expression pipeline.

3.1.10 Calculation of junction quotients for intron retention estimation

To estimate the extent of intron retention using junction counts, the ratio of unspliced (exon-intron) junction expression to total (sum of exon-intron and exon-exon) junction expression was calculated using lcpm values as a measure of library-size-corrected expression.

$$JQ_i = \frac{EI_i}{EI_i + EE_i} = \frac{unspliced\ reads}{unspliced + spliced\ reads}$$

where: EI_i = exon-intron junction read counts for gene *i*

EE_i = exon-exon junction read counts for gene *i*

JQ_i = junction quotient for gene *i*

3.1.11 Binning junction quotients and gene set enrichment analysis with clusterProfiler

Collated table of calculated junction quotients for cytoplasmic and nuclear read counts in ESCs and TSCs were split into four quantiles. Genes in each quantile were subject to over-representation analysis using GO:BP database and `enrichGO` function in clusterProfiler (using `ont = "BP"`). Visualization of GO analysis was also done with clusterProfiler's built-in graphical functions `dotplot()` and `emapplot()`.

3.2 Results

3.2.1 Experimental design and classification of subclasses of count data

Using the feautreCounts module under Subread package with exon and intron annotations resulted in exon and intron count data respectively. Both exon and intron count data were subject to differential expression and gene set enrichment analysis in separate pipelines in order to infer subcellular differences in exon-intron proportions and by extension, splicing behaviour. Furthermore, each set of count data was divided by mRNA (i.e., REFSEQ gene identifier prefix NM-) and ncRNA (i.e., REFSEQ gene identifier prefix NR-) data in order to identify whether localization behavior differs by RNA class. Splitting the data in such a way also allows for a more robust identification of enriched functional classes of genes in downstream analysis. It is of note that currently in REFSEQ database, ncRNA annotations include annotated snoRNAs, miRNAs, and lncRNAs, as well as ribosomal RNAs.

All downstream processing of count data with NOIseq, edgeR, voom, and limma was performed using both sets of count data (i.e., exon and intron) in separate analyses. For brevity, pre-differential expression results from exon count data is shown in **Sections 3.2.3**. Differential expression and functional gene set enrichment analyses were performed with both exon and intron data as well, and are presented separately in **Sections 3.2.4, 3.2.5, 3.2.7,** and **3.2.8**.

3.2.2 Quality control report for compartment normalized count data

Prior to differential expression, unsupervised clustering of the sample groups in count data was performed to ensure predictable similarities and dissimilarities in the data. As samples are expected to cluster together within the experimental condition of interest, in the current data set it is expected sample groups will cluster by parent cell type (i.e., ESC or TSC) and by cell fraction (i.e., cytoplasmic or nuclear).

Principal component analysis (PCA) plots from NOISeq generated quality control report predictably show fair clustering of samples by their cell type and cell fraction (**Figure 10, A & B**). ESC and TSC samples are separated in the first dimension while cytoplasmic and nuclear fractions are separated in the second dimension. As the first dimension in the PCA plot represents the largest variation in the data, it can be inferred that the largest source of difference in the current data set is related to differences in phenotype (i.e., samples co-vary by cell type), rather than subcellular variations. Furthermore, the clustering of cytoplasmic and nuclear samples in the count data suggest the method of normalizing for cell equivalence (explained in **Section 3.1.1**) does not introduce artifacts or biases that affect co-variance in the data.

The shape of the count distribution can be visualized using boxplots. Boxplots can reveal skewness in the data as well as the population of statistical outliers at either ends of the distribution. NOISeq generated boxplots of the count distributions across sample groups suggest skewness of data towards upper range of data with heavy tailing (**Figure 10, C & D**). This increasing variance at higher corresponding value in count data is typical of discrete count distributions, which will consequently be addressed with voom in downstream analysis.

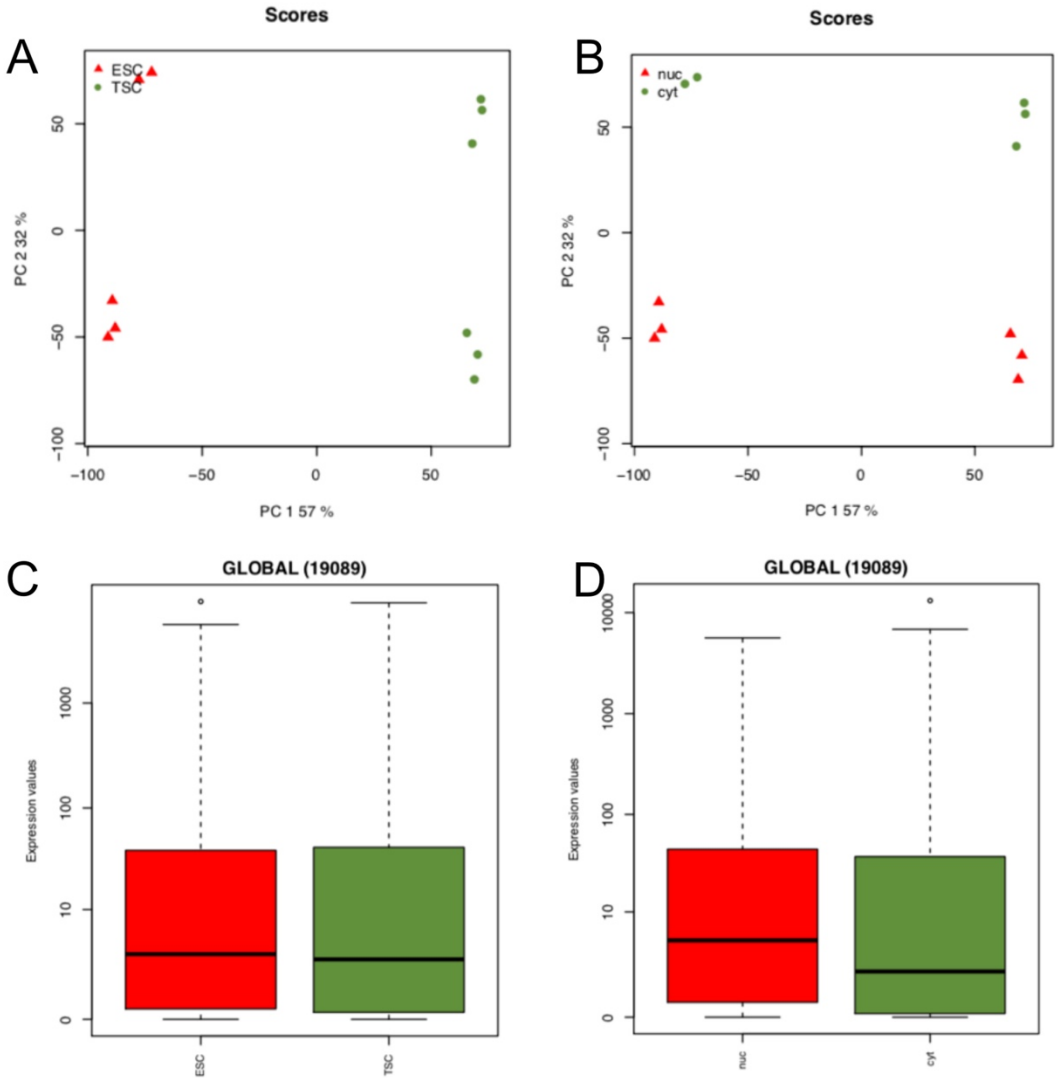


Figure 10: **NOIseq quality control plots of fractionated count data**

A) PCA result in fractionated count data show clustering of sample groups by their parent cell types in the first dimension; **B)** sample groups are clustered by their fraction components in the second dimension; **C)** in boxplots of count data, both ESC and TSC groups show skewness towards higher counts; **D)** similar result is shown across the two fraction data.

3.2.3 Pre-differential expression processing with edgeR

After mass imbalance correction, count data was converted to log-counts-per-million (lcpm) metric using edgeR. Using raw counts for differential expression analysis is not sufficient as alignment read counts are dependent on factors such as transcript lengths, size of the gene population, and sequencing artifacts. In particular, samples with greater sequencing depth will result in higher raw counts and therefore must be accounted for.

For the purpose of the current study, considerations in biases related to gene and transcript length will not be addressed as such correction is not necessary when comparing changes in expression in the same genes across samples. As such, transformations such as fragments per kilobase of transcript per million (FPKM) or reads per kilobase of transcript per million (RPKM) were not used. FPKM and RPKM transformation are more suitable for differential expression analyses comparing expression across multiple genes, or quantifying absolute levels of gene expression.

The lcpm transformation addresses differences in sequencing depth by adjusting the number of feature mapped reads to the total number of reads. After count transformation, filtering out lowly expressed reads, according to edgeR's built-in `filterByExpr` function, removes a large population of genes prior to differential expression (**Figure 11**). Filtering was necessary to remove genes which may not be biologically significant.

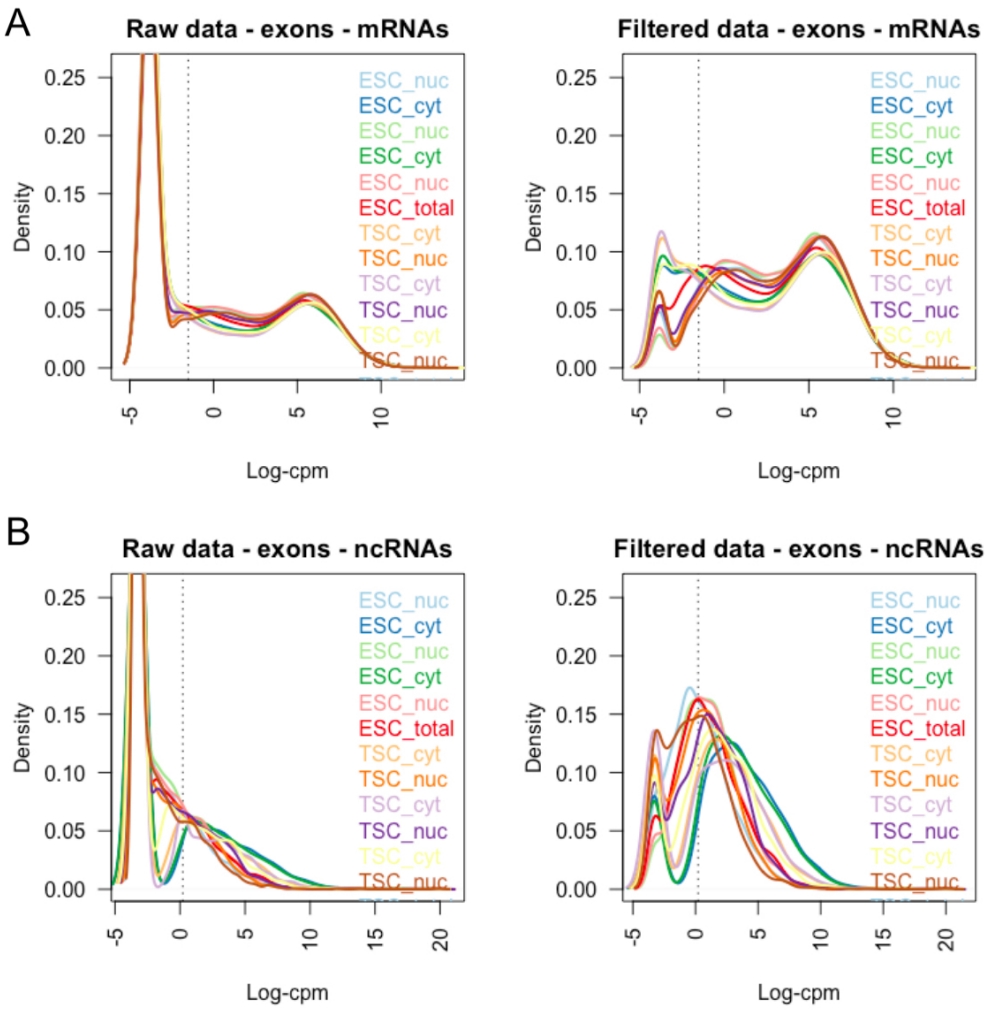


Figure 11: **The effect of filtering out lowly expressed reads in lcpm data**

A) Filtering lowly expressed counts with edgeR after conversion of raw data into the log scale leads to removal of spike in expression at negative lcpm values in mRNA data; **B)** similar effect of filtering is shown in ncRNA data; removal of lowly expressed reads prior to differential expression is necessary in order to filter out reads that may not be as biologically meaningful.

3.2.4 Differential expression profile using exon counts

A pervasive characteristic of discrete variables such as read counts (versus continuous variables such as microarray intensity measurements to measure gene expression) is that the population mean and variance are not independent. The consequence of this is that counts at higher average values tend to have higher variances – this agrees with boxplots generated by NOIseq (**Section 3.2.1**). This led to development of differential expression analysis methods specifically designed for read count data, such as the negative binomial (NB) distribution method used by edgeR. NB

method relies on a Poisson distribution model to fit count data but has been shown to be inadequate in type I error control when count dispersion is high or if the number of samples is low.

Whereas limma uses linear modeling suitable for normal distributions in the case of microarray data, in order to use discrete variables such as raw read counts (or log-transformed counts), voom attempts to address the mean-variance problem by applying mean-variance weights to individual counts. This method of 'variance modeling at the observational level' coupled to linear modeling with limma (i.e., voom-limma method) was shown to better control for type I error compared to NB or other Poisson based methods even when the number of samples was low. This advantage in performance was also shown to be even more significant when sequencing depths across samples were different.

After voom treatment, a plot showing the means (x-axis) versus the variances (y-axis) for individual genes show the removal of mean-variance dependence (**Figure 12**). voom-treated count data can then be taken into linear modeling for limma for differential expression.

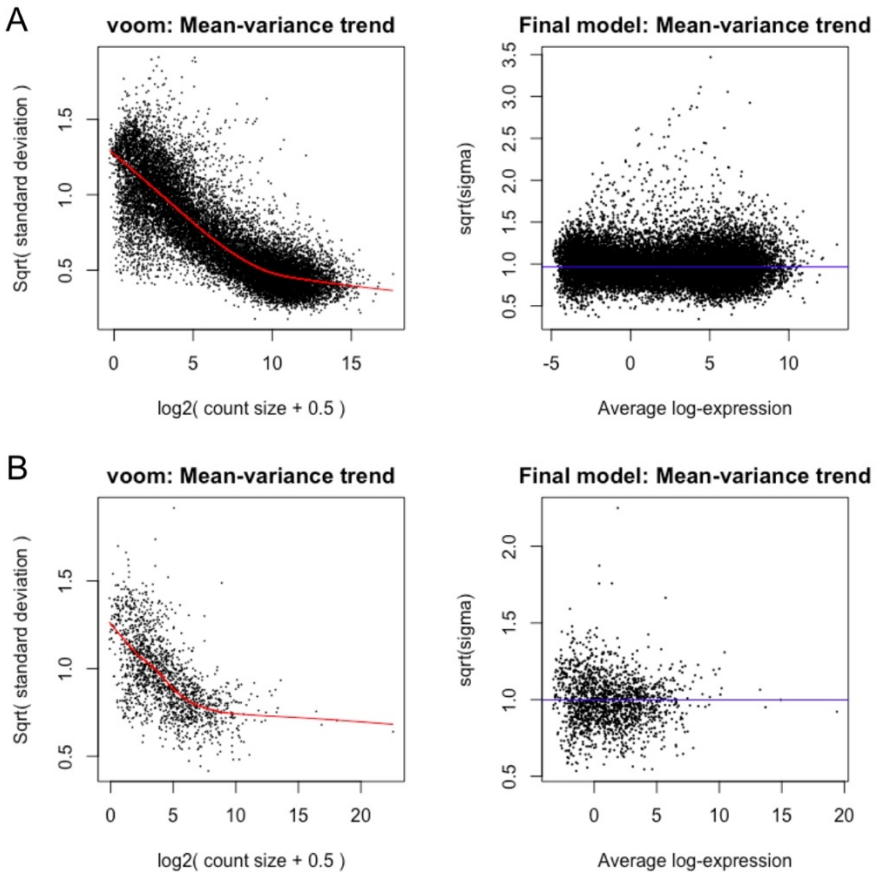


Figure 12: **The effect of voom treatment on mean-variance relationship in expression data**

A characteristic of discrete data such as in the case of expression read counts is the dependence of mean on the variance; voom applies mean-variance precision weights to remove this bias prior to ensure suitability of differential expression tools such as limma. The effect of voom treatment shows the removal of mean-variance dependence trend in **A)** mRNA data and **B)** ncRNA data.

Using limma's `decideTests()` function with `adjust.method = "fdr"` and `p.value = 0.05` returned the differential expression profile in pairwise comparisons outlined in **Section 3.1.3**. In mRNA data, the differential expression profile in the cytoplasmic vs. nuclear comparisons showed a higher number of genes down-regulated in the cytoplasmic fraction relative to the nuclear fraction (**Figure 13, A & B**). This observation held true in both ESC and TSC data. In cell-to-cell comparisons, a fairly symmetrical distribution of up-regulated and down-regulated genes was shown in the cytoplasmic and nuclear fractions (**Figure 13, C & D**).

The significantly larger population of mRNAs down-regulated in the cytoplasmic fraction versus the nuclear fraction may be explained by considerations in mRNA kinetics within the cell. As the

fate of mRNAs within a cell is governed by the rate constants of transcription, nuclear export or degradation, and translation - which in turn are governed by regulatory processes related to RNA binding proteins (e.g., poly(A) binding proteins affecting mRNA stability) and *cis*-regulatory elements – it is difficult to draw inferences on a kinetics level with a biological 'snapshot' as presented here. Nevertheless, as the nuclear export of mRNAs itself is a complex process with its rate-determining step being the passive diffusion of the mRNA complex to the nuclear pore, it is possible that these mRNA species were captured and reflected in the differential expression profile. As well, due to forgoing the use of poly(A) selection in acquisition of RNA-seq data, it is possible that the nuclear mRNA population is inflated by the presence of nascent mRNA.

Furthermore, nuclear degradation of mRNA is yet another complex, multi-faceted process which couple to every step of mRNA processing. Therefore it is also possible that detection of degraded mRNA contributed to the large number of differentially expressed mRNAs in the nuclear fraction.

In order to draw meaningful conclusions founded on kinetics, however, a time-series type of a study is necessary.

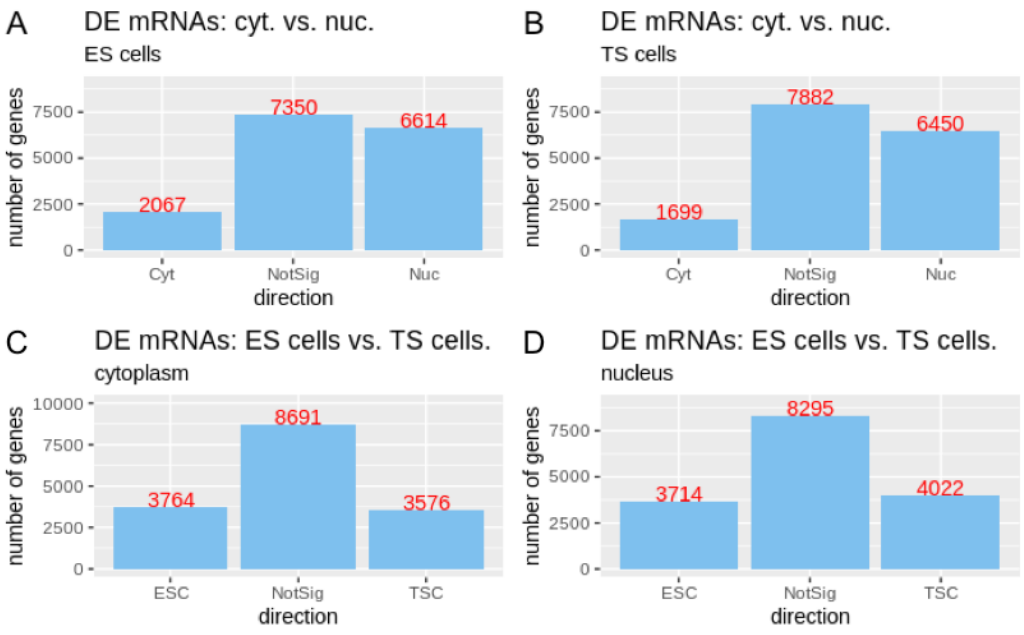


Figure 13: **Differential expression profile of mRNAs using exon count data**

Differential expression using exon counts with limma output (adjusted p-value < 0.05) show **A)** the up-regulated population of mRNAs in the cytoplasmic fraction is smaller than in the nuclear fraction in ESC data; **B)** similar trend is shown in TSC data; **C)** cell-to-cell comparisons show the up-regulated population of mRNAs in the cytoplasm show no particular bias towards either direction; **D)** similar result is shown in the nuclear fraction

In order to better understand the distribution of differentially expressed mRNAs across the cytoplasmic and nuclear fractions, exploratory plots such as M-A plots are generated. M-A plots aim to contextualize gene expression changes in terms of expression values by plotting the log fold changes from differential expression (M) on the y-axis and the log average expression (A) on the x-axis. M-A plots below reveal that the population of differentially expressed mRNAs under-represented in the cytoplasmic fraction (i.e., negative log-fold difference on the y-axis) tend to favor lower average expression values (i.e., lower value on the x-axis) (**Figure 14**). This indicates that despite a larger population of mRNAs being down-regulated in the cytoplasmic fraction, such RNAs tend to be lowly expressed. This suggests the possibility of capturing RNAs subject to degradation in a transient state.

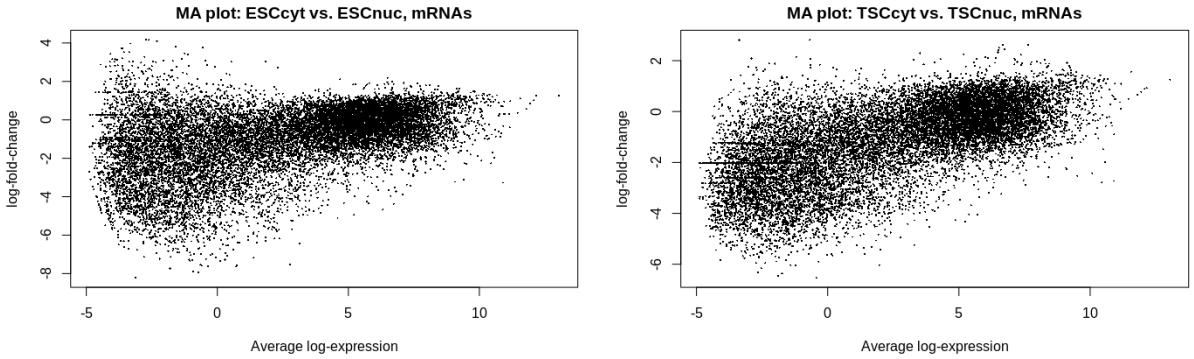


Figure 14: **M-A plots of mRNA differential expression profile using exon data**

Left: Down-regulated mRNAs in the cytoplasmic fraction in ESC data, despite its larger population, tends to be lowly expressed; **Right:** similar result is shown in TSC data.

Interestingly, an opposite trend is shown in the differential expression profile in ncRNA data. Using the identical parameters in limma as previously, the distribution of differentially expressed ncRNAs is heavily skewed towards the cytoplasmic fraction in both cells (**Figure 15, A & B**). This stark difference between mRNA and ncRNA data may hinge on their differences in post-transcriptional processing as well as in subcellular localization.

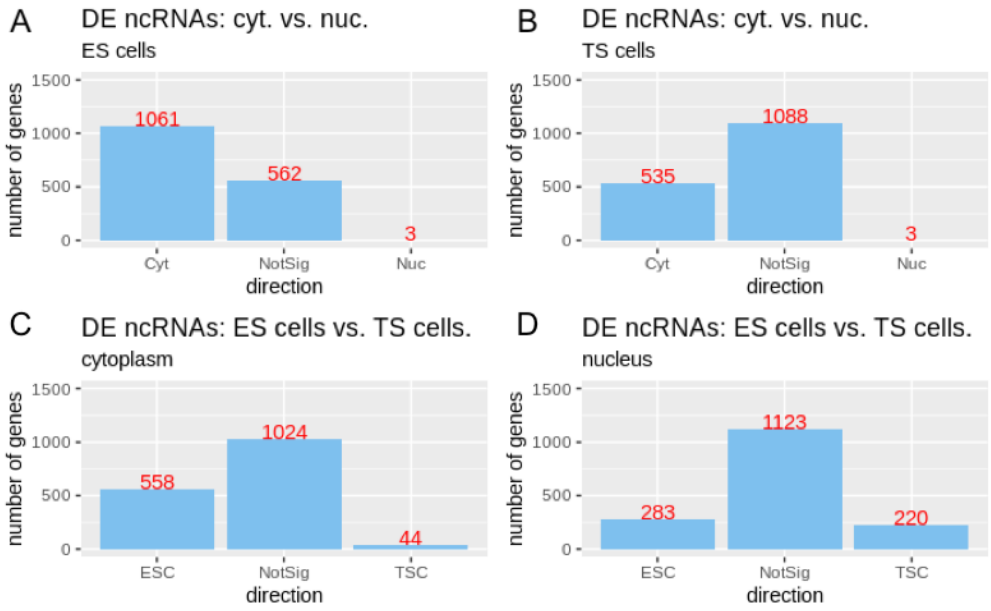


Figure 15: **Differential expression profile of ncRNAs using exon count data**

Differential expression using exon counts with limma output (adjusted p-value < 0.05) show **A)** the up-regulated population of ncRNAs in the cytoplasmic fraction is significantly larger than in the nuclear fraction in ESC data; **B)** similar trend is shown in TSC data; **C)** cell-to-cell comparisons show the up-regulated population of mRNAs in the cytoplasm in ESCs is larger than in TSCs; **D)** cell-to-cell comparisons in nuclear fractions show a more even number of up-regulated ncRNAs than in cytoplasmic data.

M-A plots on ncRNA expression profile reveals the up-regulated genes in the cytoplasmic fraction tend to show lower average expression at higher fold differences (**Figure 16**). This suggests, as before, that the significant portion of highly enriched genes (i.e., genes with the largest fold differences) may not be as biologically significant. To infer the biological significance of the population of differentially expressed genes, however, a functional gene set enrichment type of an analysis is necessary.

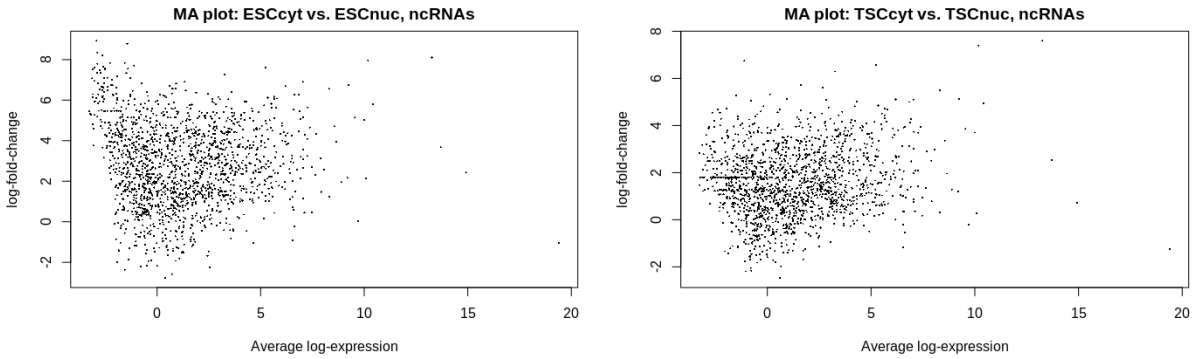


Figure 16: **M-A plots of ncRNA differential expression profile using exon data**

Left: Up-regulated mRNAs in the cytoplasmic fraction in ESC data, despite its larger population, show no particular bias towards high expression values in ESC data; **Right:** similar result is shown in TSC data.

A panel of lncRNAs was chosen to illustrate their subcellular localization behaviour; lncRNAs such as Neat1, Lncenc1, Meg3, and Bvht have been shown to influence regulation of pluripotency and self-renewal in ESCs at the nuclear level via chromatin interactions[98-100]. LncRNAs Gas5 and Snhg3 have been detected in both the cytoplasm and the nucleus in mouse ESCs to regulate the expression of Oct4, Nanog, and Sox2[101]. Finally, Malat1 has been shown to act as a competing endogenous RNA to regulate miRNA-mRNA interactions[102].

Figure 17 shows that in the differential expression profile using exon counts, lncRNAs with known nuclear functions show over-representation in the cytoplasmic fraction in both cell types. This result suggests lncRNAs do not necessarily reside in the nucleus despite associated nuclear function. As lncRNAs such as Bvht and Neat1 have been shown to regulate cell lineage commitment, it is possible lncRNAs shuttle between the cytoplasm and nucleus as a means of cell fate regulation. Interestingly, Lncenc1 shows up-regulation in the cytoplasmic fraction as well, despite literature evidence in its involvement in the nucleus to maintain cell self-renewal and key metabolic processes.

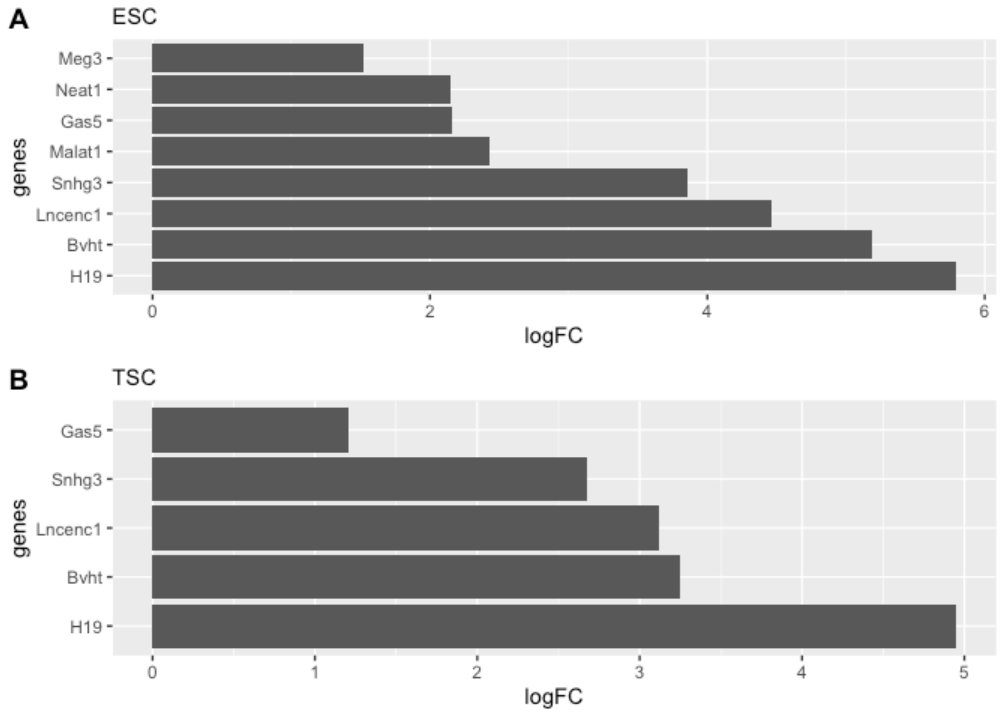


Figure 17: **Differential expression profile of a panel of lncRNAs with known involvement in pluripotency and differentiation**

Differential expression result shows cytoplasmic up-regulation of lncRNAs with both known cytoplasmic and nuclear function; this result suggests lncRNAs may shuttle between the two compartments as a means of gene regulation during cell fate commitment.

3.2.5 Functional gene set enrichments using exon expression profile

Functional gene set enrichment analysis is useful in interpreting large lists of differentially expressed genes into biologically meaningful groups. Such analysis is implemented in current study to understand whether patterns of subcellular localization is related to RNA function. Grouping related gene sets together by similarity semantics (via REVIGO) or by gene overlap (via clusterProfiler, as used in **Section 3.1.11** in network analysis) allows for inferences in which cellular processes predominate in each sample. Therefore, gene set enrichment coupled to visualization using a grouping method serves as a logical next step from differential expression profiling to uncover functional significance.

Gene set enrichment analysis result with mRNA and ncRNA data from **Section 3.2.4** show some differences in ESC and TSC data. Firstly, in mRNA data, genes related to translation and macromolecular complex assembly are predominantly enriched in the cytoplasmic fraction in

ESCs (**Figure 18**, **top**). In TSCs, instead of gene sets related to translation, genes related to metabolic function predominate (**Figure 19**, **top**). In the nuclear fractions, gene sets associated with cell division and RNA processing are most enriched in ESCs (**Figure 18, bottom**), whereas in TSCs genes related to cell structure and response to stress are most enriched (**Figure 19, bottom**). These results are summarized in **Table 3**.

Assigning functional annotations to the differentially expressed profile from **Section 3.2.4** yields a more meaningful context to the overall distribution. The up-regulated genes in the cytoplasmic fraction of ESCs are lower in overall number compared to the down-regulated genes (**Figure 13**). This suggested that the turnover rate of mRNAs localized to the cytoplasm (versus mRNAs localized to the nucleus for export) had a strong influence in shaping the uneven distribution. mRNAs that still appear up-regulated in the cytoplasm, however, seem to have functions related to the mRNA turnover itself (i.e., processes related to translation and ribosomal assembly), as well as functions related to RNA and DNA processing. In the nuclear fraction, it is interesting to note that despite a significantly larger population of mRNAs, mRNAs related to nuclear processes are the most predominant. This suggests that even though it is a possibility that the overall number of up-regulated genes in the nucleus (i.e., down-regulated in the cytoplasm) is inflated by either degraded RNA or incompletely processed and exported RNA, in an over-representation type of an analysis, the most frequently represented gene sets are related to nuclear functions. This observation is interesting as this may suggest mRNAs related to cell division are either turned over rapidly in the cytoplasm or held in the nucleus awaiting for export. In TSC data, such nuclear mRNAs include genes related to cell structure organization and response to stimuli, which indicate the possibility of such mRNAs being retained in the nucleus prior to export. The concept of transcripts retained in the nucleus awaiting for cellular signal is further addressed in the discussion of intron retention (**Sections 1.4.2** and **3.3.10**).

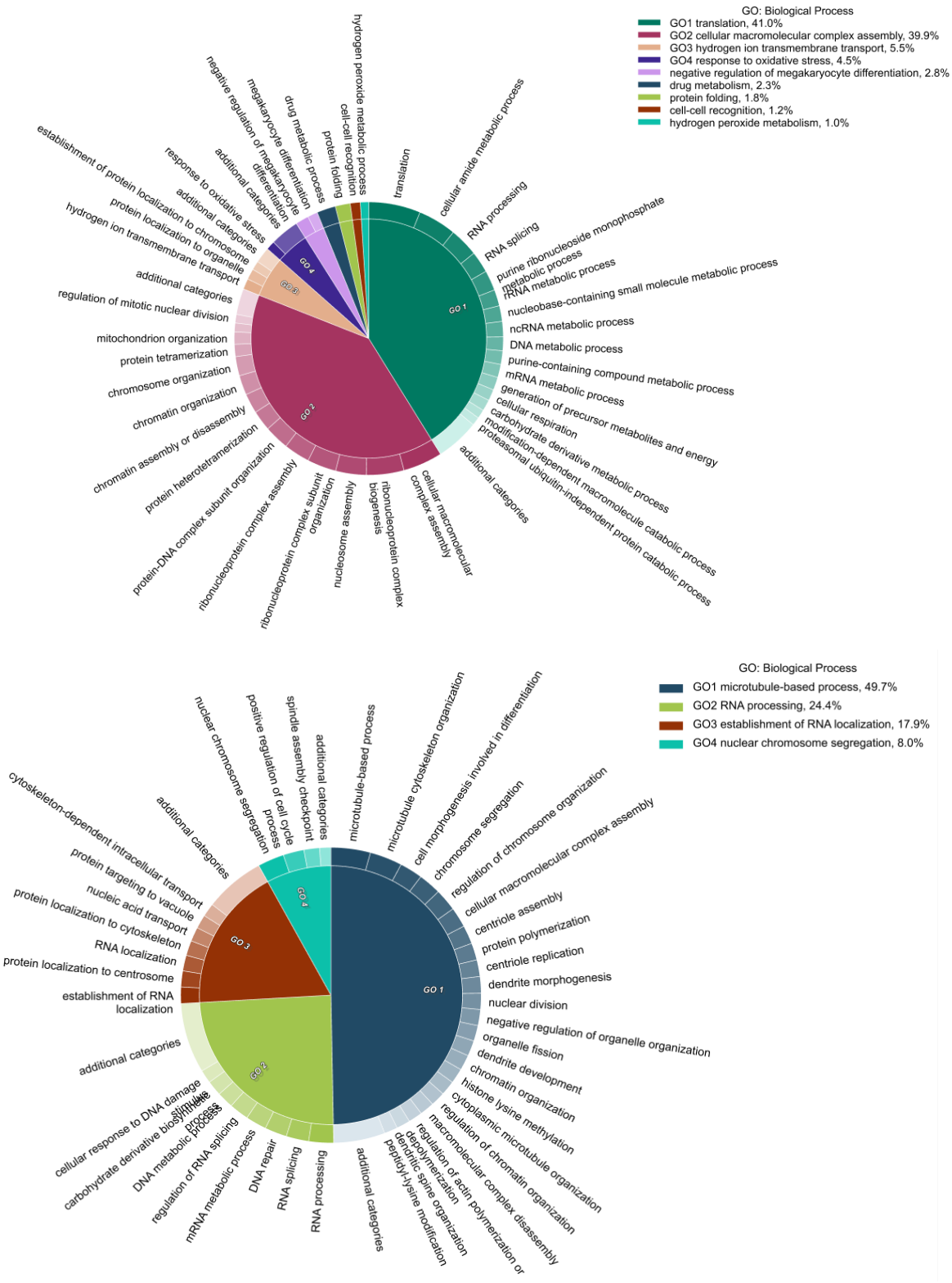


Figure 18: **GO enrichment result in ESC cytoplasmic-nuclear mRNAs using exon counts**

Top: GO terms related to translation and ribosomal processes are enriched in the cytoplasmic fraction in ESCs; **Bottom:** GO terms related to cell division and RNA processing are enriched in the nuclear fraction.

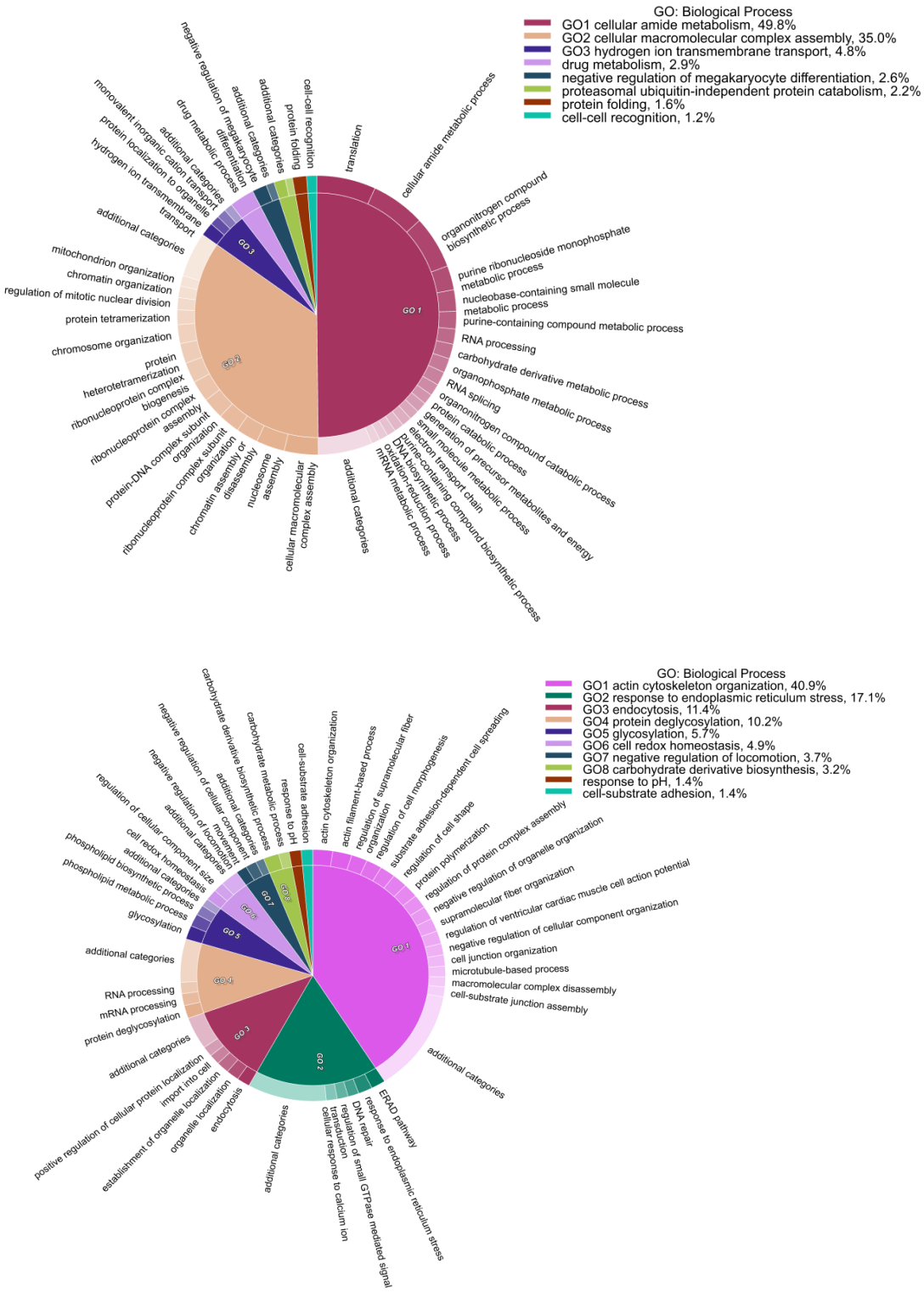


Figure 19: **GO enrichment result in TSC cytoplasmic-nuclear mRNAs using exon counts**

Top: GO terms related to metabolic and ribosomal processes are enriched in the cytoplasmic fraction in TSCs; **Bottom:** GO terms related to cell division and cell structure are enriched in the nuclear fraction.

Table 3: **Summary of functional gene set enrichment results in mRNA data**

Class	Cytoplasm (vs. nucleus)	Nucleus (vs. cytoplasm)
ESC mRNAs	• Significantly smaller population of differentially expressed counts (**Figure 13)** • Statistically significant counts tend to show lower log fold difference • Counts with relatively higher fold difference tend to show lower average log expression (**Figure 14**) • Top enriched gene sets are associated with macromolecule metabolism and ribosomal assembly (**Figure 18, top**) • Distinct clusters of ontological sets: ion transport, metabolism, RNA processing, cellular respiration, and macromolecular	• Over 3x larger population of differentially expressed counts (**Figure 13**) • Shows a wide spread of significant genes in terms of log fold difference • Shows an inverse trend between average expression and fold change (**Figure 14**) • Top enriched gene sets are related to cell division (**Figure 18, bottom**) • Distinct clusters of ontological sets: RNA processing, microtubule assembly, and perception of stimuli (**Figure 18, bottom)**

	assembly (**Figure 18, top**)	
TSC mRNAs	• Similar to ESC data, significantly smaller population of differentially expressed counts (**Figure 13**) • As with ESC data, statistically significant counts tend to show lower log fold difference • Top enriched gene sets are related to metabolism (**Figure 19, top**)	• Similar to ESC data, population of differentially expressed counts is over 3x larger (**Figure 13**) • Shows a wide spread of significant genes in terms of log fold difference • As with ESC data, shows an inverse trend between average expression and fold change (**Figure 19, bottom**) • Top enriched gene sets are related to cell structure organization and regulation of cellular processes (**Figure 19, bottom**)

In ncRNA data, gene sets in both cytoplasmic and nuclear fractions tend to be related to cell response to stimuli or a regulatory process (**Figures 20 & 21**). As the ncRNA database used in the analysis includes a wide array of RNA species such as snoRNAs, miRNAs, and lncRNAs as well as pre-ribosomal RNAs, it is difficult to generalize their localization and mode of action. In downstream analysis, considerations such as intronic content and splicing ratios will be taken into account to provide further insight into ncRNA behavior.

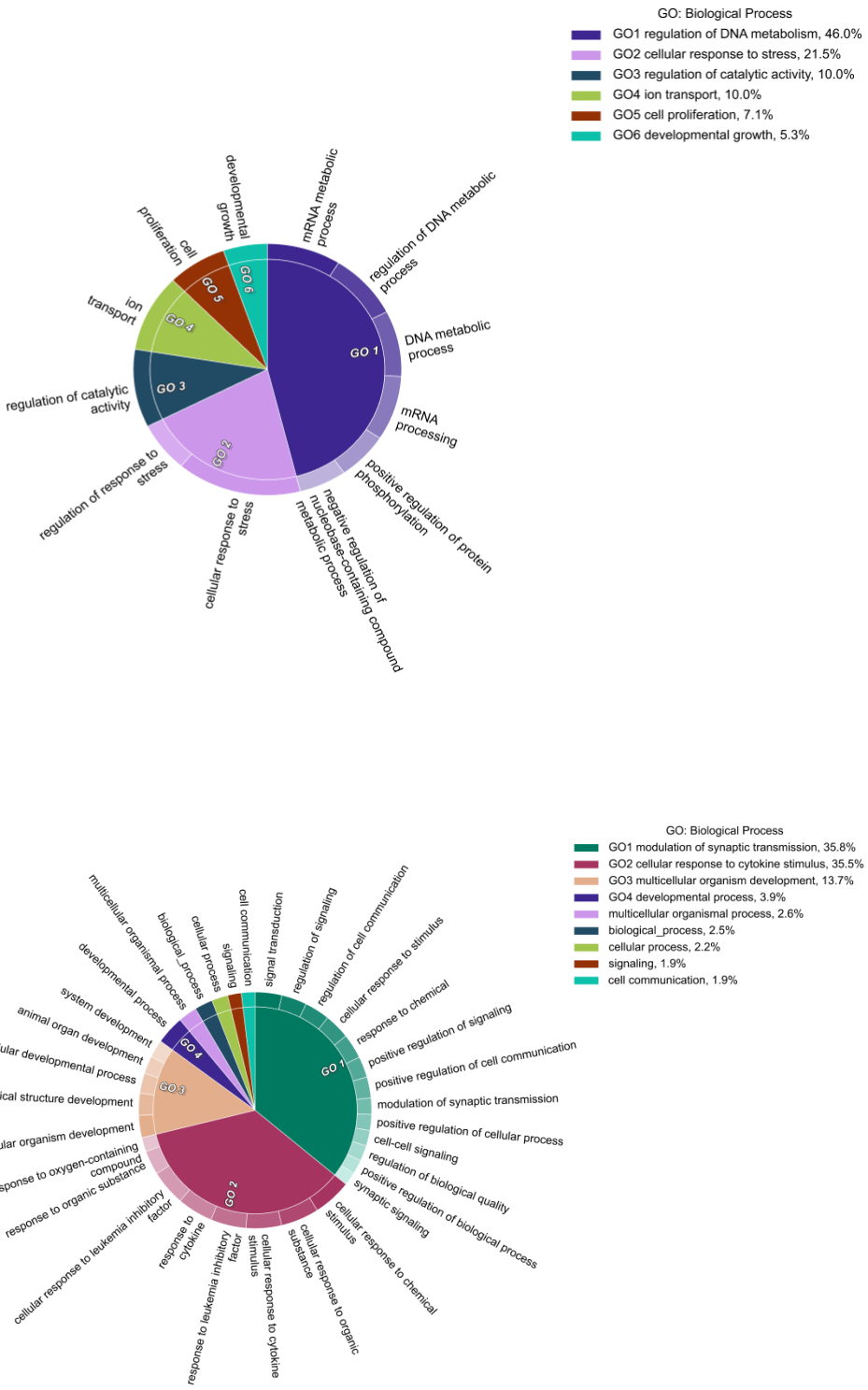


Figure 20: **GO enrichment result in ESC cytoplasmic-nuclear ncRNAs using exon counts**

Top: GO terms related to RNA and DNA processing are enriched in the cytoplasmic fraction in ESCs; **Bottom:** GO terms related to cell response to stimuli and cell signaling are enriched in the nuclear fraction.

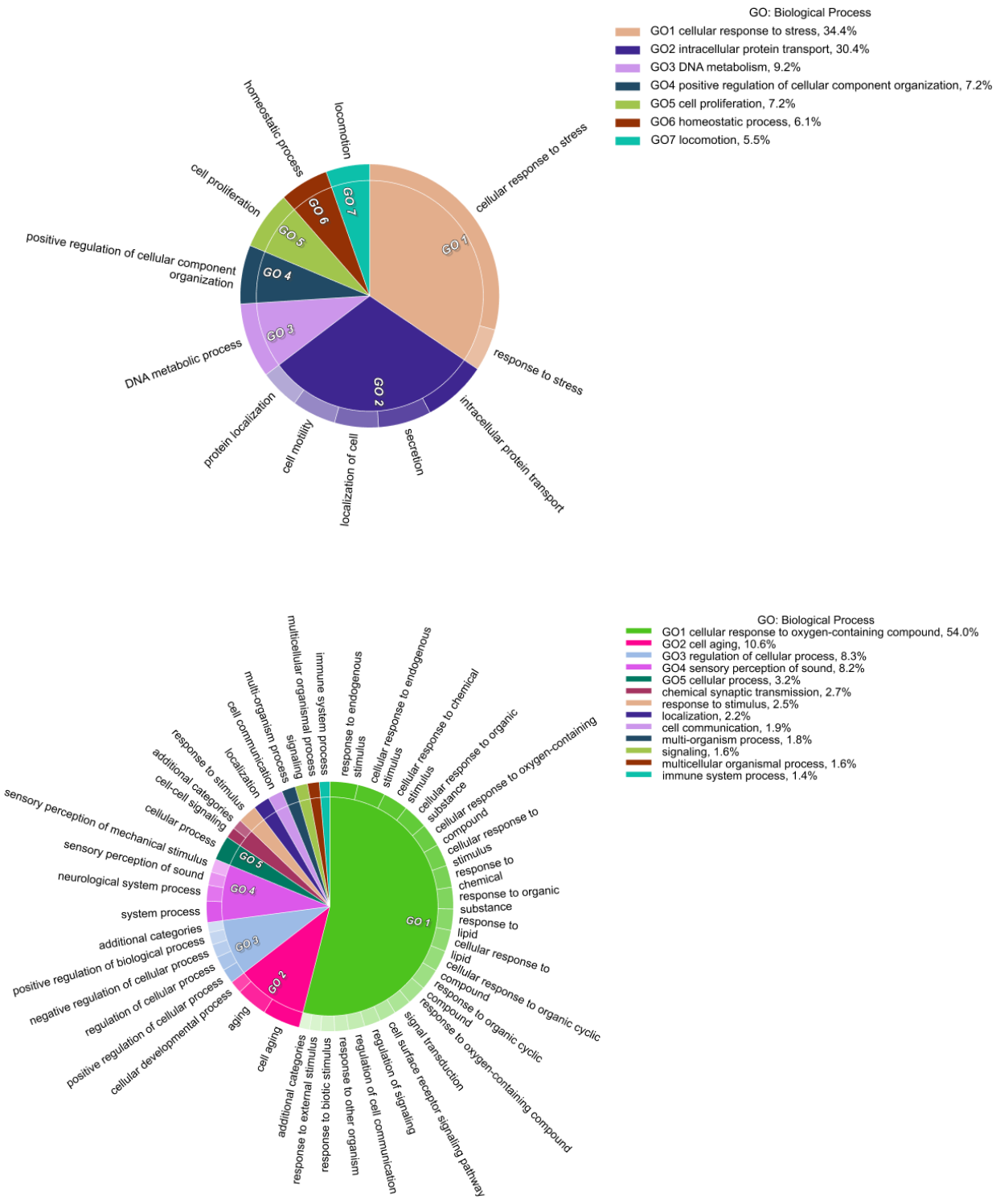


Figure 21: **GO enrichment result in TSC cytoplasmic-nuclear ncRNAs using exon counts**

Top: GO terms related to cell response are enriched in the cytoplasmic fraction in ESCs; **Bottom:** GO terms related to cell response to chemical stimuli and cell signaling are enriched in the nuclear fraction.

Table 4: **Summary of functional gene set enrichment results in ncRNA data**

Class	Cytoplasm (vs. nucleus)	Nucleus (vs. cytoplasm)
ESC ncRNAs	• Significantly larger population of differentially expressed counts (**Figure 15**; opposite trend from mRNA data) • Top enriched gene sets are associated with RNA processing, regulation of cellular processes, and metabolism of RNA/DNAs (**Figure 20, top**)	• Almost non-existent population of differentially expressed counts (**Figure 15;** opposite trend from mRNA data) • Shows an abundance of enrichment of gene sets related to cellular response and regulation of cellular processes (**Figure 20, bottom**)
TSC ncRNAs	• Significantly larger population of differentially expressed counts (**Figure 15**; opposite trend from mRNA data) • Top enriched gene sets include cell movement, proliferation, and	• Almost non-existent population of differentially expressed counts (**Figure 15;** opposite trend from mRNA data) • Two distinct enriched gene set clusters related to regulation of cellular processes and cell response to

	response to stress **(Figures 21, top)**	endogenous stimulus **(Figure 21, bottom)**

3.2.6 Identification of TSC lineage genes and cell adhesion molecules

Interestingly, the nuclear fraction in ESC and TSC data show deviating results in functional enrichment; whereas gene sets related to cell division and cell cycle are accumulated in the nuclear fraction in ESCs, genes related to cell adhesion and structure are enriched in the nuclear fraction in TSCs. In order to visualize the differential expression of cell adhesion molecules (CAMs) in particular, lcpm values of cadherins and Epcam, as well as genes related to TSC lineage are shown in a heatmap (**Figure 22**). The heatmap shows overall a higher expression of mRNAs coding for CAMs in the nuclear fraction compared to the cytoplasmic fraction, whereas mRNAs related to maintenance of TSC identity – namely Cdx2, Elf5, Sox21, and Eomes – show little difference in subcellular localization. The expression of cadherins as well as Epcam have been documented in literature to be involved in maintenance of multipotency within the trophoblast[103,104].

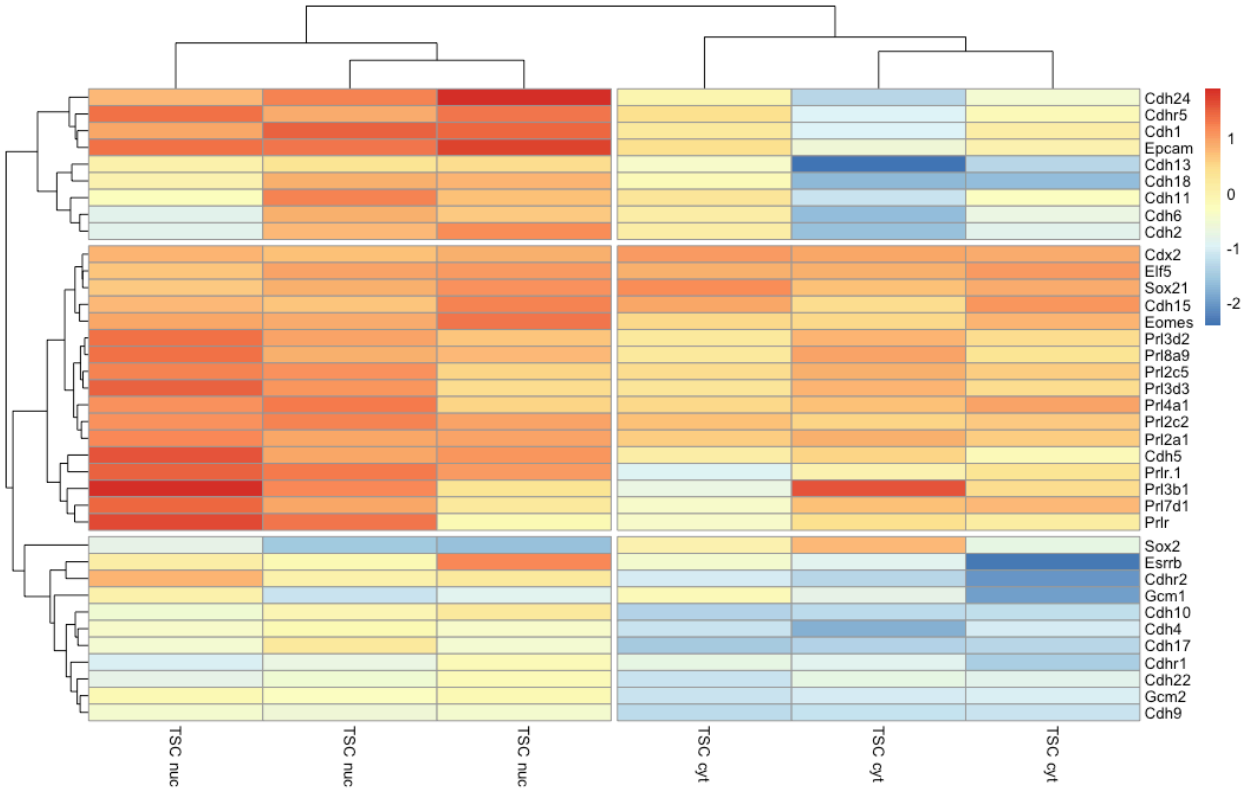


Figure 22: **Differential expression of mRNAs associated with cell adhesion and TSC lineage**

A heatmap of differential expression result shows nuclear accumulation of mRNAs associated with cell adhesion – cadherins and Epcam. Both cadherins and Epcam have been characterized to be involved in TSC lineage for maintenance of multipotency. RNAs coding for transcription factors involved in maintenance of TSC identity – Cdx2, Sox21, Elf5, and Eomes – show little difference in subcellular localization.

3.2.7 Differential expression profile using intron counts

Using custom made intron annotations, intron count data was generated and used as input for edgeR-voom-limma pipeline. This parallel analysis was done in order to better understand the context behind asymmetrical distribution of differentially expressed genes across the two fractions. As mRNAs canonically undergo splicing prior to nuclear export, one would expect cytoplasmic mRNAs to be intron-free. However, as studies in alternative splicing and in particular, intron retention would suggest, processed mRNAs can contain introns in the cytoplasm. The fate of these intron-retaining mRNAs is also multi-faceted, where some intron-retaining mRNAs can even be retained in the nucleus. Therefore, by comparing the distribution of exon and intron expression across the subcellular fractions, one can infer the splicing behavior of the RNA population in each compartment. Furthermore, as introns often harbor ncRNAs with regulatory functions, investigating the level of gene enrichment in the context of introns will provide additional insights. As more and more literature findings suggest that large numbers of intronic RNAs are expressed in comparable numbers to exonic RNA counterparts, it is important to not discard intronic RNAs as simply pre-mRNAs or excised introns destined for degradation.

The overall distribution of differentially expressed genes show deviations from exon data. In both ESC and TSC mRNA data, the distribution of up- and down-regulated genes in the cytoplasmic compared to the nuclear fraction is at a difference of less than 100 genes (**Figure 23, A & B)**. This shift in distribution is in stark contrast with the results from exon data, where the population of cytoplasmic down-regulated genes was significantly larger (**Figure 13**). This suggests the possibility that there may be nuclear detected mRNAs lacking introns – since if exons and introns were co-enriched in the nucleus, their distributions would share resemblance. The presence of up-regulated mRNAs with introns also indicate a degree of intron retention in the cytoplasm. The nuclear introns may be attributed to detained intron-retaining mRNAs, pre-mRNAs, or excised intronic elements captured before degradation.

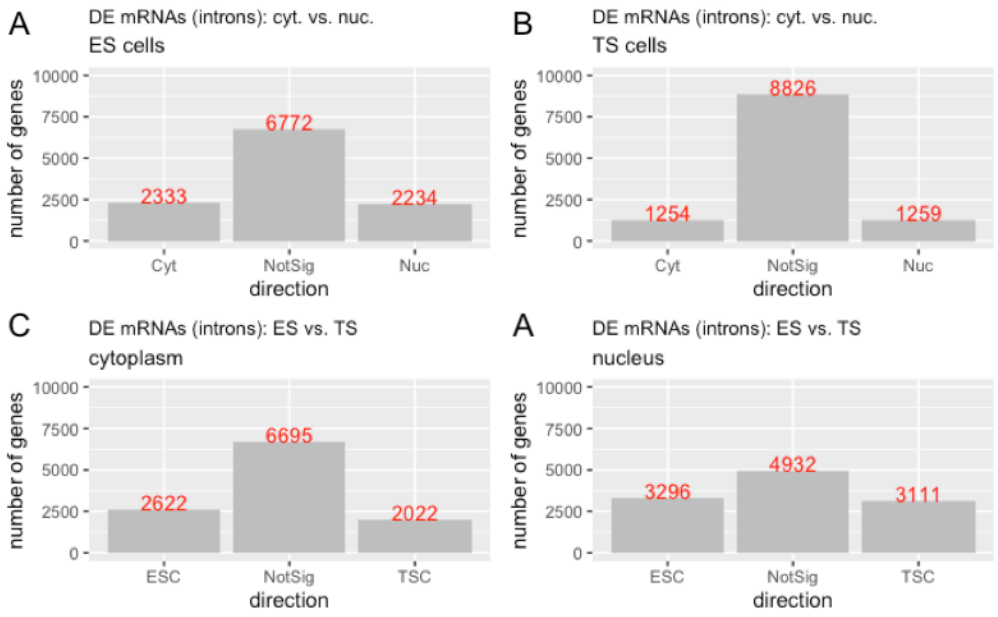


Figure 23: **Differential expression profile of mRNAs using intron count data**

Differential expression using intron counts with limma (adjusted p-value < 0.05) show relatively even number of up-regulated mRNAs in either direction in **A)** cytoplasmic-nuclear comparison in ESCs, **B)** cytoplasmic-nuclear comparison in TSCs, **C)** cell-to-cell comparisons in cytoplasmic data; **D)** cell-to-cell comparisons in nuclear data.

M-A plots reveal some symmetry in the distribution profile with respect to levels of expression; it appears that up-regulated intron-containing mRNAs in the cytoplasmic fraction exhibit lower average expression compared to down-regulated mRNAs (**Figure 24**). It is possible that this difference is related to a difference in function of intron-retaining transcripts in each compartment.

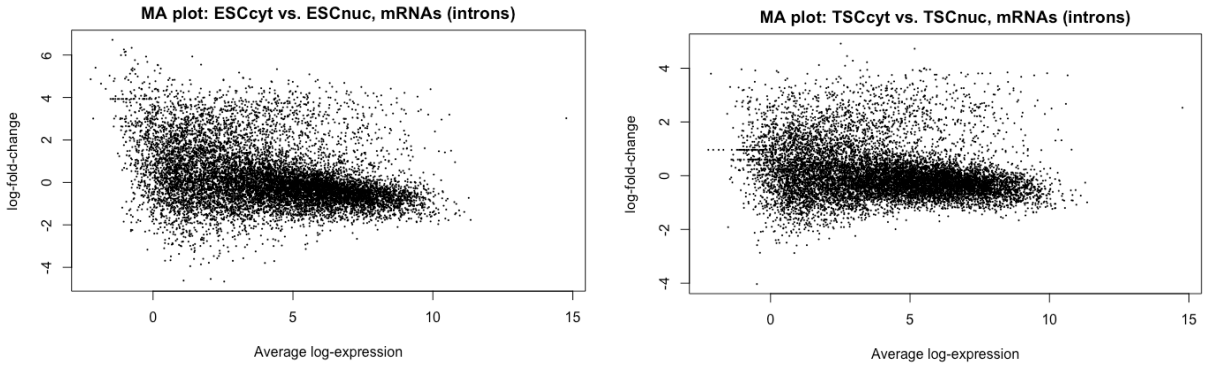


Figure 24: **M-A plots of mRNA data using intron counts**

Under visual inspection, despite the similar number of mRNAs up-regulated towards either cytoplasmic and nuclear fraction, asymmetry in fold-change with respect to average expression is shown in **A)** ESC data as well as in **B)** TSC data.

In ncRNA data, the differential expression profile distribution across the two fractions shows opposite behaviour compared to when exon counts were used instead; the population of down-regulated ncRNAs in the cytoplasm is considerably larger than the population of up-regulated ncRNAs (**Figure 25**). This suggests that in terms of ncRNAs with retained introns, the localization behaviour tends to favor nuclear enrichment – whereas when only considering exons, the opposite trend is observed. This may be attributed to different classes of ncRNA species exhibiting differential splicing and localization behaviour. A per-gene analysis of exon-intron proportions would assist in identifying this differential splicing and intron-retaining behaviour (**Section 3.2.10**).

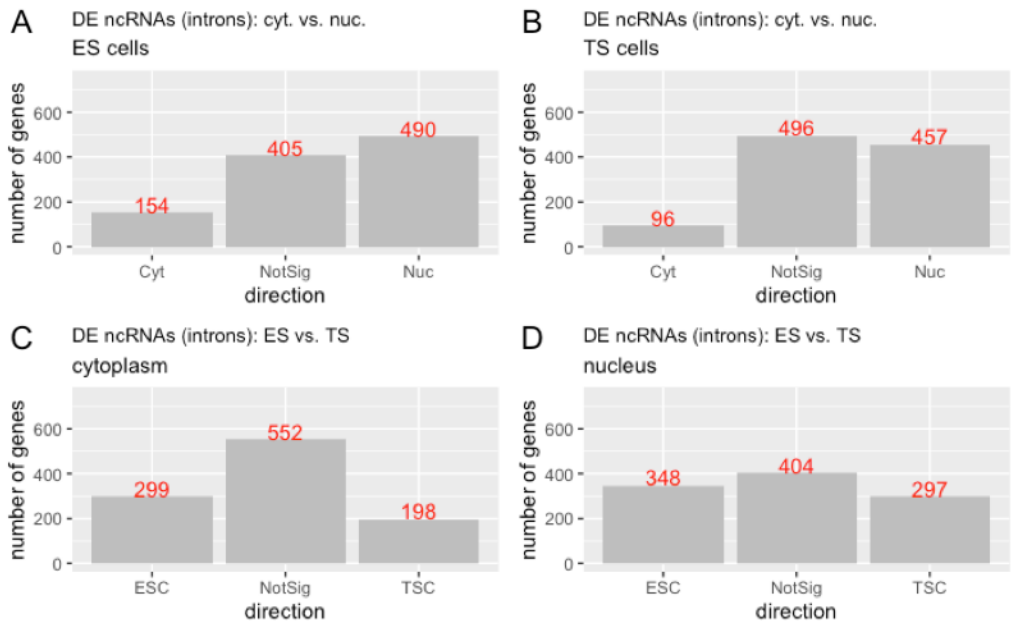


Figure 25: **Differential expression profile of ncRNAs using intron count data**

Differential expression using intron counts with limma (adjusted p-value < 0.05) show **A)** larger population of down-regulated ncRNAs in the cytoplasmic fraction relative to the nuclear fraction in ESC data; **B)** similar result is shown in TSC data; **C)** in comparison of ESC and TSC data in the cytoplasmic fraction, the up-regulated ncRNA population in ESCs is larger than in TSCs; **D)** similar result is shown in nuclear fraction data.

The same panel of lncRNAs from **Section 3.2.4** (i.e., Lncenc1, Neat1, Evx1as, Pnky, Tuna, Malat1, Xist, Meg3, Bvht, Gas5, Snhg3, and H19) was selected from the differential expression profile to examine their subcellular localization using intron counts. **Figure 26** shows only Xist, Meg3, and

Gas5 were shown to be differentially expressed in ESCs across the subcellular fractions, wherein all three lncRNAs were found up-regulated in the nuclear fraction. Snhg3 shows similar trend in TSCs, but does not pass the log-fold-change threshold. Interestingly, lncRNA Xist was not found to be differentially expressed across the two fractions when exon counts were used instead (**Figure 17**).

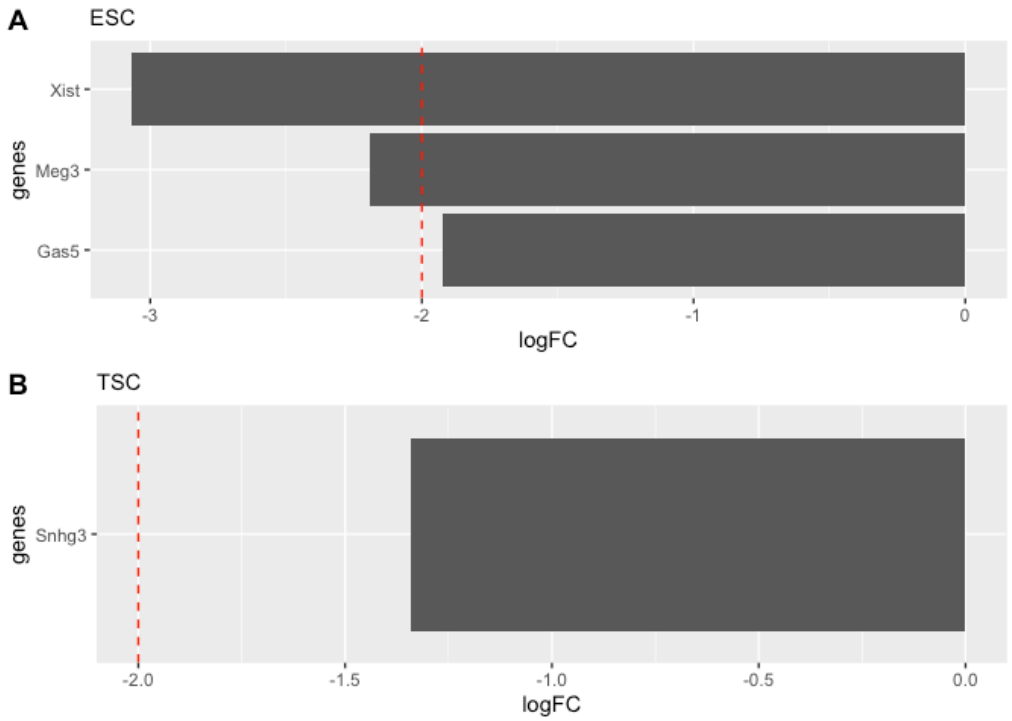


Figure 26: **Differential expression profile using intron counts of a panel of lncRNAs with known involvement in pluripotency and differentiation**

Differential expression using intron counts show fewer number of pluripotency related lncRNAs differentially expressed across the two subcellular fractions; all differentially expressed genes in the panel are up-regulated in the nuclear fraction, which is in contrast to when using exon count data

This result, in conjunction with the overall differential expression profile (**Figure 25**) suggest lncRNAs which retain introns are more likely to be localized to the nucleus. Furthermore, this suggests the lncRNAs involved in maintenance of pluripotency (i.e., Snhg3, Lncenc1, Gas5) as well as cell fate commitment (i.e., Bvht, Neat1, Meg3) may show similar splicing behaviour as mRNAs – such that after splicing they are exported into the cytoplasm. This behaviour leads to

up-regulation of lncRNA in the cytoplasmic fraction when using exon count data (**Figure 17**) and the reverse result or no differential expression when using intron count data (**Figure 25**).

3.2.8 *Functional gene set enrichment analysis using intron expression profile*

Functional gene set enrichment analysis using intron counts allows for identification of biological processes associated with intron retention. In ESC mRNA data, gene ontologies over-represented in up-regulated genes in the cytoplasmic fraction tend to be related to nuclear processes such as chromosome organization and processes associated with the cell cycle (**Figure 27, top**). For genes down-regulated in the cytoplasm versus the nucleus, processes related to cell response to stimulus and signal transduction were found to be enriched (**Figure 27, bottom**). In TSCs, gene sets related to mRNA metabolism and regulation of transcription were up-regulated in the cytoplasmic fraction (**Figure 28, top**). Similar to ESCs, gene sets related to cell response to stimulus and signal transduction were over-represented in genes down-regulated in the cytoplasmic fraction (**Figure 28, bottom**).

Overall, gene sets associated with regulatory function are more represented compared to when using exon count data. This suggests that genes whose introns are retained within the transcript and are differentially expressed across the cytoplasmic-nuclear boundary tends to show regulatory behaviour. Similar to exon data, however, over-represented gene sets in the nuclear fraction are related to cell response, indicating the possibility that these may represent nuclear retained mRNAs awaiting for a particular stimuli or a signal. It is possible that these mRNAs may show localization to the cytoplasm in a cell response feedback system. Gene sets related to protein signal transduction are enriched in the nuclear fractions as well, indicating the possibility of a relationship between intron-containing nuclear detained mRNAs and cell signaling.

Interestingly, in ncRNA data, topGO fails to return multiple enriched gene sets for ncRNAs up-regulated in the nuclear fraction of both ESC and TSC data. In the current methodology using K-S statistics to identify enriched gene sets, only one gene set (GO:0060255; regulation of macromolecule metabolic process) is found to be enriched in ESC data (KS = 0.029); no gene set was identified as enriched in TSC data.

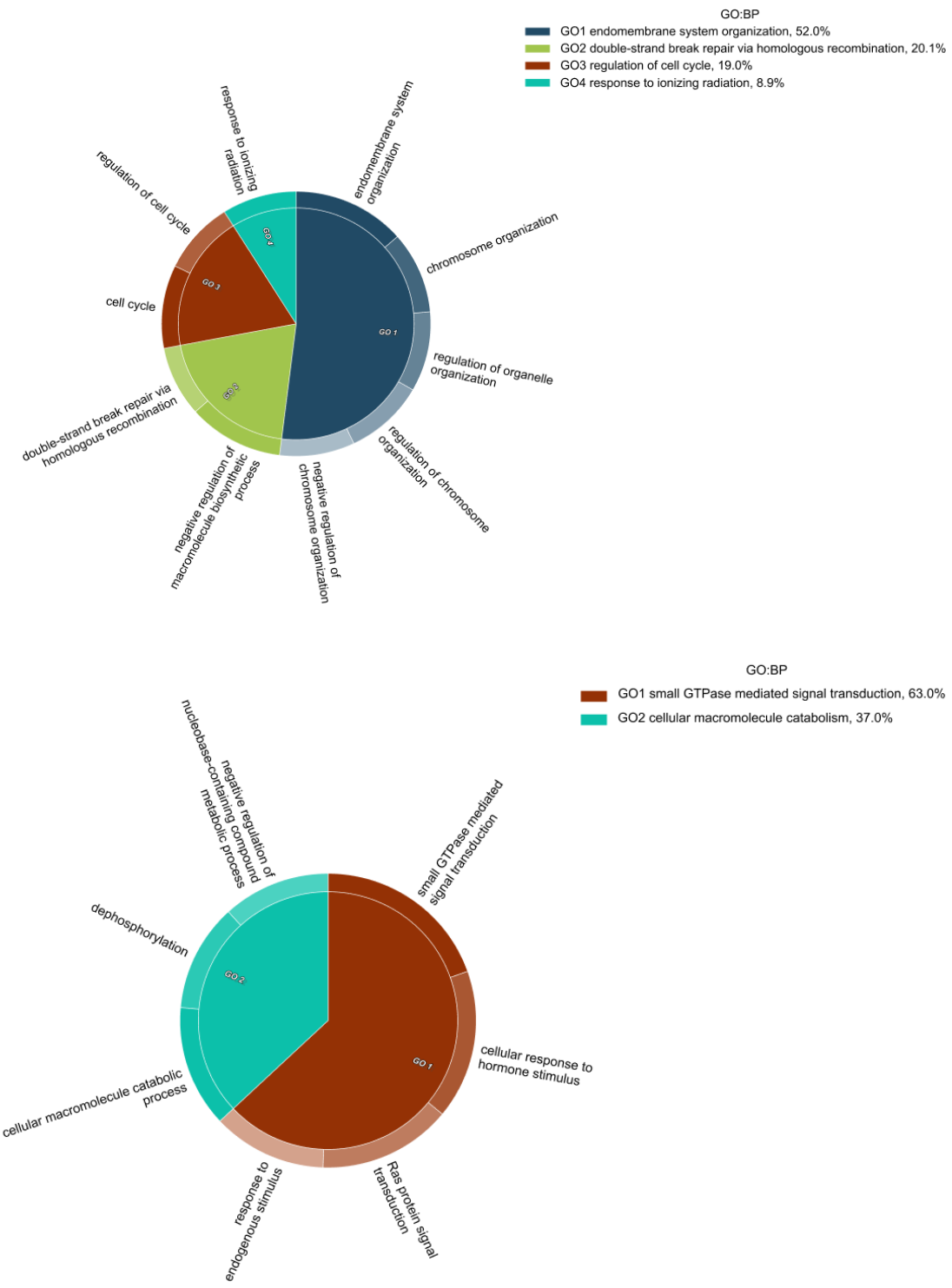


Figure 27: **GO enrichment result in ESC cytoplasmic-nuclear mRNAs using intron counts**

Top: GO terms enriched in the cytoplasmic fraction in ESCs include chromosome and cell organization; **Bottom:** GO terms enriched in the nuclear fraction are related to cell signaling and response to stimuli.

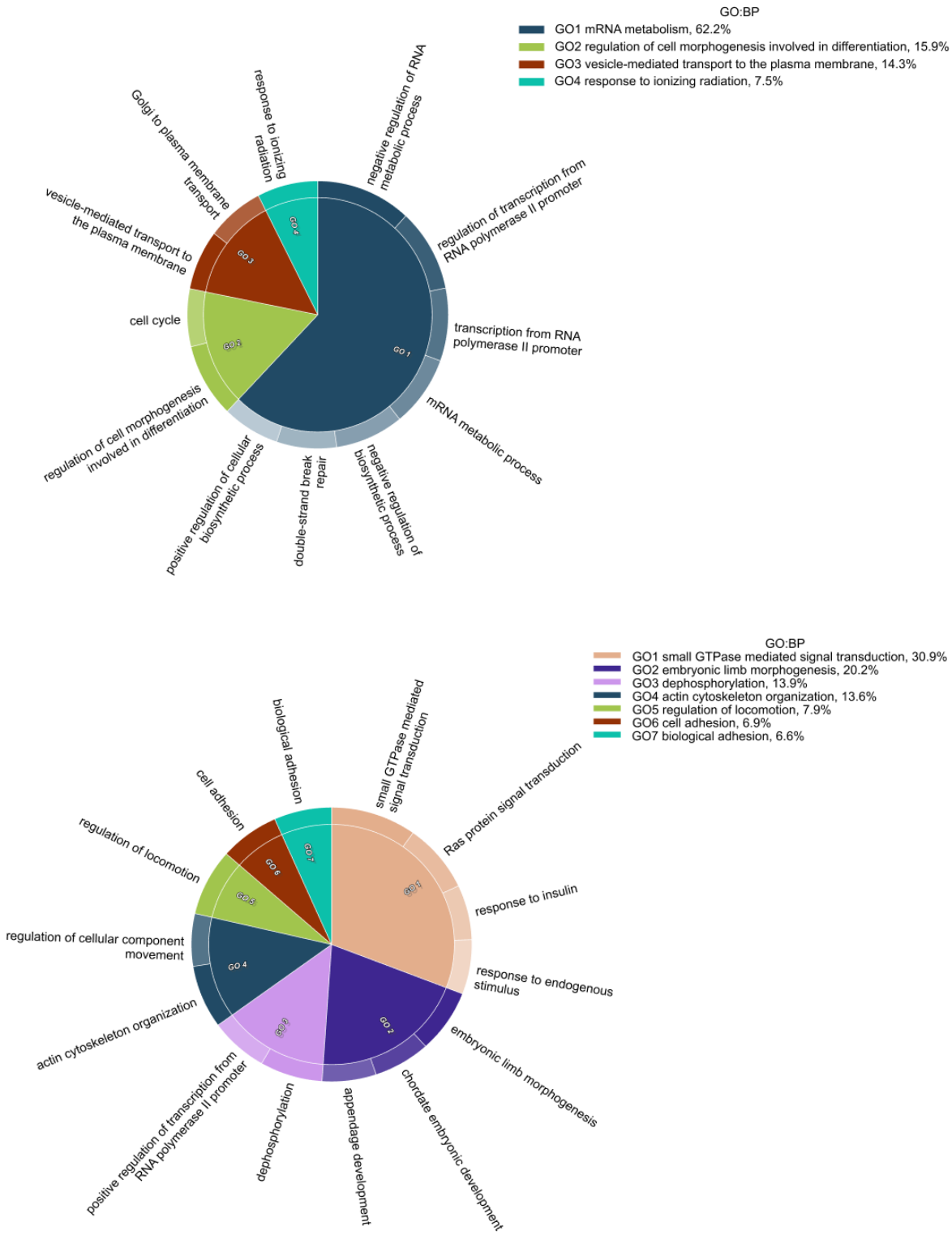


Figure 28: **GO enrichment result in TSC cytoplasmic-nuclear mRNAs using intron counts**

Top: GO terms enriched in the cytoplasmic fraction in ESCs include RNA processing and transcription; **Bottom:** GO terms enriched in the nuclear fraction are related to cell signaling and response to stimuli.

Table 5: **Summary of mRNA gene ontology enrichment results using intron data**

Class	Cytoplasm (vs. nucleus)	Nucleus (vs. cytoplasm)
ESC mRNAs	• Relatively even number of differentially expressed counts versus nuclear fraction • Top enriched terms predominantly related to negative regulatory processes • Top enriched terms include terms related to cell organization, as well as regulation of gene expression	• Relatively even number of differentially expressed counts versus cytoplasmic fraction • Statistically significant counts tend to be concentrated at lower log fold changes • Top enriched terms predominantly related to cellular response to chemical stimuli
TSC mRNAs	• Fairly even number of enriched read counts as nuclear fraction; population of differentially expressed reads in both fractions are relatively low • Wide spread of statistically significant counts across higher	• Fairly even number of enriched read counts as cytoplasmic fraction; population of differentially expressed reads in both fractions are relatively low • Statistically significant counts tend to be concentrated at

	log fold changes (similar to ESC data) • Similar to ESC data, top enriched terms are predominantly negative regulatory terms • Enriched terms include negative regulation of gene expression, as well as regulation of transcription	lower log fold changes • Top enriched gene ontology terms include morphogenesis of appendages, as well as terms related to cell structure and movement • Top terms also include response to cellular signal

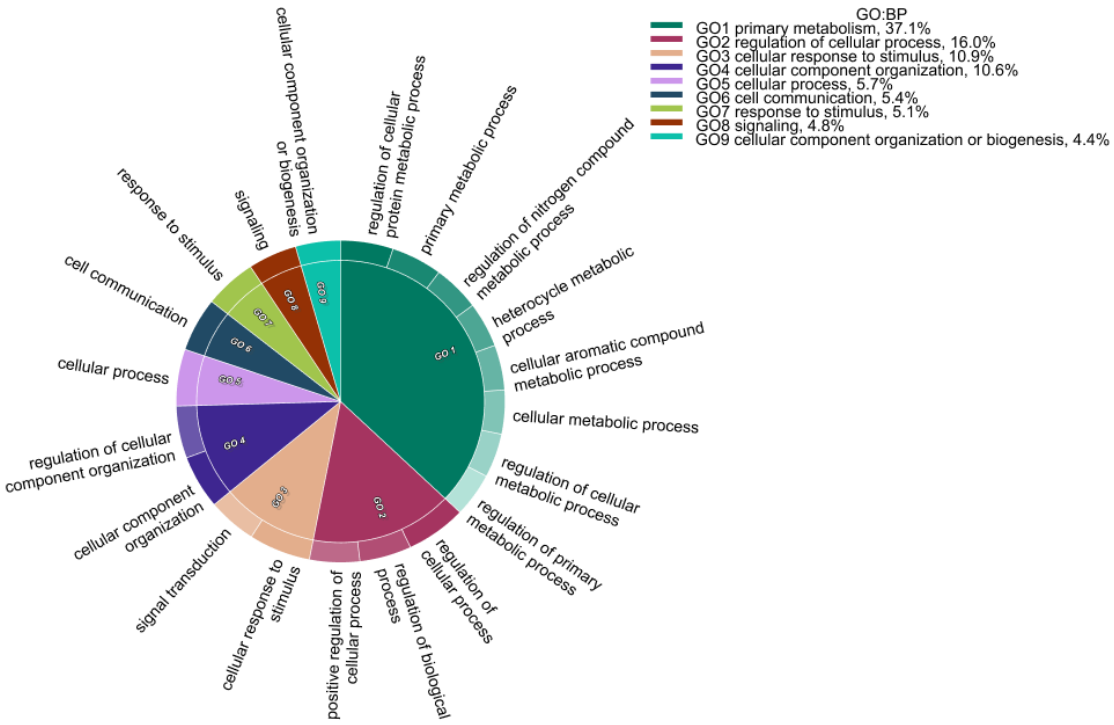


Figure 29: **GO enrichment result in up-regulated ncRNAs in ESC cytoplasmic fraction using intron counts**

Gene sets associated with up-regulated ncRNAs in the cytoplasmic fraction show an enrichment of metabolic processes.

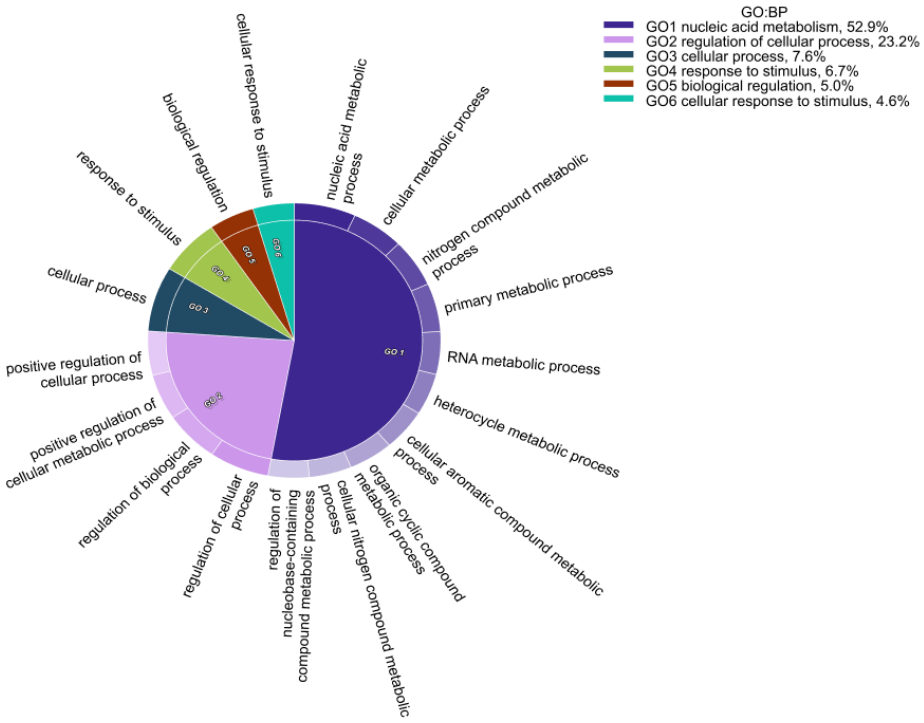


Figure 30: **GO enrichment result in up-regulated ncRNAs in ESC cytoplasmic fraction using intron counts**

Similar to ESC data, enriched gene sets in cytoplasmic ncRNA data in TSCs are predominately related to metabolic processes.

Table 6: **Summary of ncRNA gene ontology enrichment results using intron data**

Class	Cytoplasm (vs. nucleus)	Nucleus (vs. cytoplasm)
ESC ncRNAs	• Significantly lower number of enriched counts versus the nucleus • Statistically significant counts tend to be at lower logFC • Top enriched gene ontologies include terms related to metabolic processes • Enriched terms include regulatory terms related to cellular components	• Significantly higher number of enriched counts versus the cytoplasm • Wide spread of statistically significant counts, with increasing number at higher logFC
TSC ncRNAs	• As per ESC data, significantly lower number of enriched counts versus the nucleus • Statistically significant counts tend to be at lower logFC • Top enriched gene ontology terms are predominantly related to metabolic processes	• As per ESC data, significantly higher number of enriched counts versus the cytoplasm • Wide spread of statistically significant counts, with increasing number at higher logFC

3.2.9 Relationship between intron length and number of exons with differential expression

A scatterplot of log-fold-change of up-regulated mRNAs in the cytoplasmic fraction versus total intron length per gene shows a moderate negative correlation in both ESCs and TSCs ($\rho = -0.26$, $p < 2.2 \times 10^{-16}$) (**Figure 31**). This suggests that transcripts with longer introns tend to be less up-regulated in the cytoplasmic fraction compared to the nuclear fraction. This result is in agreement with findings from literature which suggest a negative correlation between total intron length and the kinetic rate of transcription[93]. A possible consequence of lower rate of transcription may manifest as lower log-fold-change values in the current fractionated differential expression analysis. Interestingly, the total intron lengths did not show a correlation with the log-fold-change of down-regulated genes in the cytoplasm. This suggests that intron lengths have an influence on the extent of cytoplasmic localization, but not necessarily localization to the nucleus.

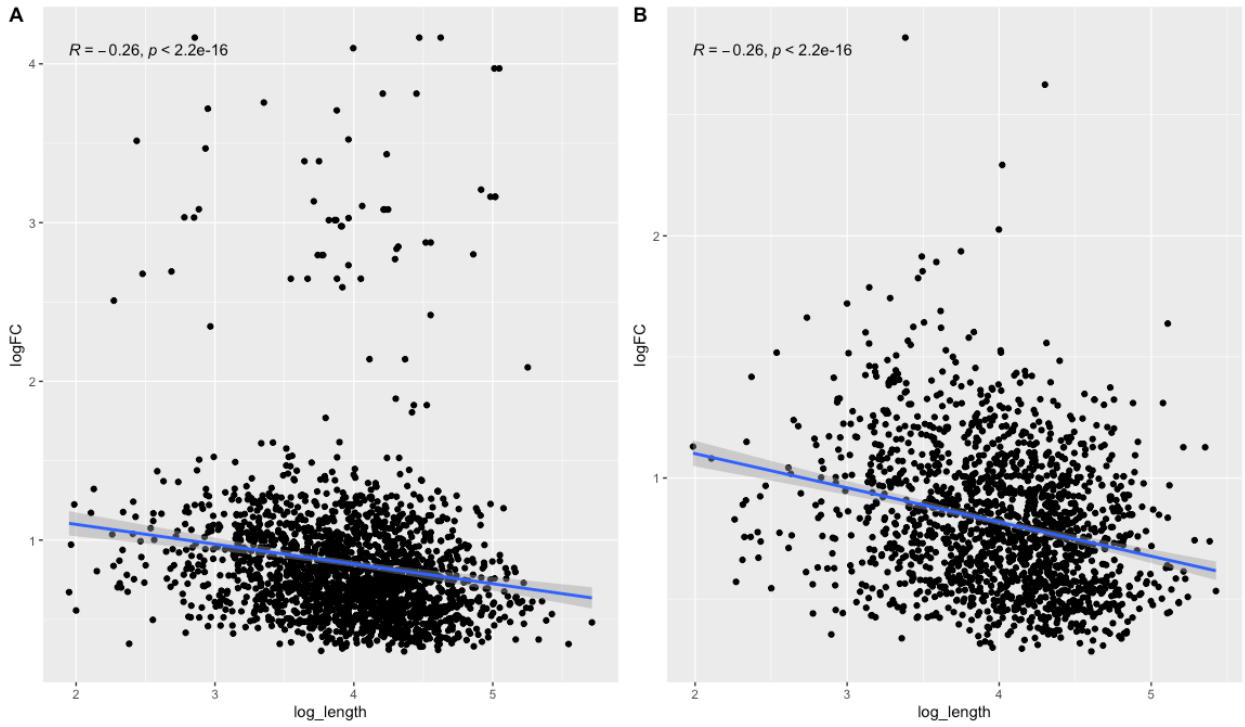


Figure 31: **Correlation between intron length and log-fold-change in differential expression**

A) A scatterplot of intron length versus log-fold-change reported by limma in differential expression profile of ESC cytoplasmic vs. nuclear comparison show a moderately negative correlation; **B)** negative correlation between fold-change and intron length is also observed in TSC data.

A scatterplot between the number of exons per gene and the log-fold-change of up-regulated mRNAs in the cytoplasm also show a moderately negative correlation in both ESC and TSC samples (ρ = -0.24, $p < 2.2 \times 10^{-16}$ and ρ = -0.29, $p < 2.2 \times 10^{-16}$ respectively) (**Figure 32**). This suggests that the level of mRNA enrichment to the cytoplasm is inversely proportional to the number of exons per gene. This result is also in agreement with the Chen *et al* study that showed a negative correlation between the kinetic rate constant of transcription versus the number of exons per gene[95].

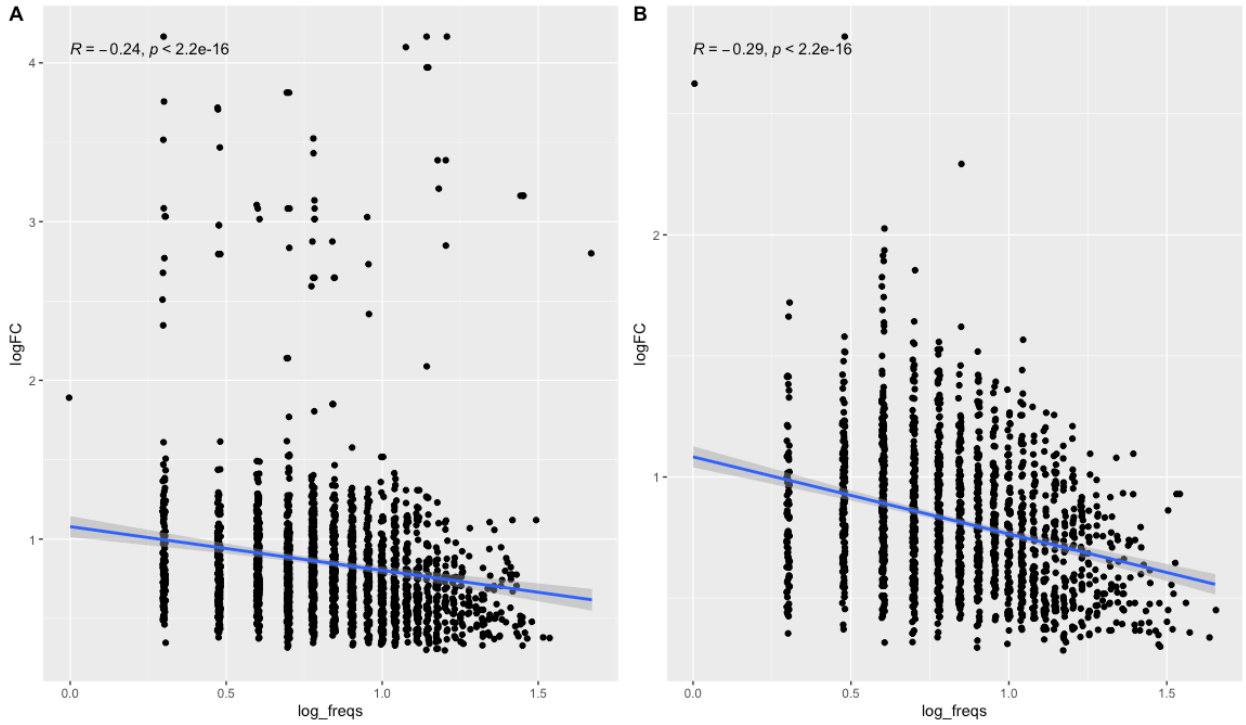


Figure 32: **Correlation between number of exons per gene and log-fold-change in differential expression**

A) A scatterplot of the number of exons per gene and the log-fold-change reported by limma in differential expression profile of ESC cytoplasmic vs. nuclear comparison show a moderately negative correlation; **B)** negative correlation between the number of exons and fold-change is conserved in TSC data.

The relationship between the length of introns and the number of exons per gene with the overall rate of transcription is reflected in the differential expression profile in the current study. This finding not only supports results from literature, but also suggests the current methodology of cell fractionation and normalizing for compartmental mass imbalance using a RNA concentration based method is appropriate in modeling transcriptional behaviour.

3.2.10 Exon-intron proportion quotient distributions

In order to draw inferences on the extent of intron retention in the data, a ratio of featureCounts generated exon counts to intron counts is calculated for mRNA and ncRNA data. An overall distribution of exon-intron count ratios (denoted Q) reveals differences in exon-intron proportions in the cytoplasmic versus nuclear fraction data, as well as differences in mRNA and ncRNA data. A Q value approaching unity suggests vanishingly low number of intron counts relative to exon counts (i.e., exon counts predominate for given gene), whereas a Q value approaching zero indicates low number of exon counts relative to intron counts (i.e., intron counts predominate for given gene).

Distribution of Q values in the cytoplasmic and nuclear count data show differences in exon enrichment. In mRNA data, the population size with respect to Q increase in an exponential-like behaviour in cytoplasmic fractions, such that a significant majority of counts show large exon to intron count proportion (**Figures 33 A & C**). In the nuclear fraction, this trend is not as apparent (**Figures 33 B & D**). This suggests that, according to simple RNA-seq count ratio metrics, the level of intron retention is higher in cytoplasmic RNA compared to nuclear RNA.

This result indicates that the extent of detectable intron enrichment relative to exon counts per gene shows: a) a cytoplasmic-nuclear asymmetry, conserved in both ESCs and TSCs, b) higher proportion of exon counts versus intron counts in the cytoplasmic fraction compared to the nuclear fraction, c) observable difference in distribution between mRNA and ncRNA data, and d) inconclusive cytoplasmic-nuclear differences in ncRNA data.

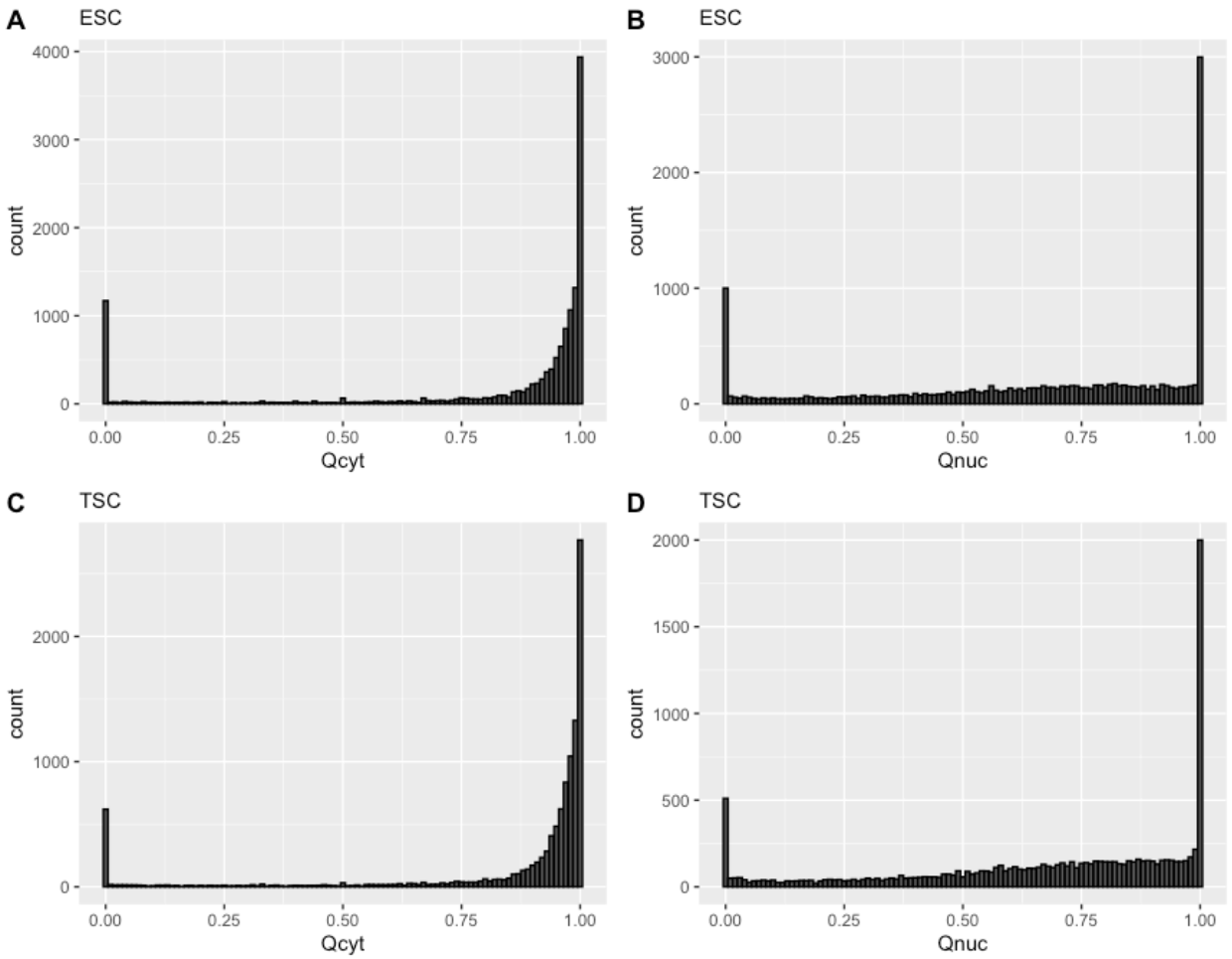


Figure 33: **Distribution of intron-exon proportions in fractionated mRNA count data**

A) The distribution of Q show an increasing trend towards Q =1 in the cytoplasmic read counts in ESCs, suggesting a significant population of genes with high exon proportions; **B)** in nuclear fraction in ESCs, such trend is not as apparent; **C)** cytoplasmic read counts in TSC data show similar results to ESC data; **D)** nuclear read counts in TSC show similar results to ESC counterpart.

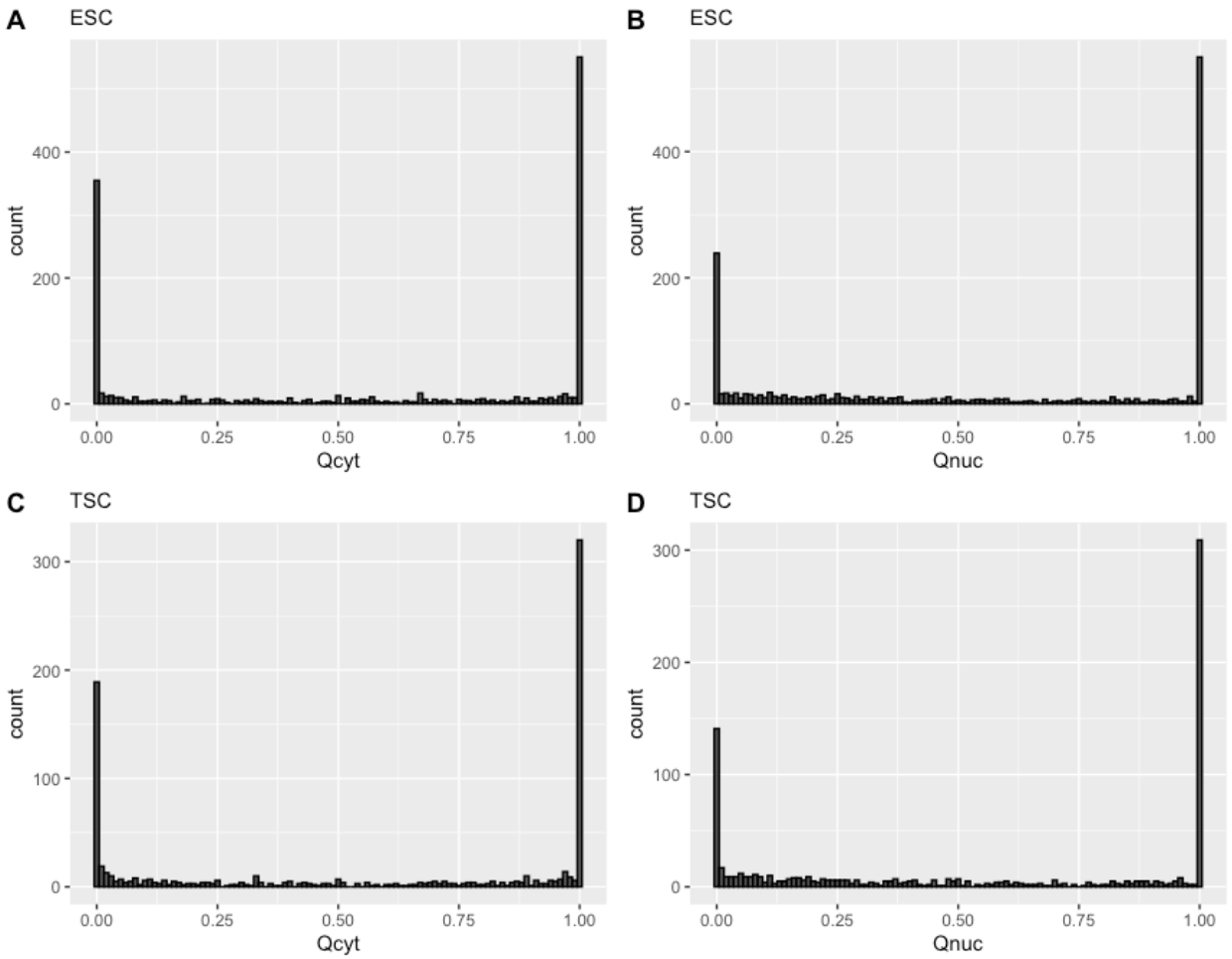


Figure 34: **Distribution of intron-exon proportions in fractionated ncRNA count data**

A) The distribution of Q in ncRNA data show binary results at either extremes ($Q = 0$ and $Q = 1$) compared to mRNA data in ESC cytoplasmic fraction; this result is also shown in **B)** nuclear fraction in ESCs, **C)** cytoplasmic fraction in TSCs, and **D)** nuclear fraction in TSCs.

An interesting observation is the largely binary nature of Q in ncRNA data; this result suggests that a significantly large proportion of ncRNAs are either largely intronic or largely exonic. This is in contrast to mRNA data, where a large population of data retain a nonzero amount of reads align to both exons and introns. In order to identify the population of ncRNAs that make up this binary behaviour, ncRNAs whose Q equals 1 and 0 were categorized by their REFSEQ annotations. **Figures 35** and **36** show in both the cytoplasmic and nuclear fractions in ESCs and TSCs, ncRNAs whose read counts suggest the ncRNA transcript is entirely exonic likely belong to genes with no splice variants - small ncRNAs such as snoRNAs and miRNAs, for example. **Figure 37 A** shows the read alignment of Snord87 in the cytoplasmic and nuclear fraction in ESCs. Other small

ncRNAs showing similar behaviour include pseudogenes such as Gm6524 and ribonucleases such as Rprl3 (**Figure 37 B** and **C**). On the other hand, ncRNAs whose read counts are entirely intronic suggest a possibility of degraded introns or DNA contamination (**Figure 37 D**).

As multiexonic transcripts such as lncRNAs have been shown to undergo processing and splicing similarly to mRNAs, it is likely that detectable functional lncRNAs in both the cytoplasm and nucleus are exon-rich. LncRNAs have been shown to undergo post-transcriptional splicing as well; therefore, intron retaining lncRNAs are most likely indicative of nascent lncRNAs at the site of transcription. Then, the distribution of exon-intron junction counts in ncRNAs across the two subcellular fractions is likely to be skewed towards the nuclear fraction – this will be addressed in **Section 3.2.11**.

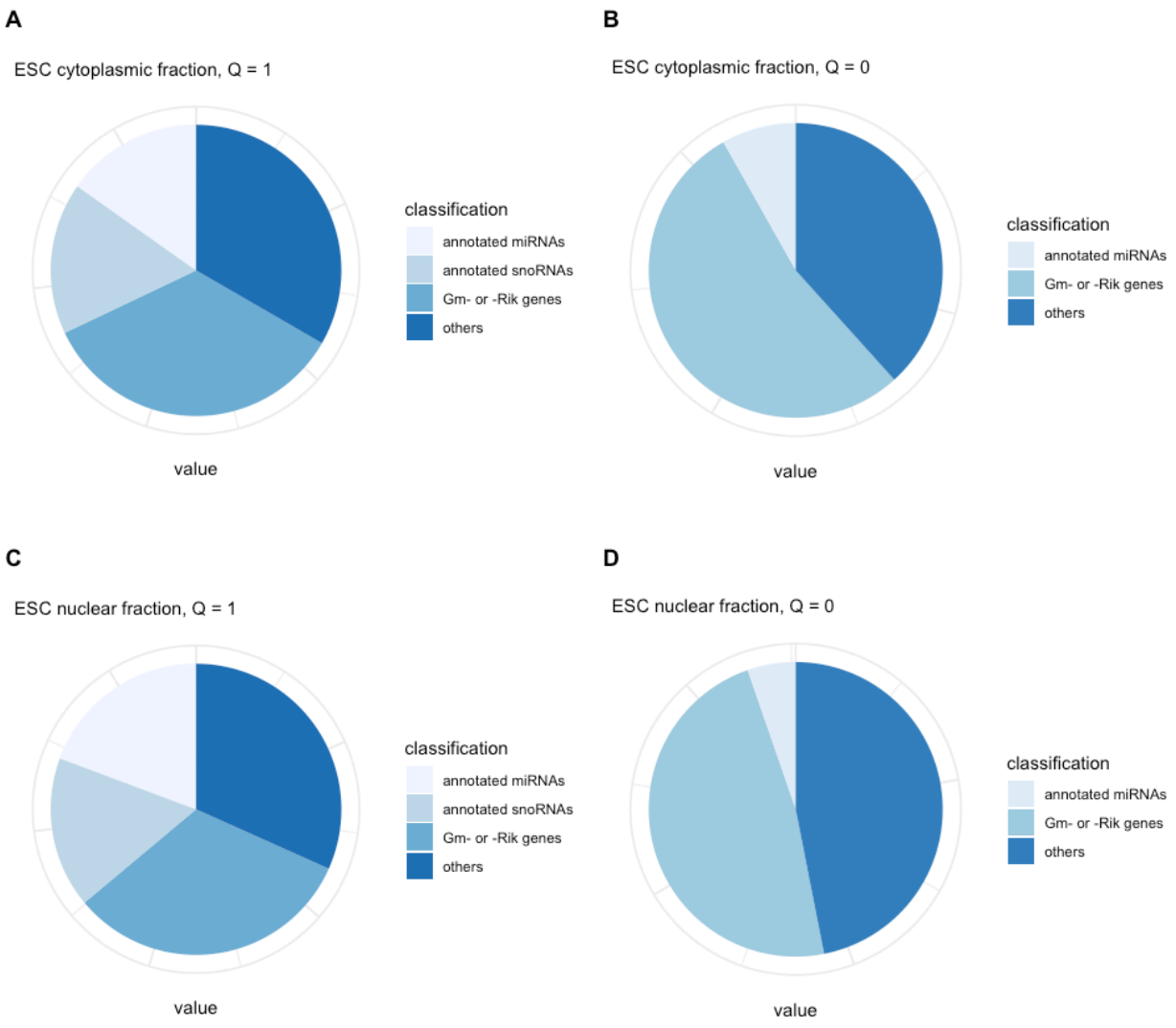


Figure 35: **Pie chart showing proportions of ncRNAs where *Q* equals 0 or 1 in ESCs**

Pie chart shows that in both cytoplasmic and nuclear fractions in ESCs, ncRNAs whose read counts are entirely exonic include small ncRNA species such as snoRNAs and miRNAs; ncRNAs whose read counts are entirely intronic or largely intronic suggest the possibility of degraded introns.

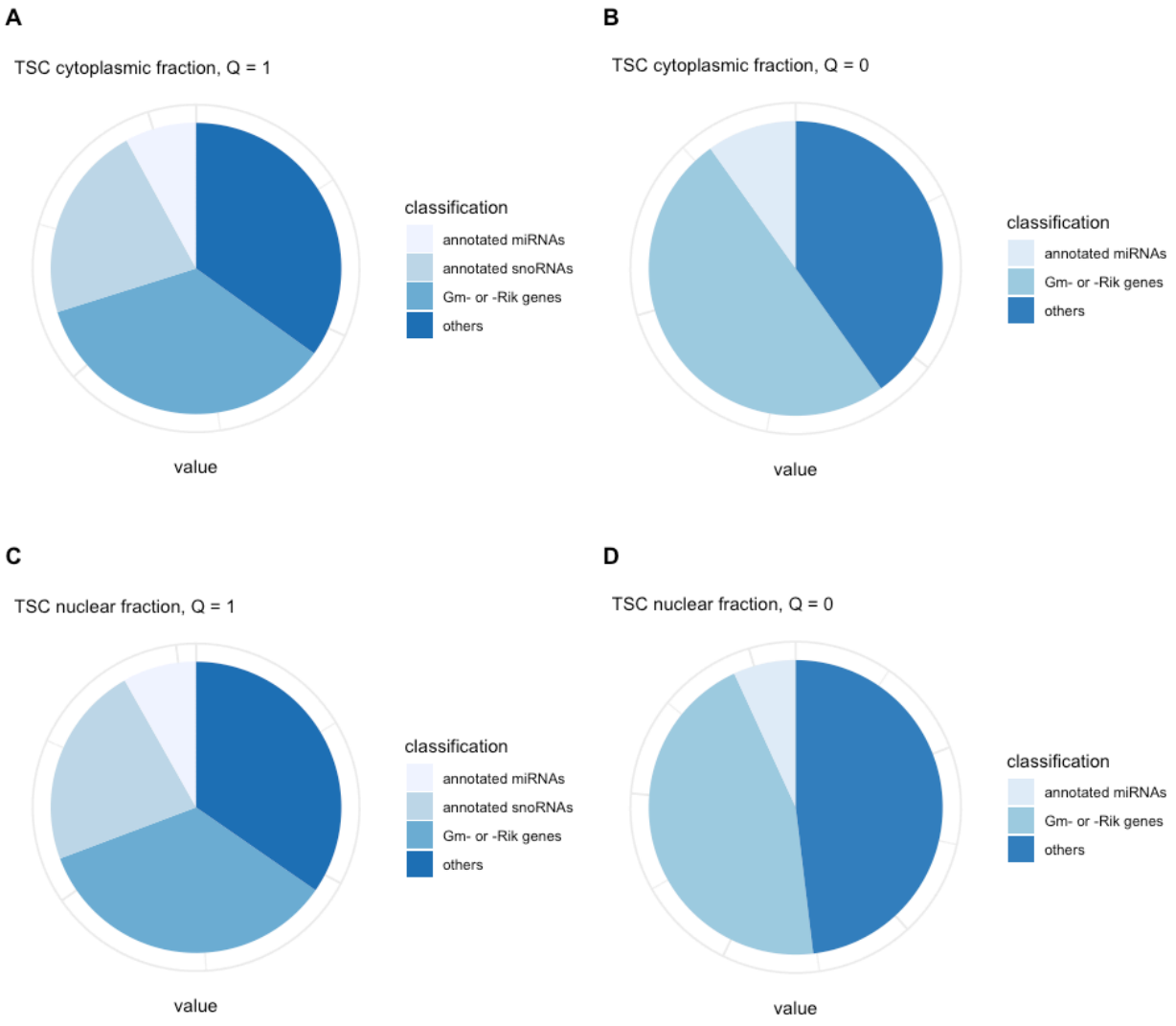
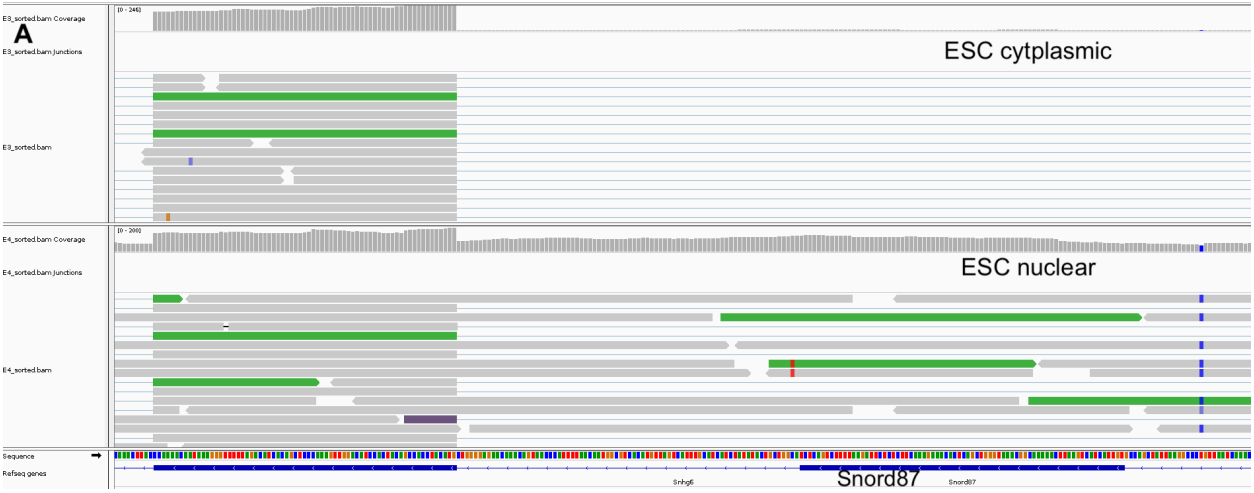
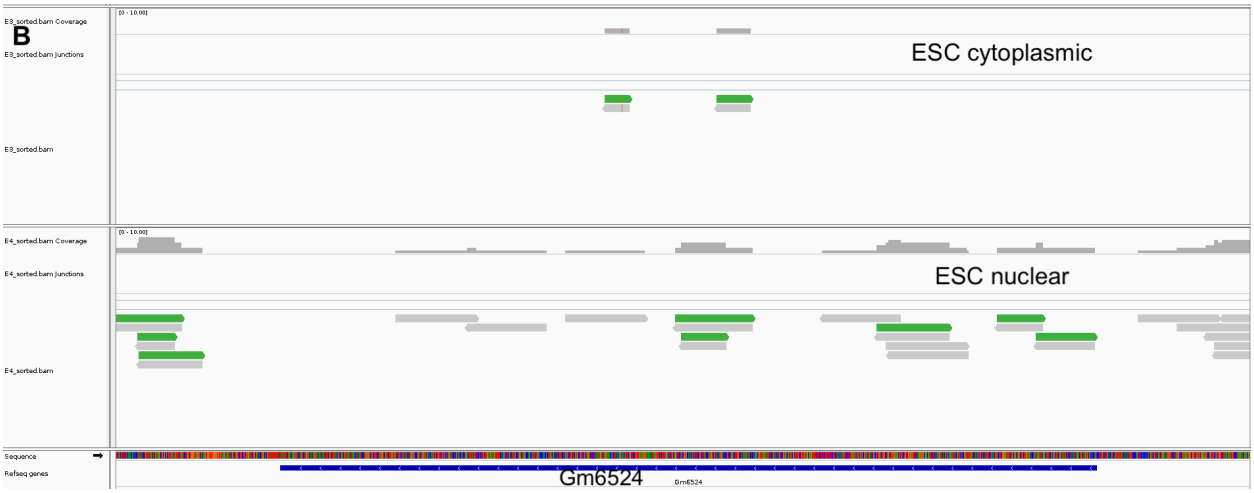
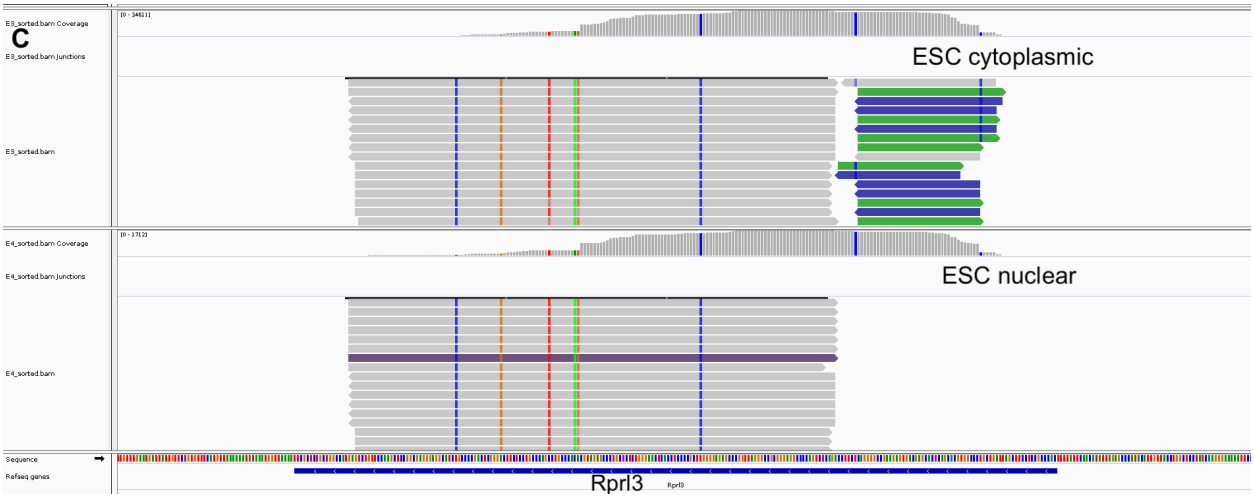


Figure 36: **Pie chart showing proportions of ncRNAs where *Q* equals 0 or 1 in ESCs**

Similarly to ESCs, pie chart shows that in both cytoplasmic and nuclear fractions, ncRNAs whose read counts are entirely exonic include species such as snoRNAs and miRNAs; ncRNAs whose read counts are entirely intronic are largely intronic suggest the possibility of degraded introns.


A
ESC cytplasmic
ESC nuclear
Snord87


B
ESC cytoplasmic
ESC nuclear
Gm6524


C
ESC cytoplasmic
ESC nuclear
Rprl3

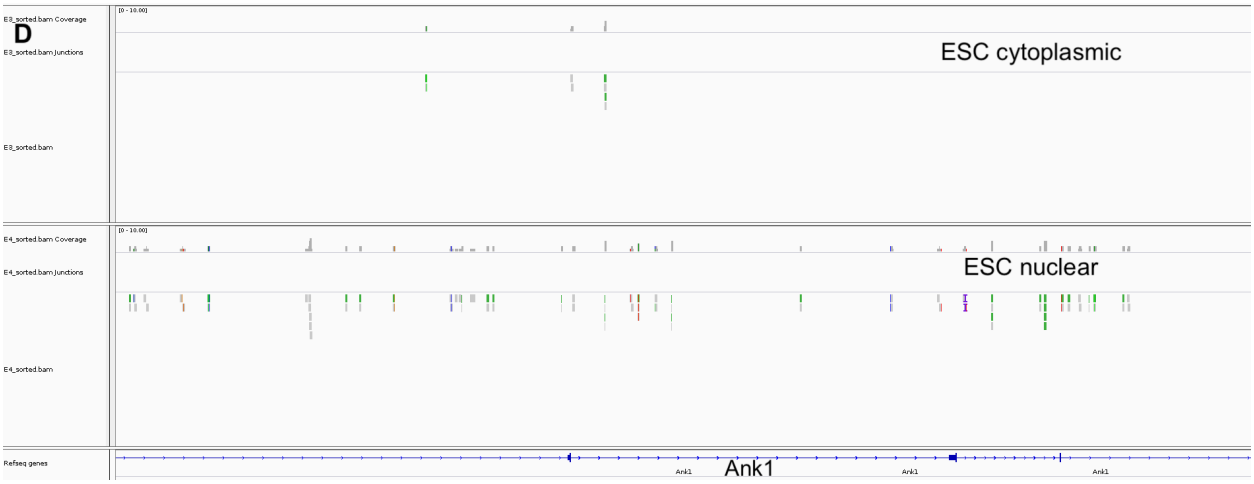


Figure 37: **Aligned read distribution in select ncRNAs according to exon-intron read count proportions**

A) Gene alignment to Snord87 shows that this small ncRNA is transcribed within the intronic region of another gene (Snhg6)**;** ncRNAs whose $Q = 1$ must encompass small ncRNAs such as snoRNAs and miRNAs; **B**) the pseudogene Gm6245 also do not show splicing variants and thus no intron counts; **C)** same can be said for the gene coding for ribonuclease Rprl3; **D)** gene alignment shows a ncRNA whose $Q = 0$ and thus no exonic counts were found; this most likely suggest degraded intronic elements or DNA contamination and is most likely not biologically significant.

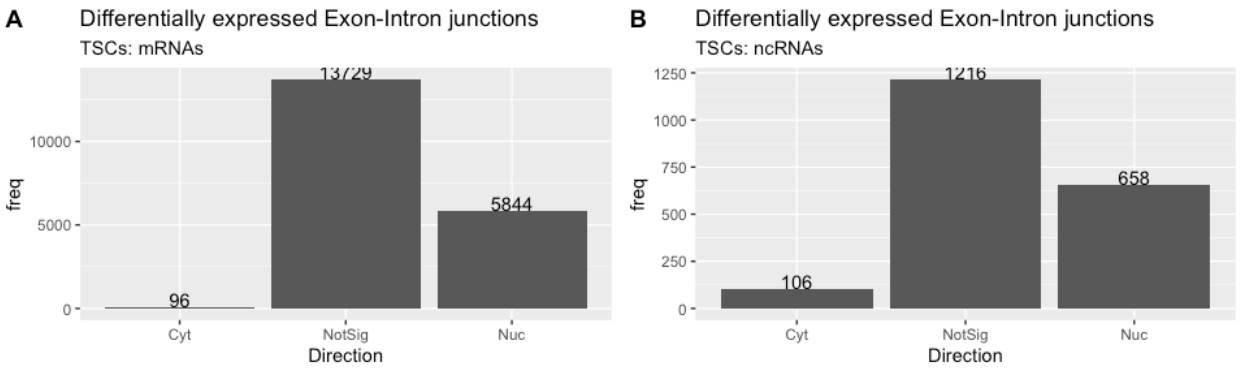
The limitation of count-based quotient metric is that genes with longer introns will tend to yield a larger number of intron aligned reads, which will consequently introduce length based biases in calculated Q values. Indeed, Spearman rank correlation test shows a moderately negative correlation between Q and intron lengths per gene in all sequencing samples ($\rho = -0.39$ and $\rho = -0.42$ in ESC cytoplasmic and nuclear fractions; $\rho = -0.42$ and $\rho = -0.45$ in TSC cytoplasmic and nuclear fractions). This correlation may in fact be attributed to an inflation of intron counts for longer introns, directly leading to lower calculated Q. Furthermore, counting sequencing reads aligned to intron elements as in the case of current analysis, as well as differential expression analysis using intron counts (**Section 3.2.7**) may not always account for retained introns. It is possible that simply counting intron reads leads to accounting for introns that have already been spliced out and thus not representative for an estimation of intron retention. In order to ensure quantification of introns flanking exons and to avoid an inflation of intron retention estimates, exon-intron junction boundaries must be quantified. Quantifying junction reads and using the junction data for differential expression then, also accounts for gene length biases as the expression of the junctions of same genes are compared across samples instead. Nevertheless, differences in

the general distribution of count ratios in cytoplasmic and nuclear samples, as well as in mRNA and ncRNA data provide a qualitative insight into general trends in intron enrichment.

3.2.11 Differential expression profile of exon-intron junctions

Quantifying exon-intron junction reads allows for a more accurate assessment of intron retention as it ensures the introns are flanked by adjacent exons. Calculating a ratio between unspliced (i.e., exon-intron junctions) and spliced junctions (i.e., sum of exon-intron and exon-exon junctions) then allows an estimation of intron retention in the form of percentage intron retention (PIR) as outlined in literature.

Obtained exon-intron junction reads in the form of a count table was used as input for edgeR-voom-limma (using same parameters as before) to make inferences on cytoplasmic-nuclear distributions of the junction counts. This differential expression profile show, in all four comparisons below, a larger population of exon-intron junctions down-regulated in the cytoplasmic fraction (**Figure 38**). This suggests in both mRNA and ncRNA data, the overall level of intron-retaining RNAs is higher in the nucleus compared to the cytoplasm, and this trend is a conserved behaviour in both ESC and TSC samples. This result does not necessarily indicate which genes are more likely to retain their introns, however, as the differential expression profile only contains information on the subcellular distribution of exon-intron junctions. Metrics such as PIR accounting for both spliced and unspliced junctions on a per-gene basis is used instead to infer the extent of intron retention.



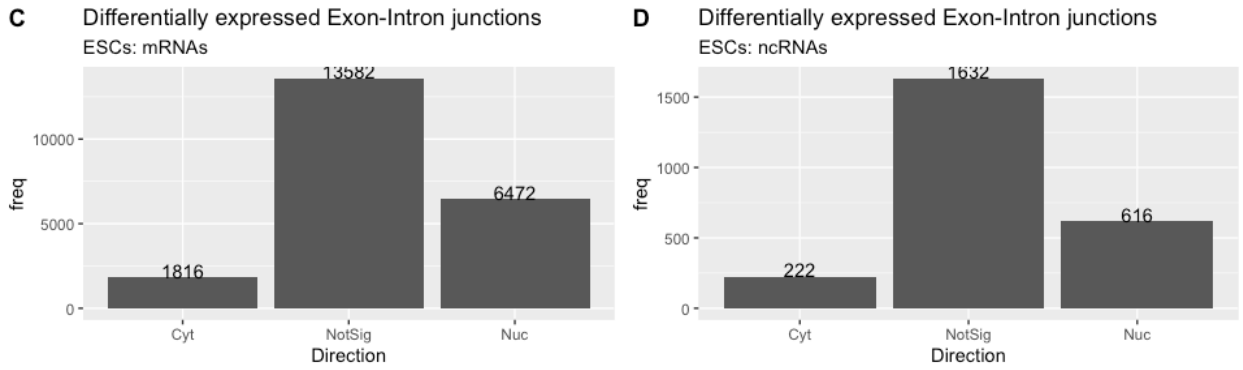


Figure 38: **Differential expression profile of exon-intron junction counts**

A) Differential expression result with limma (adjusted p-value < 0.05) show a significantly larger population of mRNAs with exon-intron junctions up-regulated in the nuclear fraction versus the cytoplasmic fraction in TSCs; **B)** larger population of ncRNAs also show up-regulation in the nuclear fraction in TSCs; **C)** similar result is shown in ESC mRNAs as TSC mRNAs; **D)** similar result is shown in ESC ncRNAs as TSC ncRNAs.

The large population of exon-intron junctions differentially expressed towards the nuclear fraction suggest a few possibilities. These genes may represent nuclear retained mRNAs as part of a gene expression regulatory mechanism. This result may be related to the differential expression analysis result using only exon counts (**Section 3.2.4**) which has shown a similarly larger population of mRNAs up-regulated in the nuclear fraction. This suggests the possibility that the nuclear population of differentially expressed mRNAs from **Section 3.2.4** encompasses intron-containing transcripts as well. In such case, a functional gene set analysis of nuclear fraction may reveal overlapping gene sets. It is also possible that exon-intron junction reads in the nuclear fraction may belong to mRNA species not fully processed, as the RNA-seq strategy used in the current study did not use polyadenylated RNA selection. The consequence of profiling non-polyadenylated RNAs extends to ncRNAs as well; in particular, lncRNAs undergo both co-transcriptional and post-transcriptional splicing and thus the nuclear enrichment of exon-intron junction reads in ncRNA data may be attributed to nascent lncRNAs. It will be useful then, to contrast this result to the distribution of exon-exon junction counts of ncRNAs in the two subcellular fractions – as mature lncRNAs have been documented to have both nuclear and cytoplasmic regulatory roles. Despite the nuclear role of lncRNAs receiving more attention in literature, recent evidence suggest in terms of subcellular localization, lncRNAs are generally more abundant in the cytoplasm. **Section 3.2.13** will show the differential expression profile of exon-exon reads.

3.2.12 *Gene ontology enrichments in exon-intron junction data*

Functional gene set enrichment analysis on the differential expression profile of exon-intron junction data allows for identification of biological processes related to intron-retaining transcripts. Even though literature evidence suggests intron retaining transcripts found in the cytoplasm are often coupled to NMD as a means of gene expression control, it is unclear whether certain function sets of genes are more likely to undergo this degradation pathway. The topGO-REVIGO result for over-represented gene sets in the exon-intron junction expression data show similar results to the gene set enrichment results from the exon count data (**Section 3.2.4**); up-regulated mRNA exon-intron junction gene sets in the ESC cytoplasm are predominated by genes related to translation, metabolic processes, and ribosomal assembly – a result which mirrors results from the exon differential expression profile (**Figure 39, top**). This reiterates the pervasive enrichment of genes related to translation and metabolic processes to the cytoplasmic fractions – an observation supported in literature. This result also suggests that high levels of intron retention may not be tied to a specific functional gene set and rather that intron retention serves as a modulatory mechanism to regulate global gene expression.

The nuclear enriched exon-intron junction gene sets also resemble results from the exon count data, such that gene related to cell division are over-represented (**Figure 39, bottom**). These gene sets may represent nuclear detained mRNAs awaiting for rapid export and translation or mRNAs awaiting nuclear degradation.

Interestingly, in the mRNA exon-intron junction data in TSCs, up-regulated gene sets in the cytoplasmic fraction include an over-representation of genes related to immune response – this result is not seen in ESC data (**Figure 40, top**). In the nuclear fraction of TSCs, genes related to the circulatory system as well as genes related to cell-cell adhesion are over-represented – a result again not seen in ESC data (**Figure 40, bottom**). In contrast, genes associated with neural development are found to be over-represented in the nuclear fraction of ESCs. These differences suggest modulatory mechanisms via intron retention plays a part in regulation of cell-specific processes as well as conserved processes such as translation and ribosomal assembly.

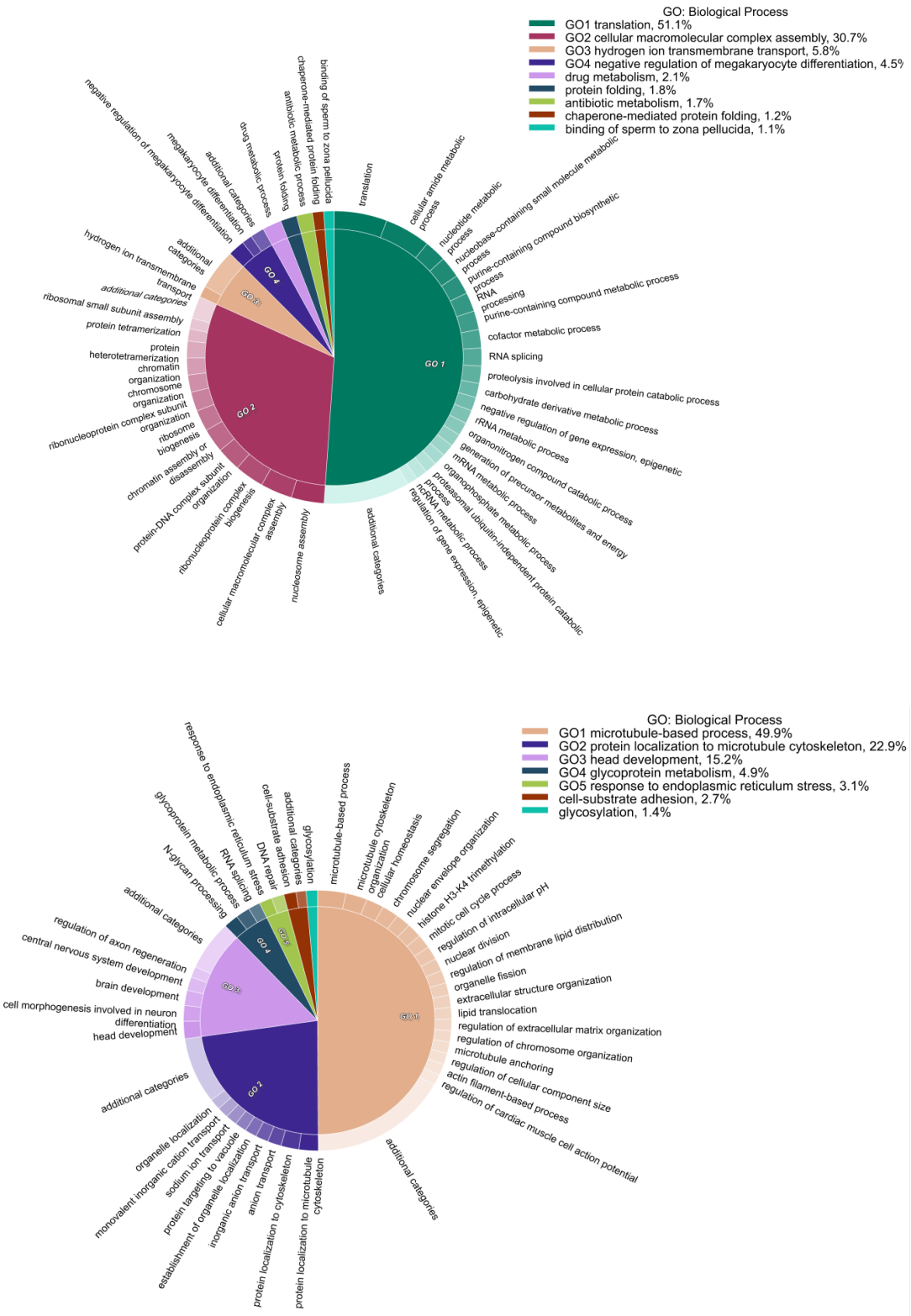


Figure 39: **GO enrichment analysis result on ESC mRNA exon-intron junction data**

Top: enriched GO terms in the ESC cytoplasmic fraction in mRNA unspliced junction data include terms related to translation and ribosomal processes; **Bottom:** enriched terms in the nuclear fraction include terms related to cell division and system development.

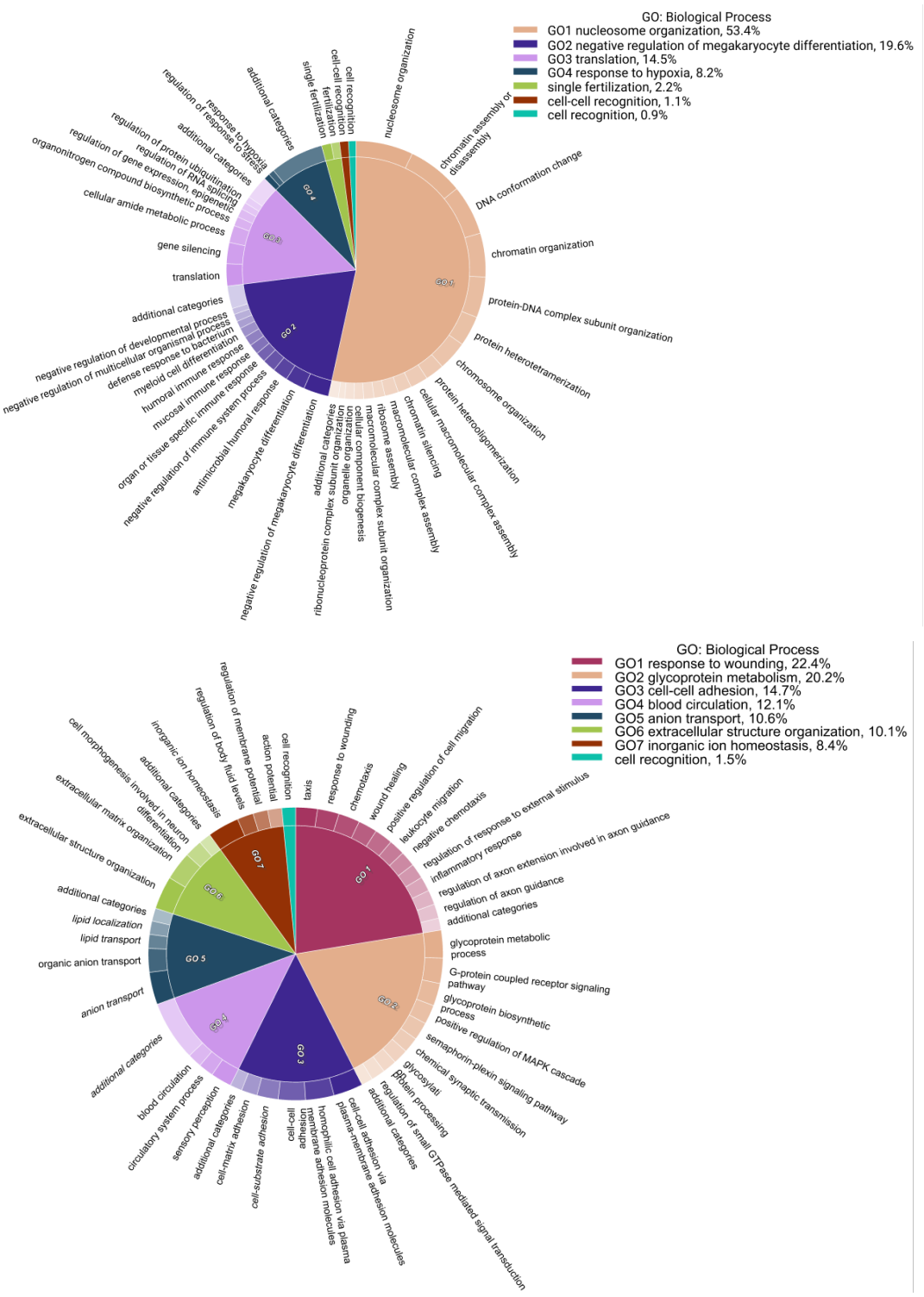


Figure 40: **GO enrichment analysis result on TSC mRNA exon-intron junction data**

Top: enriched GO terms in the TSC cytoplasmic fraction in mRNA unspliced junction data include terms related to chromosome organization and immune response; **Bottom:** enriched terms in the nuclear fraction include terms related to immune response, cell-cell adhesion, cell signaling pathways, and the circulatory system.

Introns often harbor ncRNAs and thus intronic RNA cannot simply be designated as pre-mRNA or RNAs destined for degradation pathways[105]. Functional gene set enrichment analysis of intron retaining ncRNAs reveal an abundance of genes related to cell response and cell signaling in all samples. Overall, results again resemble that from analysis with exon data (**Section 3.2.5**) which suggests differentially expressed ncRNAs in both cytoplasmic and nuclear fractions (**Section 3.2.4**) may predominantly show some levels of intron retention. Corroborating with results from **Section 3.2.9**, it is also plausible that the extent of intron retention across the population of ncRNAs is more conserved compared to mRNAs.

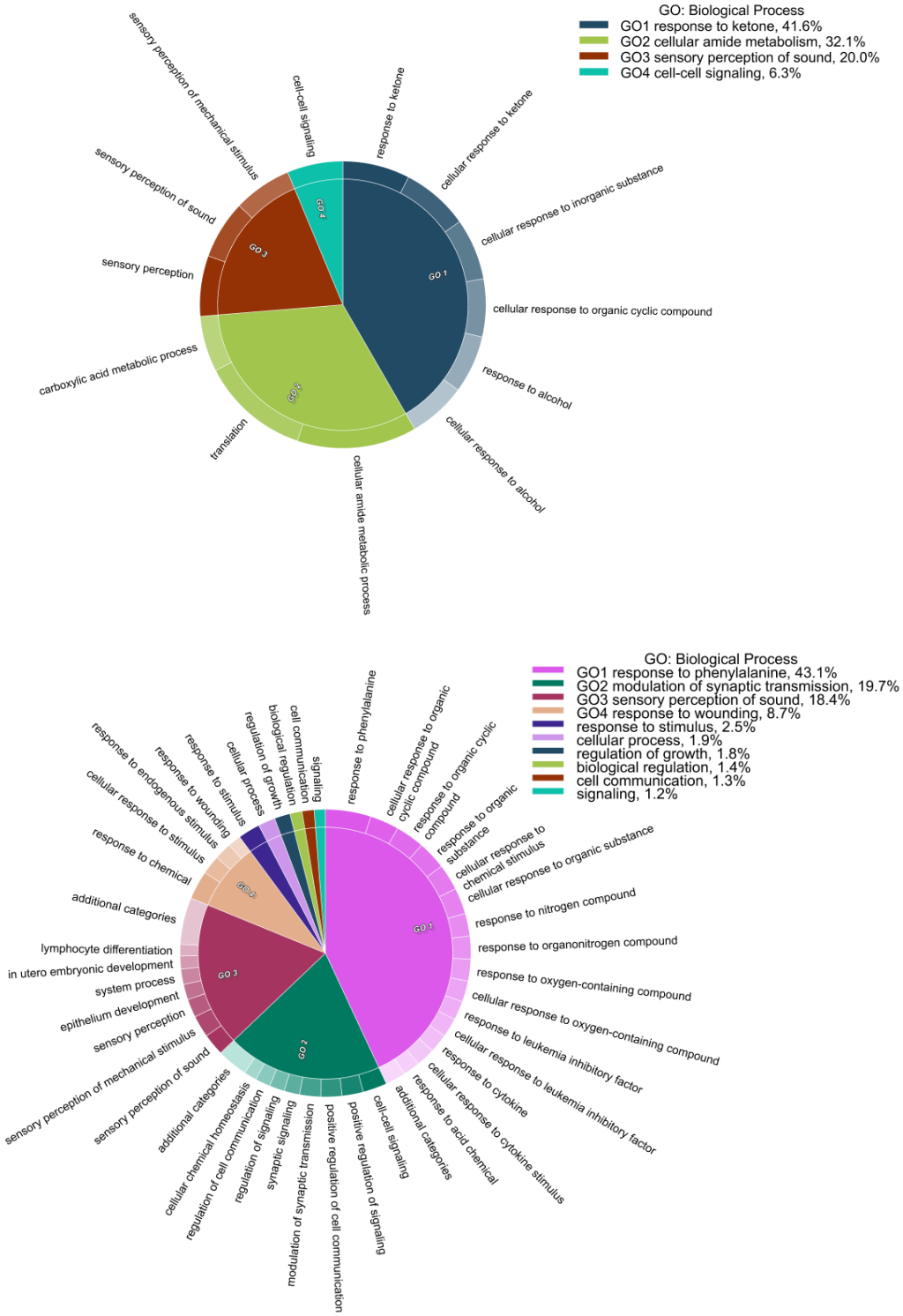


Figure 41: **GO enrichment analysis result on ESC ncRNA exon-intron junction data**

Top: enriched GO terms in the ESC cytoplasmic fraction in ncRNA unspliced junction data include terms related to cell response; **Bottom:** enriched terms in the nuclear fraction include terms related to cell response to stimuli, cell signaling, and system development.

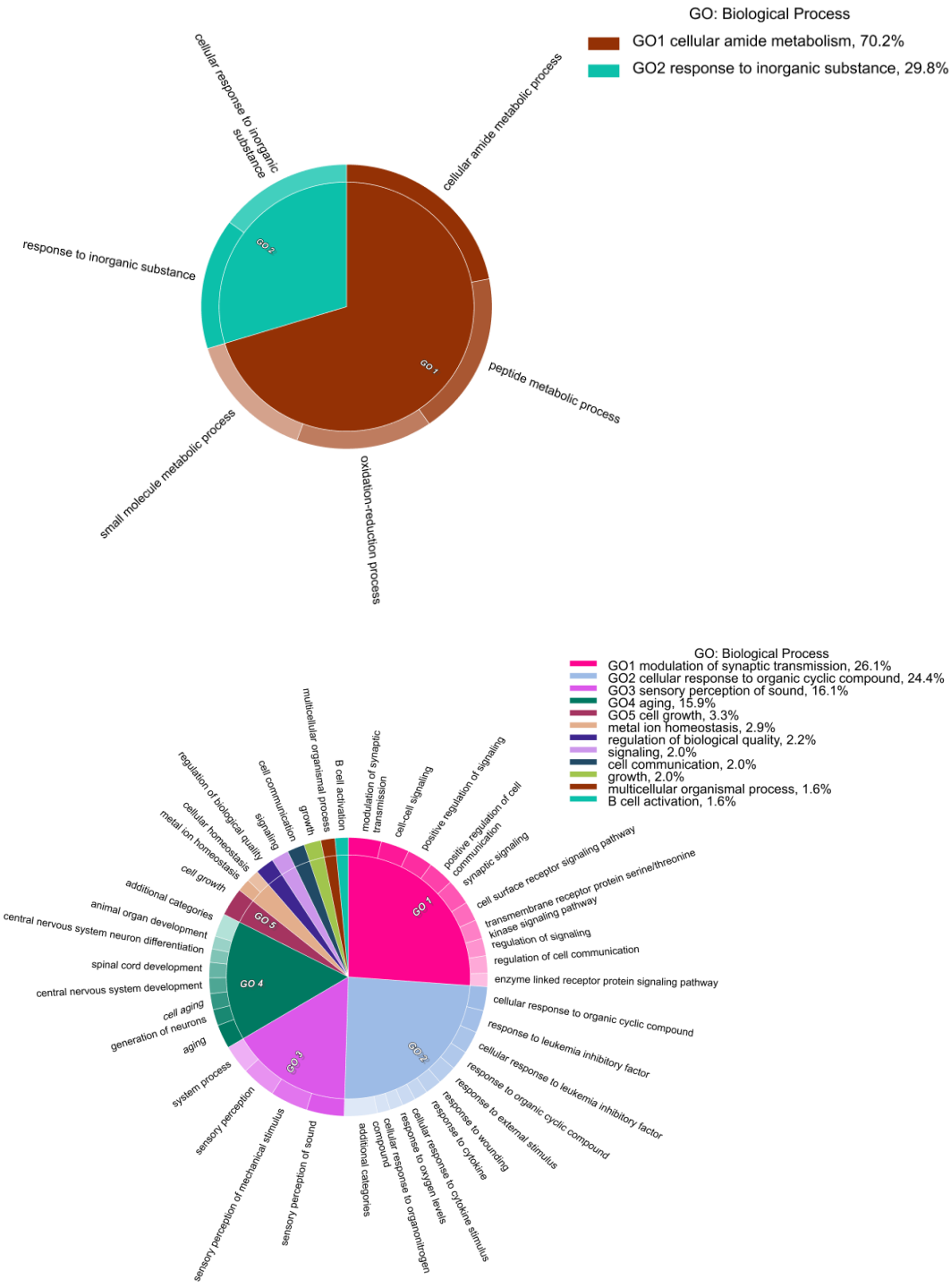


Figure 42: **GO enrichment analysis result on TSC ncRNA exon-intron junction data**

Top: enriched GO terms in the TSC cytoplasmic fraction in ncRNA unspliced junction data include terms related to metabolic processes; **Bottom:** enriched terms in the nuclear fraction include terms related to cell signaling, response to stimuli, and system development.

3.2.13 Differential expression profile of split reads

Split reads in gene alignment indicate reads spanning exon-exon junctions spanning an intron. In order to measure PIR, both unspliced (exon-intron) and spliced (exon-exon) junctions must be identified and quantified. Furthermore, the differential expression profile of exon-exon junctions itself provides insight into the subcellular distribution of spliced transcripts. Using edgeR-voom-limma pipeline as before, the distribution profile of spliced reads reveal a larger population of mRNAs down-regulated in the cytoplasmic versus nuclear fraction (**Figure 43, A & C**). This observation is similar to the result from exon-intron junction distribution (**Section 3.2.10**) which suggests a strong possibility of capturing genes with both spliced and unspliced junctions, such that there is significant overlap between the population of differentially expressed junctions. This result is due to the fact that all junction counts of a given transcript are collapsed under one unique gene identifier for the sake differential expression analysis. This is further complicated by the presence of multiple annotated exon-exon and exon-intron junctions for a given gene.

Nevertheless, the large population of spliced junction reads down-regulated in the cytoplasm is an interesting result as this suggests the presence of processed mRNAs in the nucleus. It is unclear whether this behaviour can be attributed to the kinetics of nuclear export. The nuclear export of processed mRNAs is a multi-step process (including passive diffusion to the nuclear pore) and literature in *in situ* hybridization studies have been able to detect the presence of mRNAs from the nucleoplasm to the nuclear pores. The relatively small population of spliced reads up-regulated in the cytoplasmic fraction suggest the rate of translation and cytoplasmic decay may play a significant role in the overall distribution of differential expression.

Interestingly, the spliced reads differential expression profile in ncRNA show discordant behaviour compared to the exon-intron junction data. A significant majority of ncRNAs show an up-regulation of spliced reads in the cytoplasmic versus the nuclear fraction (**Figure 43, B & D**). This suggests a significant portion of ncRNAs found in the cytoplasm may exist as spliced forms.

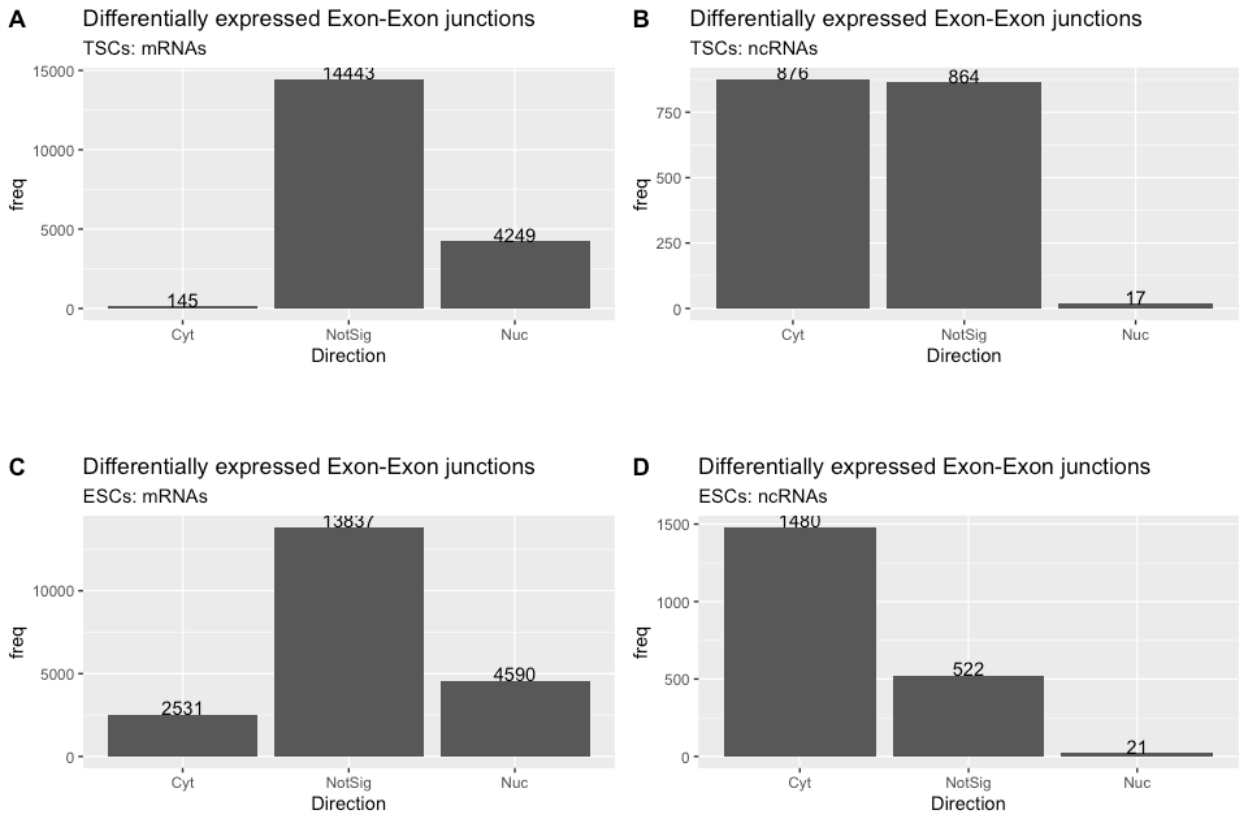


Figure 43: **Differential expression profile of exon-intron junction counts**

A) As in the case of exon-intron junction data, the population of up-regulated mRNA exon-exon junction reads in the TSC nuclear fraction is significantly larger than the population in the cytoplasmic fraction; **B)** in contrast to the exon-intron junction data, the population of up-regulated ncRNAs in the TSC cytoplasmic fraction is significantly larger; **C)** similar result is shown for mRNAs in ESCs as in the case of TSCs; **D)** similar result is shown for ncRNAs in ESCs as in the case of TSCs.

3.2.14 Gene ontology enrichments in split read data

A significant overlap in over-represented functional gene sets between spliced mRNA reads and unspliced reads (**Section 3.2.11**) suggest a large population of genes containing both types of junctions in sequenced RNA. In ESCs, the population of up-regulated spliced reads in the cytoplasmic fraction show similar ontological enrichment results to the enrichment results in unspliced reads; predominantly cytoplasmic biological processes tend to be related to metabolism of macromolecules, translation, and ribosomal assembly (**Figure 44, top**). This result suggests mRNAs with such functions show some level of alternative splicing events. Furthermore, this reoccurring enrichment of such functional gene sets in the cytoplasm suggest a strong relationship between the rate of transcription and the level of cytoplasmic enrichment – as literature suggest

gene sets related to metabolism and translation show the highest kinetic rates of transcription. This observation, along with the negative correlation result between log-fold-change in the cytoplasm and intron length (**Section 3.2.8**) suggest a relationship between the rate of transcription, cytoplasmic enrichment, and intron proportions.

In mRNA gene sets down-regulated in the cytoplasm in ESC data, genes related to cell division, chromosome organization, and cell cycle are over-represented in both spliced and unspliced data (**Figure 44, bottom**). However, the over-representation of gene sets related to ion transport is not observed in spliced reads, despite its presence in unspliced data (**Section 3.2.11**). This suggests that mRNAs related to ion transport function show a high level of intron retention but not necessarily a high level of splicing events – hence a high value of PIR. Downstream analysis in PIR using junction quotients will investigate this level of intron retention in a given gene set.

In TSC mRNA data, gene sets related to metabolic processes in the cytoplasmic fraction show a higher level of over-representation in spliced read data versus unspliced reads (**Figure 45, top**). Genes related to nucleosome organization and immune response are both up-regulated in the cytoplasm in terms of spliced and unspliced reads. In the nuclear fraction, genes related to circulatory system, cell adhesion, and cell movement are over-represented in both spliced and unspliced data (**Figure 45, bottom**), whereas genes related to wound healing and axon guidance show over-representation above threshold (i.e., top 40 ontological categories calculated by REVIGO – see **Section 3.1.4**) in unspliced reads but not in spliced reads.

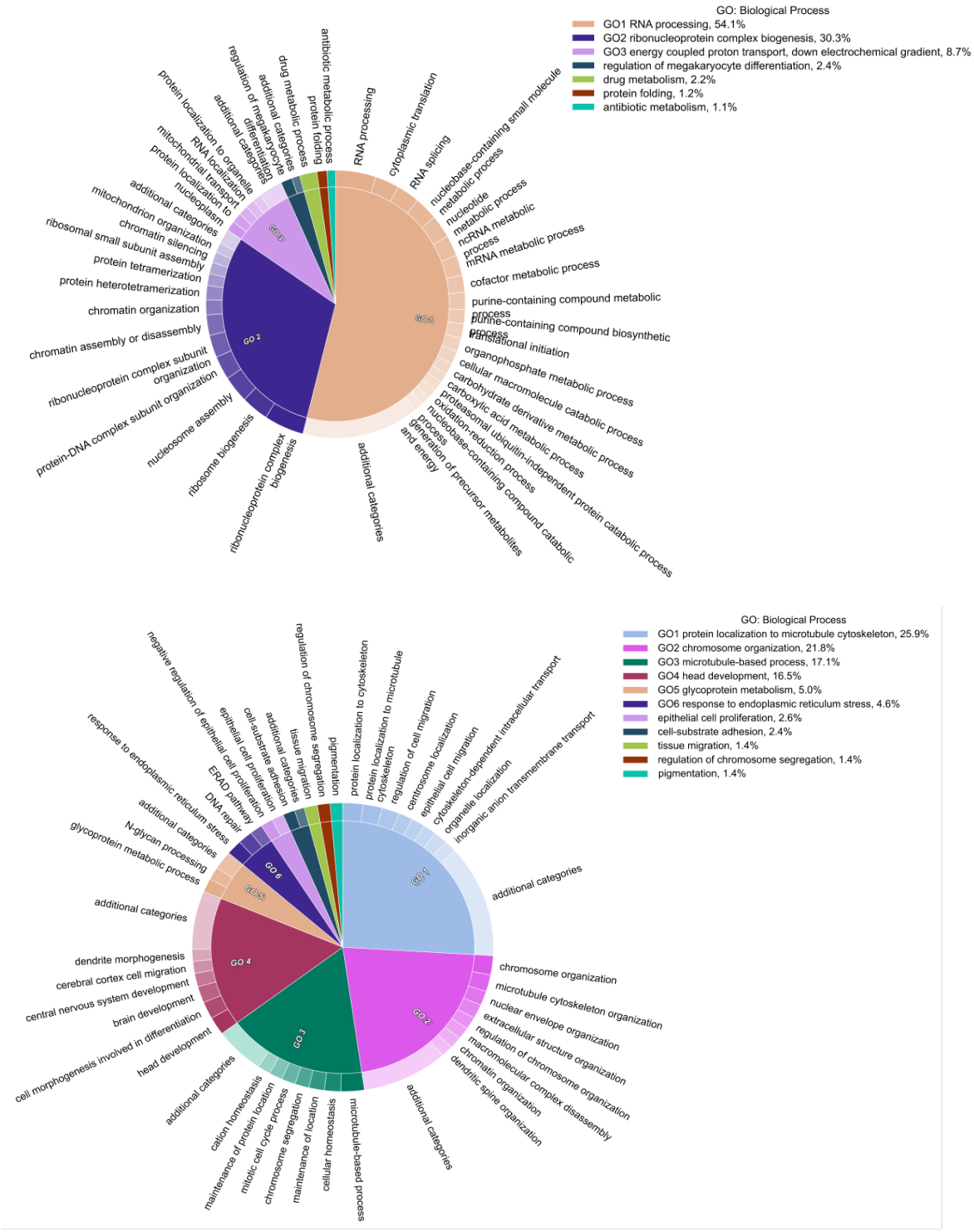


Figure 44: **GO enrichment analysis result on ESC mRNA exon-exon junction data**

Top: enriched GO terms in the ESC cytoplasmic fraction in mRNA spliced junction data include terms related to translation, metabolic processes, and ribosomal assembly; **Bottom:** enriched terms in the nuclear fraction include terms related to cell division, chromosome organization, and system development.

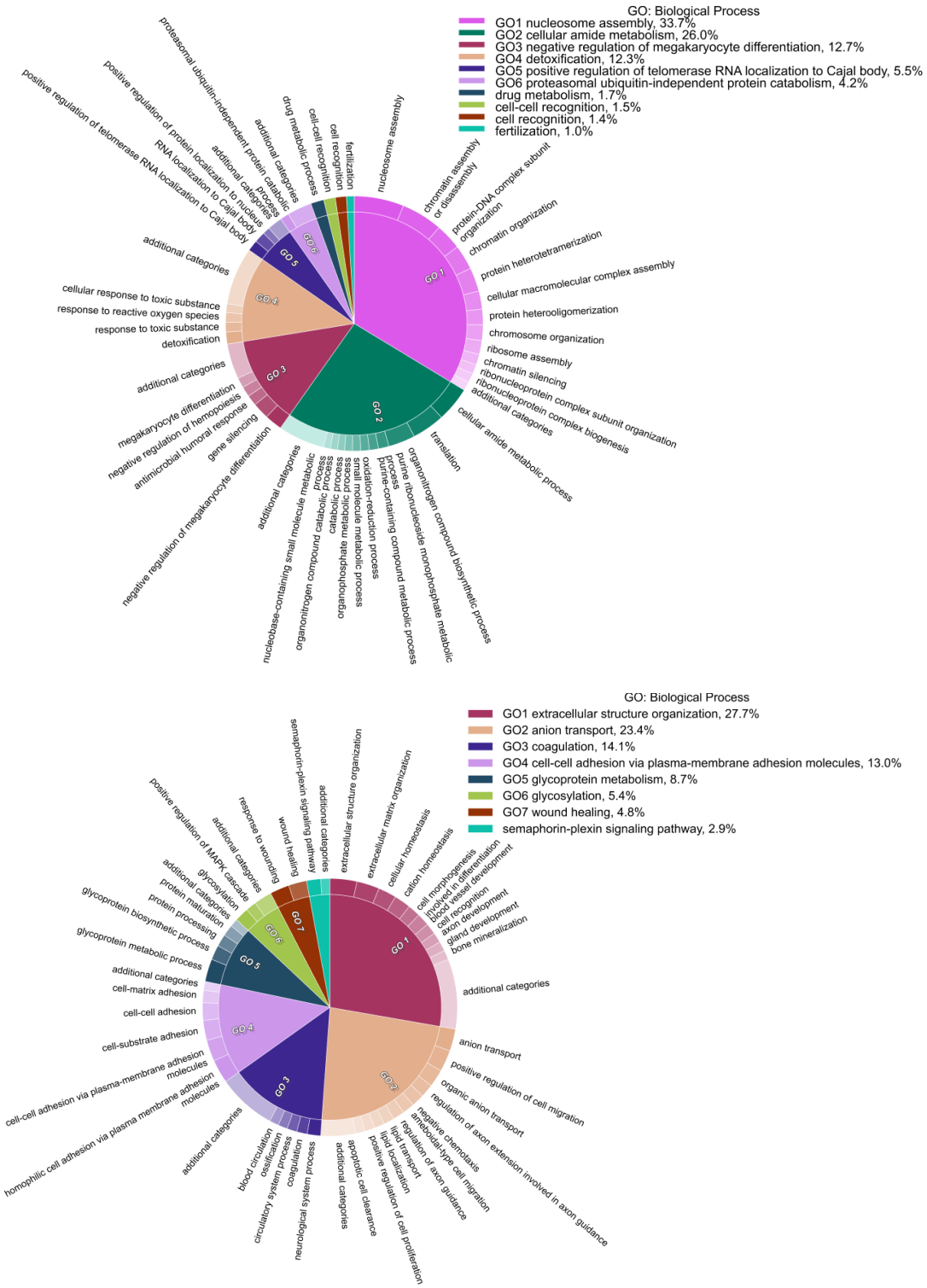


Figure 45: **GO enrichment analysis result on TSC mRNA exon-exon junction data**

Top: enriched GO terms in the TSC cytoplasmic fraction in mRNA spliced junction data include terms related to chromosome assembly, ribosomal processes, and metabolic processes; **Bottom:** enriched terms in the nuclear fraction include terms related to cell structure, ion transport, and circulatory system.

The distribution of spliced reads in ncRNA is interesting as it shows a drastic shift of differentially expressed read towards the cytoplasmic fraction. This suggests the possibility that ncRNAs show more of a binary behaviour in terms of splicing and intron retaining events compared to mRNAs – such that the population of RNAs containing both exon-intron and exon-exon junctions is smaller. This results in a lower overlap between up-regulated unspliced reads and spliced reads, unlike the observation shown in mRNA data above. This observation is reflected in the results from exon-intron count proportions analysis using quotient values (**Section 3.2.10**) where unlike in mRNA data, the Q values for ncRNA show largely binary outcomes (i.e., $Q = 0$ or 1). This suggests the possibility that unlike the phenomenon of alternative splicing events and intron retention as pervasive mechanism to modulate mRNA expression, in ncRNAs the level of exon-intron proportion is generally related to its unique identity.

The population of nuclear exon-exon junction reads in ncRNA data is too small for a meaningful gene set enrichment analysis. In cytoplasmic ncRNA, over-represented gene sets in spliced reads are predominantly related to translation, metabolic processes, and ribosomal processes – a result similar to over-representation analysis in cytoplasmic mRNAs (**Figure 46 & 47**). This result is in contrast with gene sets enriched in the cytoplasm in unspliced reads (**Section 3.2.12**) wherein ncRNAs related to cell response were found to be over-represented. This observation reiterates the possibility of ncRNAs and their exon-intron proportions being related to the function and identity of given ncRNA, whereas intron retention and alternative splicing is a pervasive event in mRNAs.

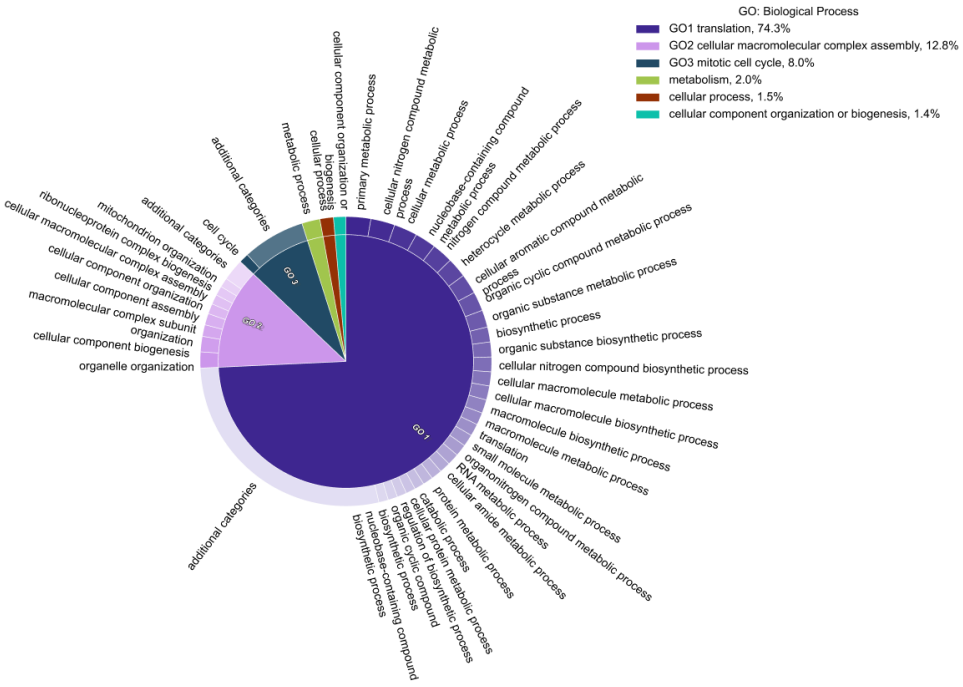


Figure 46: **GO enrichment analysis result on spliced ncRNAs up-regulated in ESC cytoplasmic fraction**

GO terms enriched in the ESC cytoplasmic fraction in ncRNA exon-exon junction counts are predominantly related to metabolic processes and macromolecular complex assembly.

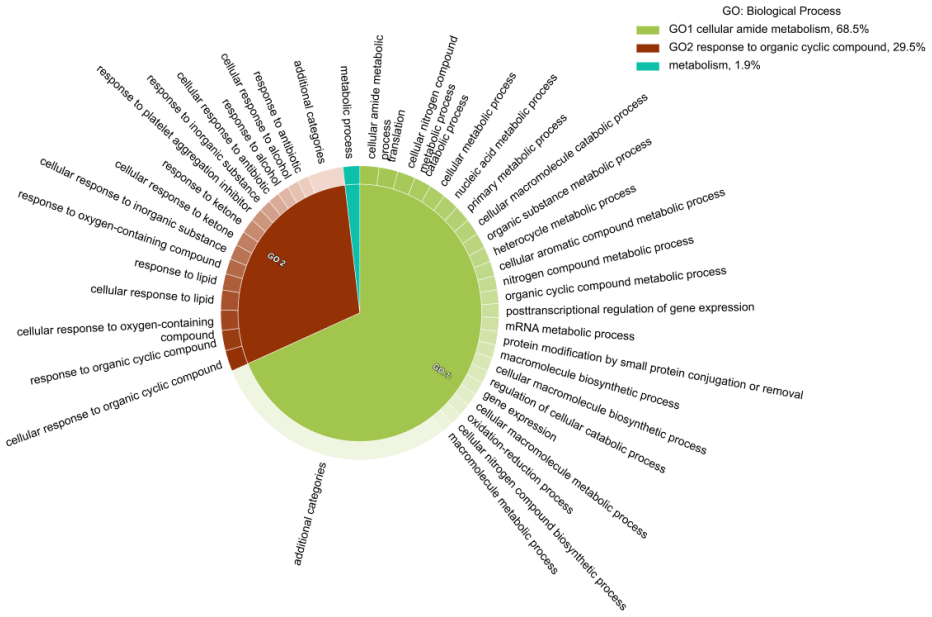


Figure 47: **GO enrichment analysis result on spliced ncRNAs up-regulated in TSC cytoplasmic fraction**

GO terms enriched in the TSC cytoplasmic fraction in ncRNA exon-exon junction counts are predominantly related to metabolic processes and cell response to stimuli.

3.2.15 Calculation of junction quotients for intron retention

In order to estimate the extent of intron retention per gene, the ratio of unspliced to total (i.e, sum of unspliced and spliced) read counts is calculated as a measure of PIR. This ratio – denoted JQ - uses exon-intron and exon-exon junction counts from **Section 3.2.11** and **Section 3.2.13** for annotated genes used in current analysis. JQ is directly proportional with PIR such that genes with relatively high JQ denotes genes with high PIR.

The overall distribution of JQ in cytoplasmic and nuclear fractions show statistically distinguishable means in both ESC and TSC data (**Figure 48**). In mRNAs, the mean JQ is statistically lower in cytoplasmic ($x = 0.355$, $s = 0.135$) versus nuclear fraction ($x = 0.419$, $s = 0.181$) in ESCs with $t = -43.73$ and $p < 2.2 \times 10^{-16}$ using Welch's two sample t-test at 95% confidence interval. This result is shown in TSCs as well, with cytoplasmic ($x = 0.360$, $s = 0.145$) JQ lower on average than nuclear ($x = 0.426$, $s = 0.189$) with $t = -40.56$ and $p << 2.2 \times 10^{-16}$ at 95% confidence interval.

This suggests on average, mRNAs in the nuclear fraction show higher levels of intron retention. This result is in agreement with literature findings that suggest higher levels of intron retention detected in the nucleus than the cytoplasm due to association of intron retention with either NMD or nuclear detainment[56].

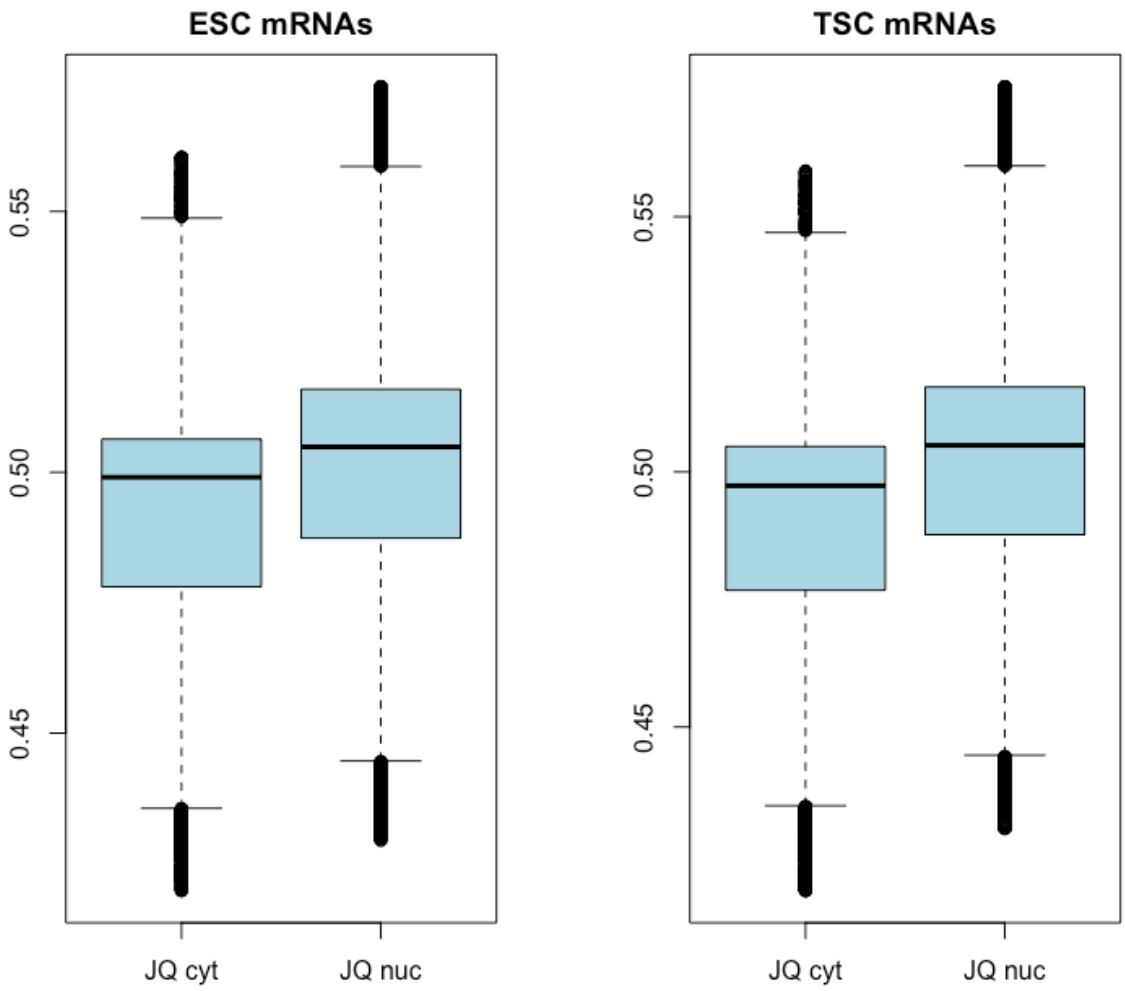


Figure 48: **Boxplot of junction quotient distributions in all sample groups**

In all sample groups, the junction quotient values show higher median values in the nuclear fraction versus in the cytoplasmic fraction.

3.2.16 Relationship between junction quotients and gene expression

A scatterplot between mRNA expression in lcpm and calculated JQs per gene shows a fairly positive correlation in the cytoplasmic fraction ($\rho = 0.47$) as well as in the nuclear fraction ($\rho = 0.41$) (**Figure 49**). This result suggests in both subcellular fractions, the level of detectable mRNAs increases with extent of intron retention. This finding in the cytoplasmic fraction possibly takes into account the intron retention role in stabilization of mRNA, as well as in translation into isoforms. Hence for intron retaining mRNA transcripts that did not undergo degradation via NMD,

the transcript may indeed be stabilized and accumulate in the cytoplasm. It should be noted that depletion of mRNA via NMD cannot be effectively described using read counts without the presence of a NMD negative control.

This finding also supports literature findings that suggest intron retaining mRNA transcripts may be actively detained in the nucleus and avoid nuclear degradation.

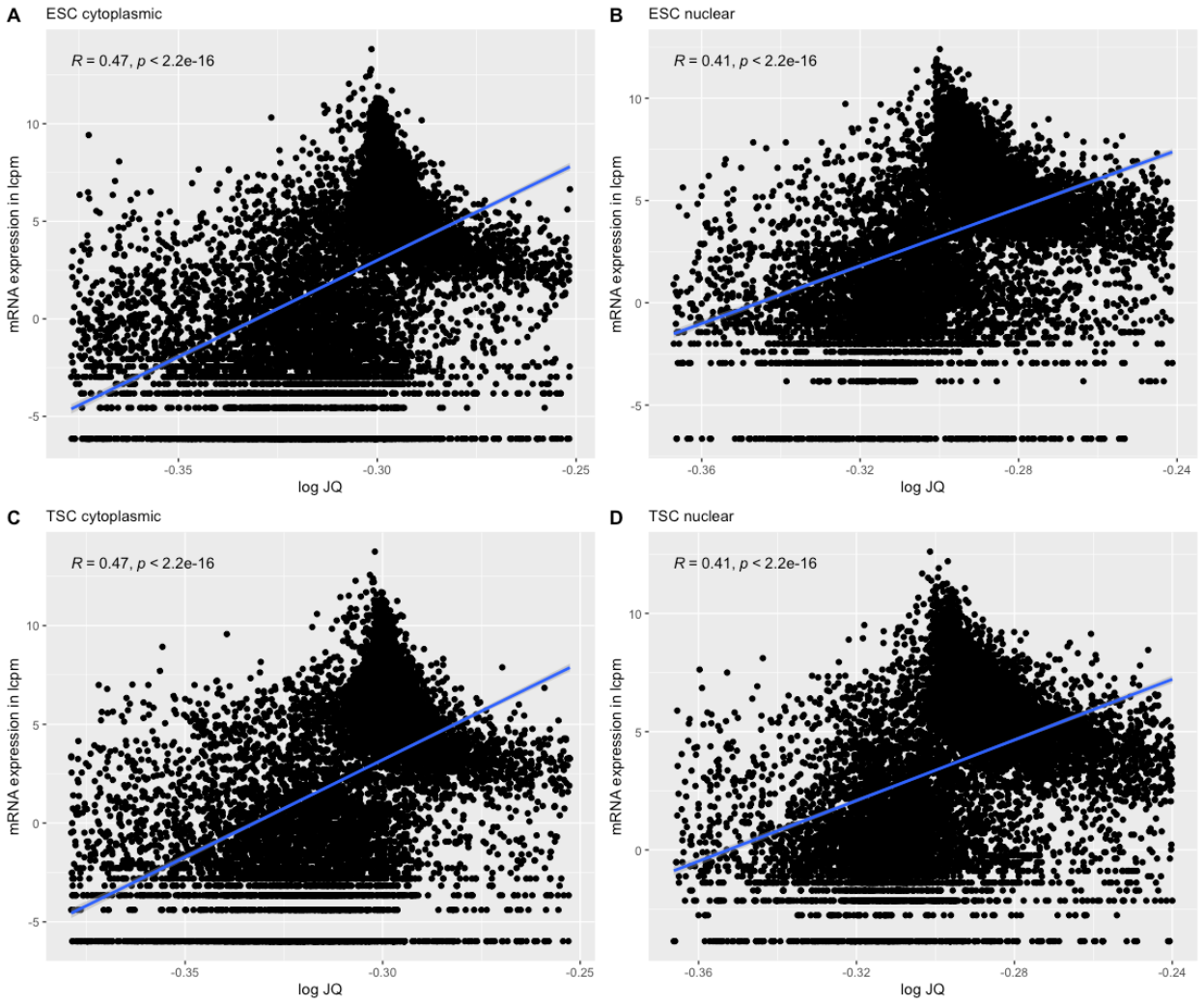


Figure 49: **Relationship between gene expression and percent intron retention measured via junction quotients**

In both the cytoplasmic and nuclear fractions in ESCs and TSCs, the relationship between calculated junction quotients (estimating for percent intron retention) per gene and its expression in lcpm shows a moderate positive correlation; this result suggests that in both fractions, accumulation of transcripts may be associated with intron retaining behaviour. This suggests that in the cytoplasm, the intron retaining mRNA transcripts that did not undergo degradation via nonsense mediated decay may in fact be stabilized. Furthermore, this finding reiterates literature findings that nuclear detained intron retaining transcripts may be actively detained.

3.2.17 Over-representation analysis using junction quotient quantiles

Dividing the distribution of JQs into equal quantiles enables segregated analysis of over-represented gene sets by PIR. Identification of biological processes with each quantile of PIR provides insight into whether certain gene sets are more predisposed to undergo intron retention. In the current analysis, in *n* quantiles with increasing *n* indicate higher values for JQ and larger extent of intron retention

In cytoplasmic ESC mRNA data, gene sets related to catabolic processes are over-represented in the first quantile, suggesting a relatively low degree of PIR (**Figure 50, A**). The second quantile shows an over-representation of genes related to RNA processes, as well as genes related to the chromosome and histone modifications (**Figure 50, B)**. Such gene sets are also over-represented in the third quantile, as well as terms associated with the cell cycle, cell division, and DNA repair (**Figure 50, C**). Gene sets related to DNA processes, cell organization, and cell signaling show the highest degree of PIR, as shown by the result in the fourth quantile (**Figure 50, D**).

Gene sets related to cell organization are over-represented in the first JQ quantile in the nuclear fraction, suggesting relatively low degree of PIR in the nucleus (**Figure 51, A**). As well, terms related to catabolic processes are enriched in the first quantile, similar to results from the cytoplasmic fraction. The second quantile shows an abundance of gene sets related to metabolic processes, as well as chromatin and histone modifications (**Figure 51, B**). The third and fourth quantile show similar results to the cytoplasmic fraction, with an abundance of gene sets related to cell division, cell cycle, and RNA processes (**Figure 51, C & D**). Gene sets associated with methylation processes also show enrichment in the fourth quantile, suggesting a high degree of PIR.

Trends in PIR show some similar results in TSCs, suggesting IR may be a conserved mechanism in both cell types. As in the case of ESCs, the two subcellular fractions in TSCs show similarities in the over-representation of gene sets across the four quantiles. The first quantile shows an enrichment of genes related to catabolic processes as well as cell signaling pathways (**Figure 52, A**). Genes related to metabolic processes and chromatin modifications are enriched in the second quantile, whereas genes related to RNA processes are enriched in the third and fourth quantiles

(**Figure 52, B, C, & D**). The third quantile also shows over-representation of terms related to the cell cycle and cell division, while the fourth quantile shows an abundance of genes related to cell organization.

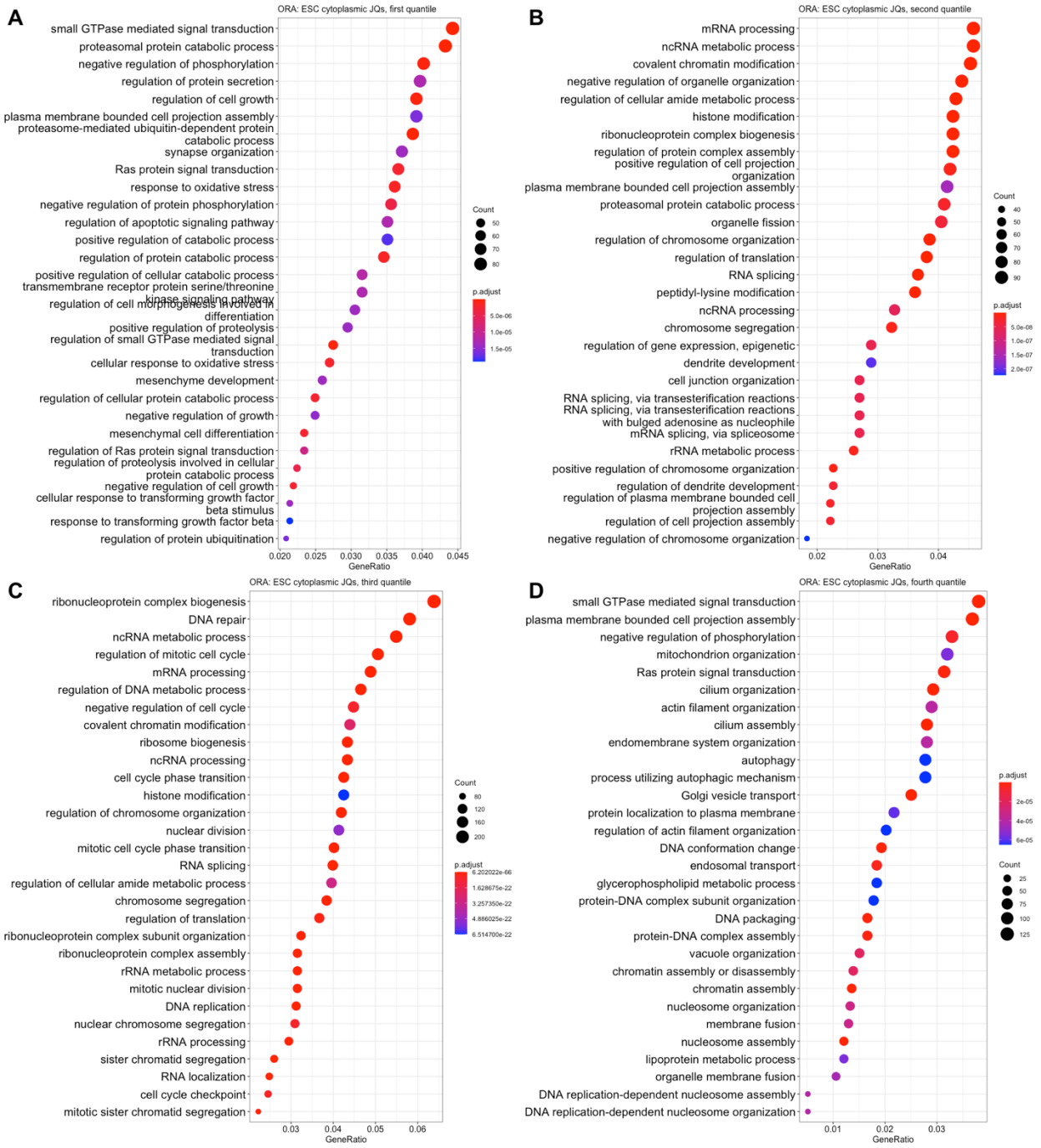


Figure 50: **Over-representation analysis result on binned mRNA junction quotient distribution in ESC cytoplasm**

A) Top over-represented GO terms in the first junction quotient quantile include gene sets related to cell growth and neuronal development **B)** second quantile show similar results to first quantile, with enrichment of gene sets related to tissue development; gene sets related to ion transport are also over-represented **C)** third quantile show an abundance of terms related signal transduction as well as cell organization D) fourth quantile show an abundance of terms related to cell division and cell cycle

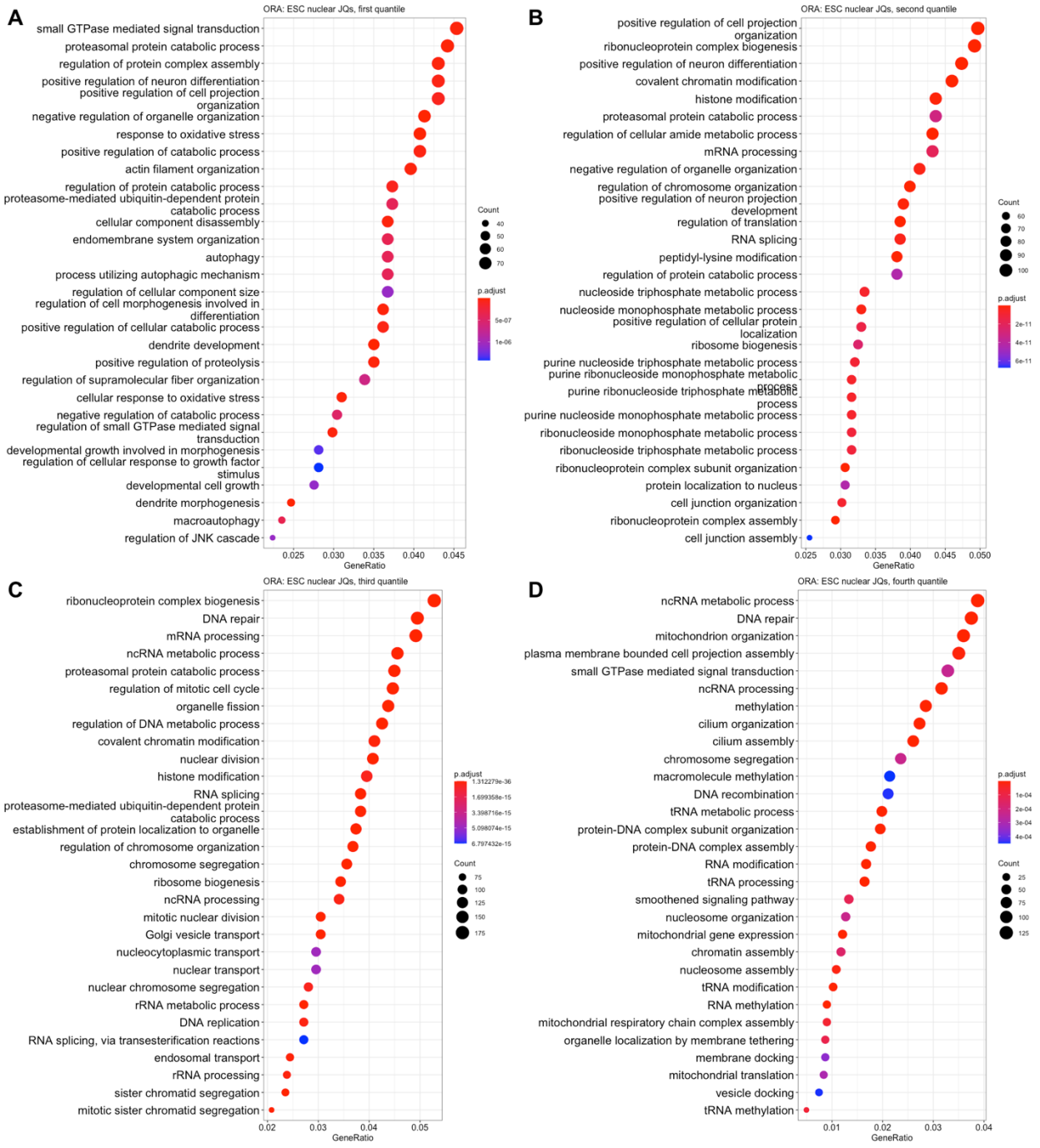


Figure 51: **Over-representation analysis result on binned mRNA junction quotient distribution in ESC nucleus**

A) Top over-represented GO terms in the first junction quotient quantile include gene sets related to cell differentiation and growth **B)** second quantile show enrichment of terms related to development and ion transport **C)** third quantile show an abundance of terms related to signal transduction as well as biosynthetic processes; D) fourth quantile show an enrichment of genes related to cell cycle and DNA/RNA processes

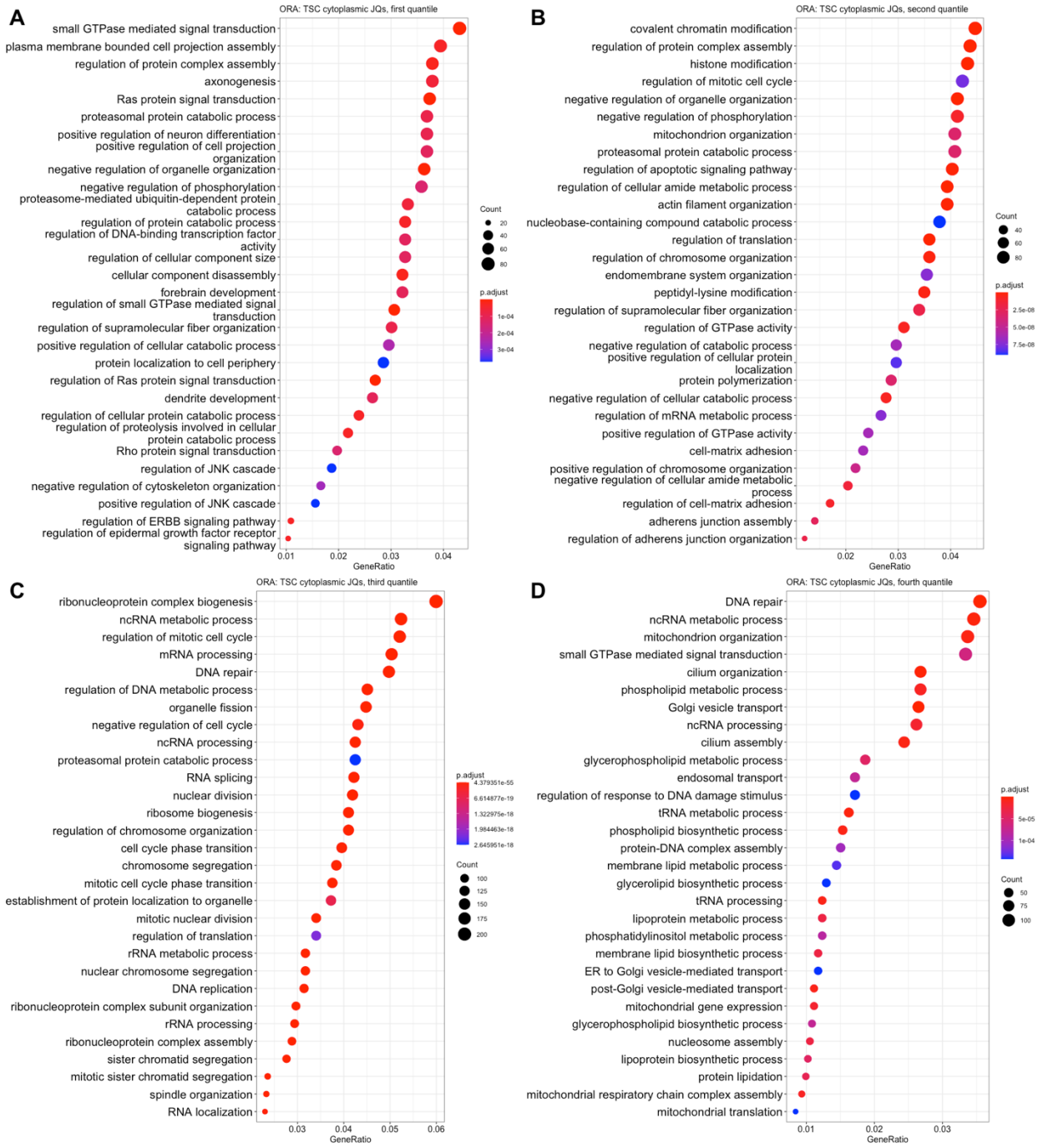


Figure 52: **Over-representation analysis result on binned mRNA junction quotient distribution in TSC cytoplasm**

A) GO enrichment in first quantile show similar results in ESC data; terms related to neural development are over-represented; terms related to signaling pathways are also enriched; **B)** second quantile show similar results to first quantile, with enrichment of terms related to system development, as well as terms related to metabolic processes **C)** third quantile show an abundance of terms related to RNA and DNA processes, as well as ribosomal processes; D) fourth quantile show an abundance terms related to cell structure, as well as cell signaling

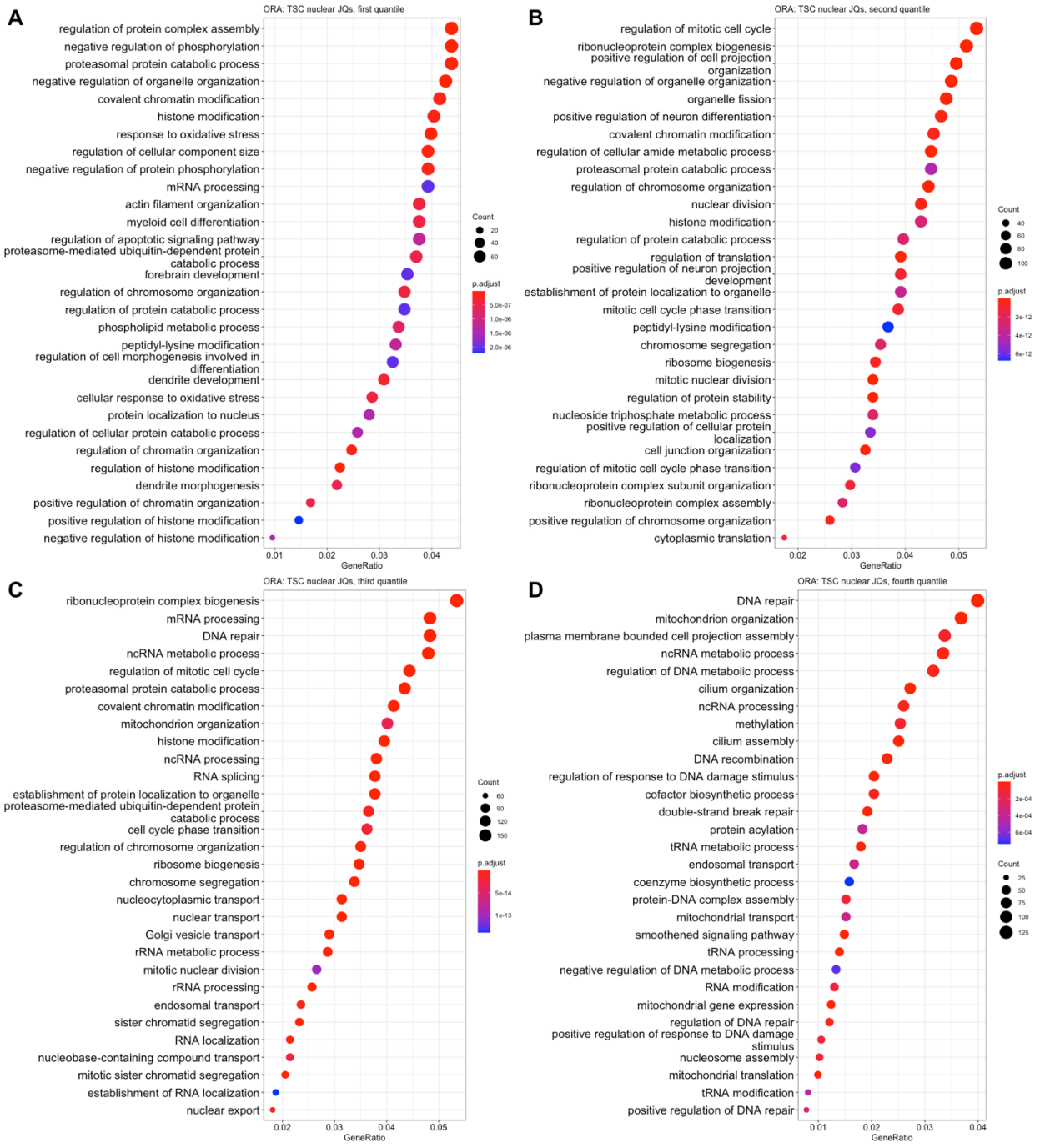


Figure 53: **Over-representation analysis result on binned mRNA junction quotient distribution in TSC nucleus**

A) Top over-represented GO terms in the first junction quotient quantile include gene sets related regulation of development; **B)** second quantile show enrichment of terms related to development and metabolic processes **C)** third quantile show an abundance of terms related to RNA and DNA processes, as well as ribosomal processes; D) fourth quantile show an enrichment of terms related to cell division and RNA processes

The role of Clk kinases in particular are of interest due to their relationship with nuclear detained retained introns; a study by Boutz *et al* showed that both Clk1 and Clk4 themselves in mouse ESCs show intron retaining behaviour in the nucleus and splicing of these introns due to heat shock or osmotic stress leads to alteration of other nuclear detained intron retaining transcripts. This suggests Clk role in finetuning the expression level of transcripts via post transcriptional splicing control. **Figure 54** shows in terms of JQs, both Clk1 and Clk4 mRNAs in the subcellular fractions of ESCs and TSCs show some degree of PIR, with Clk4 showing calculated JQ above the sample median.

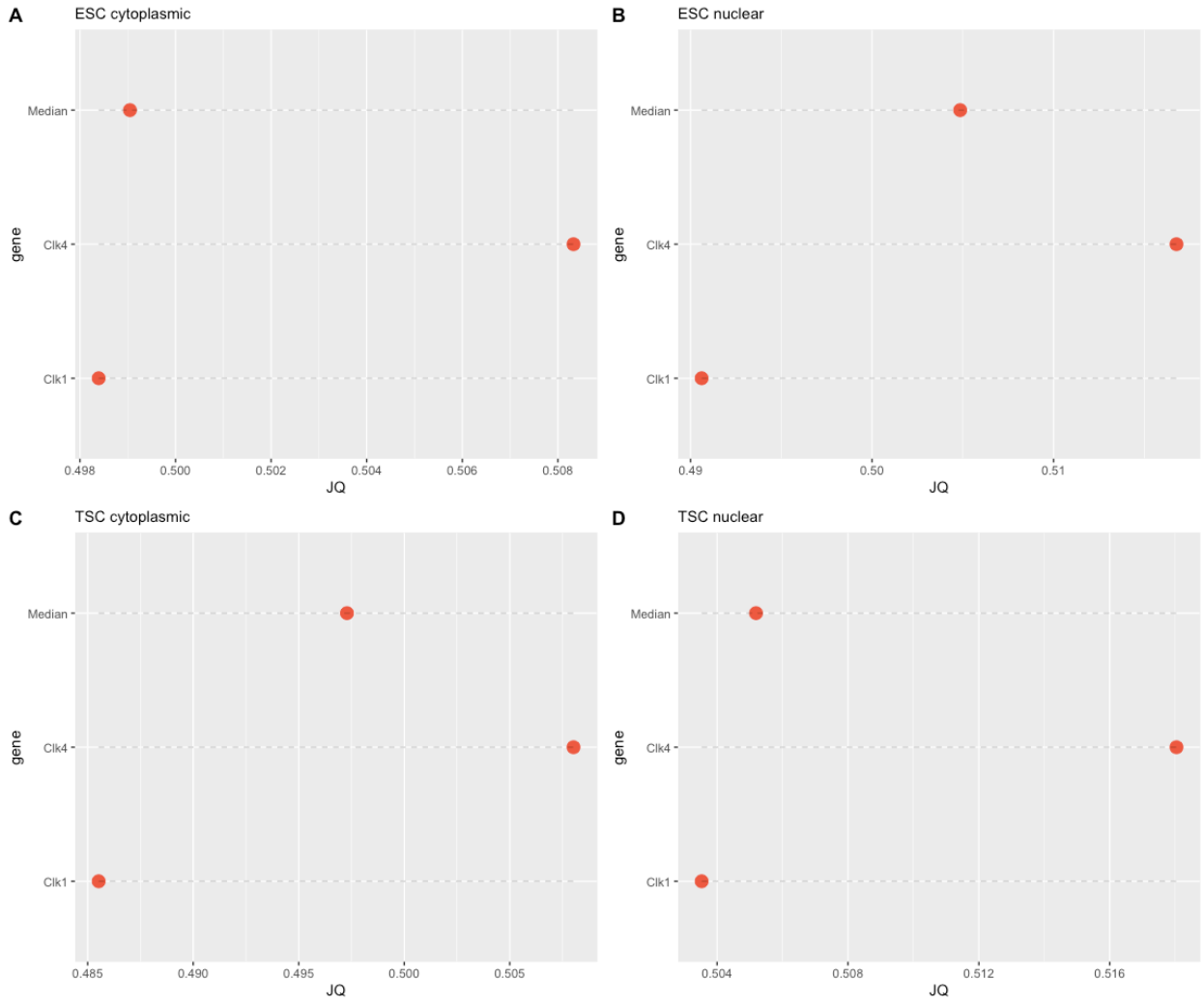


Figure 54: **Junction quotients for Clk1 and Clk4 kinases**

Both Clk1 and Clk4 show some degree of intron retention as per the calculated PIR, with Clk4 showing junction quotient value higher than the median value; the intron retaining behaviour of Clk transcripts have been documented to be involved in signal transduction and processing of other intron retaining transcripts in the nucleus.

3.3 Discussion

3.3.1 Normalization for cytoplasmic-nuclear mass imbalance is required

A complication often overlooked in literature is the problem of cell equivalence. Results from RNA concentration measurements with Qubit suggest that when normalized for unit cell amount, the two subcellular fractions yield unequal amounts of RNA (**Table 2**). The consequence of this is that in comparisons of RNA expression between cytoplasmic and nuclear fractions, a normalization step is required to account for this RNA imbalance.

An appropriate normalization step is required in every RNA-seq experiment in order to infer direct relationships between read counts across multiple samples. The consequence of a suitable normalization, which Evans *et al* claim in their extensive review of normalization methods, is that differentially expressed genes across sample groups should have normalized read counts whose differences can be attributed to true differences in RNA content per cell[106]. Non-differentially expressed genes, on the other hand, should have comparable read counts across the sample groups. The authors note achieving such conditions is affected by multiple factors in an RNA-seq experiment, such as sequencing depth and within-sample variations in gene lengths and GC content.

Within-sample variations, in particular variations due to differences in gene lengths, is due to the relationship between number of mapped reads and increasing gene length. A longer gene leads to a higher number of mapped read fragments, which leads to an inflated read count that cannot be attributed solely to the abundance of the gene itself[107]. In order to account for this length bias, normalization methods such as RPKM and FPKM transformations are used to scale each read count to the gene feature length as well as the library size. Such within-sample corrections, however, are not required when performing differential expression within the same genes across samples[108]. Therefore for the purpose of the current RNA-seq experiment, RPKM and FPKM transformations were not considered.

In their review of normalization methods, Evan *et al* outline three main modes of read count correction: 1) normalization by library size, 2) normalization by distribution, and 3) normalization

by spike-in controls[104]. In a review of synthetic spike-in controls in RNA-seq experiments, however, Risso *et al* found that obtained individual read counts of the spike-in controls showed high variability compared to their expected counts and as well, variations in library preparations affected the controls differently than the bulk of the genes[109]. The efficacy of using external controls as the sole normalization method in RNA-seq has been brought into question in literature[110-112]. For that reason, normalization at the feature count level was considered for the current study.

Normalization by distribution, in the case of Trimmed Means of the M-values (TMM) method used by edgeR, shifts the distribution of read counts by trimming relative to the fold changes and absolute expression values of a reference sample[113]. Evans *et al* notes that even though quantile normalization methods does well in equalizing distribution of read counts, a symmetric differential expression must exist across the samples for the normalization to be valid[106]. This means that quantile normalization assumes equal number of up- and down-regulated genes in differential expression, which cannot be made in the case of cytoplasmic-nuclear fraction samples. In typical whole-cell comparisons in RNA-seq, the dynamic localization of mRNA transcripts from the cytoplasm to the nucleus is entirely contained within each whole-cell data; as such, considerations in within-cell dynamics can effectively be 'canceled out' as such process is occurring in both samples. This is not the case in comparison of fractions, as subcellular differences will be preserved and reflected in differential expression. As such, it is disingenuous to make the assumption that the differential expression profile is symmetric.

Another important factor in normalization is the total amount of mRNA per cell. Quantile normalizations such as TMM or Upper Quartile method used by DESeq performs well in situations where amount of mRNA per cell is unequal, but fails to account for asymmetry in differential expression as outlined above[114,115]. Normalization by library size can handle asymmetry, but assumes equal mRNA per cell across the sample groups[106]. This assumption is not valid in RNA-seq with fractions, as seen in the total RNA measurements in the two fractions (**Table 2**). This problem of unequal RNA per cell equivalent unit in fractionated RNA-seq is often overlooked in literature; the violation of equal RNA assumption prior to library size normalization can lead to an increased likelihood of both type I and type II error in identifying differentially expressed genes.

In order to address this mass imbalance, read counts must be scaled in the subcellular fractions to equalize the effective mRNA per cell equivalence so that library size normalization can be used.

The method of cytoplasmic-nuclear fraction normalization presented in **Section 3.1.1** uses precisely measured RNA concentration from obtained cell fraction samples. Cell equivalence is preserved as approximately 40 million cells were used in the cell fractionation protocol across all technical samples and each aliquot was stored at equal volumes prior to measurement of RNA concentration. Scaling the read counts of one of the fraction read counts relative to the other effectively equalizes RNA per cell metric and allows for normalization by library size without violating its core assumption.

NOIseq generated PCA plots of corrected read counts show that this method of scaling counts has preserved the relatedness in the data, such that cytoplasmic and nuclear fraction samples cluster well within themselves. Transformation to lcpm values with edgeR then normalized read counts by library size prior to differential expression. This method of adjusting RNA content per cell value across the subcellular fractions and subsequently normalizing for library size allows for analysis without biases caused by unequal total RNA population, while acknowledging that asymmetric differential expression may exist. Indeed, in a subsequent analysis with simulated data, Evans *et al* found that normalization by library size performed well when mRNA per cell were equal across sample groups despite asymmetry in differential expression[106].

3.3.2 Differential expression profile at the exon level show nuclear enrichment

Literature in microarray and RNA-seq experiments pervasively use whole-cell RNA content to determine differential expression of mRNAs. Trask *et al*, however, in a study of cytoplasmic fraction and whole-cell RNA-seq data, challenged that in studies of steady-state mRNA population, the contribution of nuclear RNA is not negligible[116]. The authors found that differential expression profile of the cytoplasmic fraction discovered differentially expressed mRNAs which would not have been found in the analysis using whole-cell data. This finding suggests that in order to effectively capture the gene expression profile, cytoplasmic and nuclear RNA population should be examined as two separate fractions.

The role of nuclear RNAs in influencing the gene expression profile is supported in literature outlining the lifecycle of a given transcript[95]. After transcription and polyadenylation, the rate of nuclear export and cytoplasmic turnover affects the amount of RNAs detected in the cytoplasmic and nuclear fractions. Furthermore, literature has previously shown nuclear role of retaining RNA transcripts not needed for immediate translation until rapid export under physiological stress or response to stimuli[46]. Finally, the nuclear RNA exosome has been shown to affect gene expression via regulation of RNA maturation and degradation[117]. As such, it is no surprise that in a study of three human cell lines, Solnestam *et al* found that 400 to 1400 out of 15000 genes showed differential expression between cytoplasmic fraction and total RNA samples[118].

The nuclear presence of mRNAs is reflected in the differential expression profile of the two subcellular fractions in **Figure 13**, where a total of 6614 and 6450 mRNAs are down-regulated in the cytoplasmic fraction relative to the nuclear fraction in ESCs and TSCs, respectively (Benjamini-Hochberg adjusted p-value < 0.05). A rather surprising observation in the differential expression profile is the larger population of mRNAs down-regulated in the cytoplasm, suggesting a larger proportion of mRNAs – counting at the exon level – were found to be up-regulated in the nucleus. In a study of nuclear enriched RNAs, Halpern *et al* found that species detected in the nuclear fraction predominantly consisted of lncRNAs, hyper-edited dsRNAs, and incompletely spliced mRNAs[119] – a result consistent with literature evidence that shows mature, fully spliced mRNAs are predominantly found in the cytoplasm. It is important to note, however, that the differential expression profile in **Figure 13** does not aim to only encapsulate polyadenylated mRNAs; poly(A) selection was not used in the preparation of the sequencing library in order to capture a more complete lifecycle of mRNAs. The consequence of this may be that the number of differentially expressed mRNAs in the nuclear fraction is inflated by the processes associated with the nuclear lifetime of mRNAs – namely transcription, post-transcriptional processing, and nuclear degradation. As such, the expression profile in **Figure 13** does not necessarily reflect the steady state cytoplasmic population of mRNAs, but instead incorporates the nuclear contributions as well.

The inflated number of nuclear enriched mRNAs may also be attributed to hyperactive transcriptional activity; in a review of hyperactive transcription and stem cell biology, Percharde *et al* suggest that hyperactive transcription activity is a pervasive mechanism in mediating cell fate

transitions and early development[120]. Indeed, the authors found that such coordinated amplification of the transcriptome is seen throughout development such in the case of somatic stem cells, hematopoietic stem cells, and ESCs in the peri-implantation epiblast. The authors hypothesize that this amplification of nascent transcripts is due to the biosynthetic demands of rapidly proliferating cells, as in the example of ESCs. Furthermore, literature evidence suggests pluripotent stem cells exhibit permissive, open chromatin structure due to the interaction with chromatin regulators required for lineage-specific gene silencing, as well as due to overall elevated levels transcriptional activation[121]. This association of transcriptional hyperactivation and chromatin structure reiterates the requirement of stem cells in early lineage specification – as in the case of ESCs and TSCs – to exist in hyperactivated transcriptional state. Such hyperactivation of the transcriptome leads to a global elevation of nascent transcripts present, which may manifest in the inflated population of nuclear detected mRNAs.

As such, the large number of differentially expressed mRNAs in the nuclear fraction in both ESCs and TSCs may be related to their ability to retain pluripotency. The global amplification of the transcriptome associated with self-renewal and gene silencing of the embryonic and extraembryonic lineages in early development may indeed be reflected in the differential expression profile of the two subcellular fractions in ESCs and TSCs.

The contribution of nuclear retained mature mRNAs to the overall differential expression profile may also be significant; in a study of polyadenylated mRNAs in the nucleus, Halpern *et al* identified a wide range of nuclear-retained mature, fully spliced mRNAs[119]. Such retained mRNAs were shown to reside in the nucleus for a longer period of time than their lifetime in the cytoplasm prior to turnover. This result suggests that nuclear retention as a phenomenon has a non-negligible impact on the overall population of mRNAs in the cell, and thus may be reflected in the differential expression profile in **Figure 13**.

3.3.3 The rate of transcription may be correlated with differential expression

The differential expression profile across the cytoplasmic and nuclear fractions gives insight into the localization behaviour of mRNAs. The accumulation of mRNAs detected in either fraction, as Chen *et al* claims in their study of nucleocytoplasmic dynamics in mRNAs, is affected by three

kinetic events: 1) transcription, 2) nuclear export, and 3) cytoplasmic decay. These three cellular processes then must in return influence the differential expression profile in **Figure 13**.

In their study, Chen *et al* performed a cell fractionation coupled to deep sequencing in a time-series using *Drosophila* Kc167 cells[95]. Then the authors used mathematical modeling using first-order kinetic rate law to fit mRNA expression values with respect to change in time, thereby calculating the rate constant for each of the three processes. As such time-dependent experiment design was not used in the current study, it is not possible to employ a similar model to deduce the rate constants. However, the authors showed a negative correlation between the rate constant of transcription and genomic features such as transcript length, intron length, and the number of exons. In the differential expression profile result in **Figure 13**, it turns out the intron length and the number of exons per gene both show negative correlation with the fold-change (**Figure 31** and **32**). This result suggests that factors that affect the efficiency of transcription has an influence in the accumulation of mRNAs in the cytoplasm as well. Indeed, Chen *et al* showed that the rate of transcription has the biggest influence on the variance of overall steady state mRNA population in the cytoplasm[95].

Thus in differentially expressed mRNAs in the cytoplasmic fraction, the extent of their expression is dependent on the efficiency of transcription. The hyperactive transcriptional activity characteristic of ESCs, then, should lead to a large population of steady state mRNAs detected in the cytoplasmic fraction. For example, in a study of transcriptional activity in ESCs, Efroni *et al* found elevated levels of total RNA and mRNAs. This observation, however, is not as apparent in **Figure 13** due to the population of nuclear mRNAs. The elevated levels of total RNA as shown by Efroni *et al* may be correlated to an influx of nascent mRNA – which in turn inflate the number of nuclear detected mRNAs[94]. Regardless, the negative correlation between factors affecting transcription and the fold-change in differentially expressed cytoplasmic mRNAs (**Figure 31** and **32**) show that a total RNA survey in subcellular fractions has the ability to model transcriptional behaviour. It is possible then, to make inferences on genes and their subcellular dynamics by their transcriptional behaviour.

3.3.4 Genes related to cell cycle and the chromatin are differentially expressed

In a study of single-cell sequencing data in hematopoietic stem cells, Tsang *et al* showed that genes related to cell cycle and nuclear division are predominantly over-represented[122]. This finding is echoed in another single-cell sequencing study, where Kolodziejczyk *et al* showed an enrichment of cell cycle related genes in mouse ESCs grown in 2i and a2i cell culture conditions[123]. The abundance of gene sets related to such function can be attributed to the self-renewal and proliferative characteristic in each cell lines, which in turn can be related to the concept of hyperactive transcriptional activity.

As the transcriptional activity is modeled by the differential expression profile, profiling the over-represented gene sets should yield mRNAs that are subject to elevated transcriptional activity. Indeed, in gene set enrichment analysis of the fractionated differential expression data, gene sets related to cell division and cell cycle are predominantly over-represented in the nuclear fraction of ESCs (**Figure 18**). This result suggests the elevated transcription of genes related to cell cycle and cell division results in an accumulation of mRNA in the nucleus. This supports the theory that a large population of nuclear detected mRNAs can be attributed to a hyperactive transcriptome, and as well, that the kinetic rate of transcription influences the overall subcellular dynamics of mRNAs. Furthermore, Efroni *et al* showed that transcriptional hyperactivity in ESCs also features a disproportionate amount of genes related to chromatin-remodeling[94]; this is due to the maintenance of a permissive chromatin in ESCs, which allows for increased nascent transcript output. Indeed, **Figure 18** shows an enrichment of gene ontology terms related to chromatin organization as well as the regulation of chromatin.

3.3.5 Self-renewal and proliferation depends on metabolic processes

A review of the relationship between metabolism and pluripotency by Tsogtbaatar *et al* found that pluripotent stem cells exhibit a high demand for anabolic and catabolic processes to maintain their self-renewal and proliferation[124]. This high biosynthetic demand can be attributed to the sheer energy needed for the propagation of cell content and genetic materials during cell cycles. However, as the authors suggest in their review, the link between metabolism and stem cell biology also exist in the context of epigenetics.

Epigenetic regulation of gene expression, such as via histone methylation and acetylation, is responsible for regulating pluripotency and differentiation in cell fate decisions. For example, in a study of isogenic mouse ESCs, Juan *et al* showed that the distribution of histone H3 lysine 27 di- and tri-methylation (H3K27Me2 and H3K27Me3) was responsible for exerting regulatory control on cell lineage specification and transcriptional programs[125]. In another review of epigenetics and TSCs, Kohan-Ghadr *et al* showed that acetylation of histones H2A and H2B reduced invasiveness in mouse TSCs and retained multipotency[126]. Such findings reiterate the interplay of epigenetic modifications and the transcriptome in cell fate decisions.

Tsogtbaatar *et al*, in their review, suggests that the link between metabolism and cell fate decisions is mediated by epigenetics. Indeed, acetylation has been shown to be dependent on cellular levels of acetyl-CoA, which is derived from glycolysis[124]; production of S-adenosylmethionine (SAM) from tetrahydrofolate in carbon cycles was shown to contribute to methylation via SAM acting as a methyl donor[127] lysine-specific demethylase I (LSD1) was characterized to require flavin adenine dinucleotide (FAD) – which is produced via the citric acid cycle – in order to catalyze demethylation of histones[128].

This behaviour of elevated metabolism then must be reflected in a transcriptomic survey of ESCs and TSCs. Indeed, in the mRNA differential expression profile of both ESCs' and TSCs' subcellular fractions, gene sets related to both metabolism of macromolecules and chromatin assembly are over-represented in the cytoplasmic fraction (**Figures 18** and **19**). Gene sets related to metabolic processes, in particular, are predominantly enriched in large proportions. This result suggests mRNAs associated with metabolic function may be subject to higher rate of nuclear export and possibly protein synthesis. The detection of differentially expressed mRNAs in the cytoplasmic fraction may not necessarily suggest higher rate of transcription for the select genes, but may be related to a higher demand of efficient translation and turnover. It is possible that mRNAs related to metabolism are preferentially localized to the cytoplasm due to the high demand of proliferating cells to require glycolysis – due to the anabolic demand as well as for production of metabolites for maintenance of stemness. Indeed, Yu *et al* was able to show Sox2, Oct4, and Nanog binding sites at the GLUT1 enhancer, which led to an elevation of GLUT1 expression and subsequent glycolytic flux in ESCs[80]. Additionally, a review of stem cell metabolism and cell fate

control by Folmes *et al* showed that the expression of hexokinase and pyruvate kinase – enzymes directly involved in glycolysis – are under direct transcriptional control of Oct4[129]. The dependence of mouse ESCs and amino acid metabolism has also been outlined in literature; for example, Alexander *et al* found that threonine dehydrogenase expression is elevated in mouse ESCs and shows decreased expression during differentiation – which suggests threonine metabolism is also crucial in self-renewal[130]. Therefore, it is clear that cell fate decisions require significant metabolic demands – which, as Folmes *et al* suggests in their review, enables stem cell population to prioritize metabolic pathways in order to meet such demands[129].

The differential expression profile between two subcellular fractions was able to uncover the subcellular localization behaviour of mRNAs related to processes directly involved in cell fate decisions. Both ESCs and TSCs employ gene regulatory networks linked with metabolism, and causes an accumulation of related mRNAs at the site of protein synthesis.

3.3.6 Genes related to cell adhesions are differentially expressed in TSC fractions

The link between metabolic processes and epigenetic regulations is also present in trophoblast populations. A study of epigenetic marks in TSCs by Senner *et al* found that DNA methylation patterns on the loci of lineage-specific transcription factors play a crucial role in determining cell fate[131]. For example, the hypomethylation of the Elf5 promoter in TSCs led to an establishment of a positive feedback system between Elf5, Cdx2, and Eomes and maintenance of trophoblast lineage. The accumulation of gene sets related to metabolic processes in the cytoplasmic fraction of TSCs (**Figure 19**) suggest their role in maintenance meeting such biosynthetic demands.

The gene set enrichment analysis result for the nuclear fraction in TSCs show results deviating from that in ESCs (**Figures 18** and **19**). In TSCs, an accumulation of mRNAs related to cell structure, cell junctions, and cell shape is observed. Indeed, a heatmap of gene expression in lcpm for cell adhesion molecules (i.e., cadherins, Epcam) show up-regulation in the nuclear fractions compared to cytoplasmic fractions (**Figure 22**). Literature in TSCs have documented the role of cell adhesion molecules in maintenance of trophoblast lineage; in a study of the mouse placenta, Ueno *et al* were able to isolate multipotent precursors with high levels of Epcam expression[132]. These cells were able to differentiate into all lineages within the trophoblast (i.e.,

syncytiotrophoblast layers and trophoblast giant cells). Cdh3 was identified as a trophoectoderm marker, whose expression is induced by an undifferentiated TSC marker in Sox21[133]. Foxd3 – a transcription factor required in trophoblast lineage – has shown to regulate the expression of Cdh7 during epithelial-mesenchymal transition[134]. The role of cell-cell adhesion in tissue patterning is well documented in literature, in particular in context of placenta development and acquisition of invasiveness; as Latos *et al* notes in their review, cell-to-cell fusion and communication is pertinent in syncytiotrophoblast formation and eventual establishment of maternal-fetal exchange surface[133].

The gene set enrichment related to cell adhesion and integrity in the nuclear fractions suggest a few possibilities; this nuclear accumulation of mRNAs may be attributed to intentional retention of mature transcripts awaiting physiological stress or stimuli. It may also be attributed to a high rate of transcription and subsequent accumulation prior to nuclear export. Another possibility is the detection of pre-mRNAs or mRNAs designated for nuclear degradation. Regardless, the asymmetric distribution of mRNAs across subcellular fractions suggest functional differences in gene regulation; the cytoplasmic accumulation of mRNAs related to metabolism suggested a relationship with the high demand of proliferating cells and possibly a higher rate of protein synthesis and turnover. The nuclear accumulation of mRNAs may be functionally related to developmental timing and nuclear retention – such that mRNAs related to cell division and cell cycle in ESCs and cell-cell adhesion in TSCs are subject to a modulatory mechanism. The asymmetric accumulation of functional gene sets give insight into the modulatory pathways governed by subcellular localization.

3.3.7 Genes related to immune function show intron-retaining behaviour in TSCs

Quantification and differential expression of exon-intron junctions in subcellular fractions allows for identification of intron-retaining mRNAs and their localization behaviour. In literature of intron retention, the fate of intron retaining transcripts is largely binary: nonsense mediated decay (NMD) in the cytoplasm due to a premature termination codon – as a means of gene expression regulation - or nuclear retention as *detained introns*[46, 50, 56]. However, in the cytoplasm, intron retaining transcripts have also been shown lead to protein diversity via production of isoforms, as well as activation or suppression of translational efficiency[54,55]. In some cases, intron retaining transcripts

can also avoid NMD and be stabilized. As such, intron retention serves as a complex means of gene regulation machinery.

The differential expression profile in TSCs show deviations from the result from exon data; an enrichment of gene sets related to immune processes, as well as genes related to chromatin and nucleosome organization in the cytoplasmic fraction is observed (**Figure 40**). This result suggests an accumulation of intron-retaining mRNA transcripts related to immune and chromatin function in the cytoplasm of TSCs. The immune properties of the trophoblast lineage in development have been documented in literature; TSCs *in vitro* have been shown to express interferons (IFNs) in response to viral stimuli – an observation not seen in ESCs. In fact, Fendereski *et al* suggest that ESCs interact with the trophoectoderm in the blastocyst via paracrine signaling to gain IFN innate response and antiviral protection[136]. Another study by Aikawa *et al* found that the addition of poly I:C – an immunostimulant – induced the production of IFN-β in TSCs, therein which an exposure to IFN-β in ESCs led to the expression of antiviral genes[137]. Such literature findings suggest the trophoectoderm role in immune function within the blastocyst. Corroborating with the results from **Figure 40** then, it can be inferred that TSCs respond to viral infections and immune function may be modulated via intron retention events. Indeed, literature findings indicate intron retention as a major component in gene regulation machinery during development, as well as in response to stress and disease.

3.3.8 Intron retention may modulate lineage-specific processes in development

A study of intron retention in mouse ESCs by Boutz *et al* found a functional enrichment of intron retaining gene sets whose products' cellular abundance must be tightly controlled[56]. A survey of nuclear-localized intron retaining transcripts in the same study found that gene sets related to DNA damage response showed high levels of detainment in the nucleus[48,49]. These genes were shown to undergo rapid splicing and expression upon induced DNA damage, suggesting the role of nuclear intron retention as a means of cell response to stress and stimulus. Indeed, gene set enrichment analysis of exon-intron junctions show an enrichment of genes related to DNA repair and response to endoplasmic reticulum in the nuclear fraction of ESCs (**Figure 39**). Gene sets related to wound healing is also over-represented in the nuclear fraction of TSCs, as well as genes related to immune response, reiterating the possible role of nuclear intron retention and cell response (**Figure 40**).

Another observation from **Figure 39** and **40** is the differences between ESC and TSC data; the over-representation of gene sets related to neural development in ESCs' nuclear fraction and gene sets related to circulatory system and cell-cell adhesion in TSCs' nuclear fraction suggests intron retention may influence lineage-specific processes in development as well. Indeed, intron retention has been shown to be involved in both granulocyte and B-cell differentiation[48,49,138]. As well, the role intron retention in cell fate decisions by modulating mRNA levels has also been well documented in literature of hematopoietic stem cells[139]. Findings presented in **Figure 39** and **40** suggest possible modulation of gene expression via intron retention in the embryonic and trophoectoderm lineages as well.

3.3.9 Accumulation of metabolism related mRNAs persist in spliced data

The common theme across the differential expression profile in exon data, exon-intron junction (i.e., unspliced) data, and exon-exon junction (i.e., spliced) data in ESCs is the accumulation of gene sets related to metabolism in the cytoplasmic fraction relative to the nuclear fraction. This result suggest mRNAs related to metabolic function may exhibit alternative splicing events (ASEs); the link between ASEs and metabolism has been investigated in literature, and a review from Biamonti *et al* suggests ASEs can modulate the transcriptome depending on the cells' demands; an example is the ability for cells to finetune their metabolic function in response to stress or external stimuli via expression of pyruvate kinase isoforms[140]. As such, it is possible that ESCs' heterogeneity in cytoplasmic transcript population in terms of splicing variants is related to their metabolic plasticity and biosynthetic demands.

Furthermore, literature in transcription kinetics suggest genes related to metabolic function show highest kinetic rates of transcription, which is corroborated by the consistent accumulation of such mRNAs in the cytoplasmic fraction in ESCs[95]. This finding, along with the negative correlation between cytoplasmic enrichment and intron length (**Figure 31**), establishes the relationship between transcription rates, intron length, and the level of differential expression in the cytoplasmic fraction.

3.3.10 Intron retention is related to steady state population of mRNA

Percent intron retention (PIR) in literature is defined as the ratio of unspliced junctions and the sum of unspliced and spliced junctions per gene[56]. A calculation of PIR from junction expression in lcpm and junction quotients in the two subcellular fractions then allows for an estimation of intron retention in each compartment. The divergent fate of intron retaining transcripts in either compartment, however, makes it difficult to make overarching conclusions; for example, intron retaining mRNAs in the cytoplasm may be subject to gene regulation via NMD or translated to functional isoforms. In either scenario, however, intron retention serves an important regulatory function in development; studies show both alternative splicing for protein diversity and gene modulation via NMD plays an important role in cell differentiation and organogenesis.

Indeed, intron retention has been shown to modulate global mRNA expression levels during development. A study of mouse ESCs by Braunschweig *et al* found that intron retention can lead to a suppression of the expression of genes not relevant for particular cell fate; such that genes with a high level of intron retention in neurons tended to be related to cell cycle, DNA repair, and pluripotency[56].

Interestingly, results from junction quotient (JQ) count data in ESC cytoplasmic fraction show an over-representation of gene sets related to cell cycle, nuclear division, and DNA repair in the third quantile (**Figure 50 D**). As each succeeding quantile represents a higher value of JQ and thus higher estimated PIR per gene, this result suggests such gene sets show a relatively high level of intron retention per gene.

These results, however, must be interpreted in context of overall gene expression and steady state mRNA levels. **Figure 50 A** showed that the level of mRNA expression in the cytoplasmic fraction shows a moderately positive correlation ($\rho = 0.47$) with JQs – suggesting mRNAs with a higher level of expression may also show a higher level of retained introns within the transcript. This suggests a higher expression of intron-retaining mRNAs related to the cell cycle and cell division in the cytoplasm versus over-represented gene sets in the first and second quantiles. This observation holds true in the differential expression profile between the two subcellular fractions using exon-intron junction count data, where gene sets related to cell cycle and cell division were up-regulated (**Figure 39**). This observation indicates that for mRNAs pertinent in maintenance of

ESC lineage as outlined before, such as those related to cell cycling as well as metabolic and biosynthetic processes, intron retaining behaviour may also lead to stabilization of the transcript.

The role of intron retention in regulation of cytoplasmic mRNA levels can indeed involve stabilization of transcript, as well as production of isoforms[48-50]. The consequence of a retained intron – whether an intron retaining transcript is likely to be degraded or stabilized – depends on the location of the retained intron within the transcript; a premature termination codon, for example, can be introduced and lead to NMD – as shown in Braunschweig *et al* study where neural progenitors suppressed the expression of pluripotency genes via intron retention and NMD[56,139,142]. As depletion of an mRNA nor a presence of a premature termination codon cannot be elucidated in the current method of JQs and read counts, the role of intron retention in suppression of mRNA governing cell fate cannot be deduced. Corroborating with findings from Braunschweig *et al*, however, it can be inferred that intron retention can affect the population of cytoplasmic mRNA by both depletion and stabilization of transcripts – such that intron retaining mRNAs can still be detected in the cytoplasm and such mRNAs are likely to code for proteins related to maintenance of ESC identity.

3.3.11 Intron retaining transcripts can be nuclear detained

Over-representation analysis of JQ data in the cytoplasmic fraction of ESCs showed that up-regulated genes (i.e., genes related to cell division and cell cycle, as well as metabolic processes) show intron-retaining behaviour. This result, in addition to literature evidence in intron retaining transcripts and NMD, suggests intron retention is a pervasive mechanism that influences a large portion of the transcriptome. Indeed, Braunschweig *et al*, in their study of 40 human and mouse tissues, concludes that intron retention occurs frequently and affects as much as three quarters of multi-exonic genes and their transcripts[56].

Another mechanism in which intron retaining transcripts can evade NMD is via nuclear retention; Jacobs *et al* suggests in their review that nuclear detained intron retaining transcripts are either degraded within the nucleus or stored awaiting signal-induced splicing and export[50]. Multiple sources in literature suggest the overall level of detectable intron retention is generally higher in

the nucleus relative to the cytoplasm – suggesting nuclear storage and NMD in the cytoplasm as the predominant fate of intron retaining transcripts in each compartment[48-50,56,142].

As evident in **Figure 38 A** and **C**, the population of retained introns is significantly larger in the nuclear fractions compared to the cytoplasmic fractions; the median PIR per gene is also higher in the nuclear fraction (**Figure 48**). This observation suggests both differential expression using exon-intron junction reads and calculation of JQs using junction counts across the two subcellular fractions was able to describe the effect of intron retention on mRNA subcellular localization behaviour.

Furthermore, gene set over-representation analysis of exon-intron junction reads showed that genes related to cell-specific functions show intron retaining behaviour in the nuclear fractions (**Figures 39** and **40**); this suggests nuclear retention of intron retaining transcripts are under cell-specific regulation control. This finding is supported in literature, where Boutz *et al* found a substantial number of nuclear detained intron retaining transcripts expressed in only one or two cell types in a survey of adult tissues and ESCs[56]. Therefore, nuclear detained introns allow cells to control the overall expression levels of the transcriptome via post-transcriptional, regulated splicing of physiologically relevant transcripts.

Clk kinases such as Clk1 and Clk4 have been documented to play a role in regulated splicing of nuclear detained introns[143]. Clk mRNAs have been shown to exhibit intron retaining behaviour in the nucleus of mouse ESCs and the inhibition of Clk kinase activity led to splicing of Clk transcripts[144]. This led to subsequent alteration of intron retention events in ~10% of total ~3000 observed retained intron events, which suggests splicing of retained intron in the Clk mRNA modulates post-transcriptional splicing[56]. Indeed, in both JQ data of ESC and TSC subcellular fractions, both Clk1 and Clk4 mRNAs showed observable levels of PIR in both the cytoplasmic and nuclear fractions (**Figure 58**). Furthermore, gene sets related to RNA splicing and mRNA processing were found to be over-represented in the third JQ quantile in all sample groups (**Figures 50 – 53**), indicating mRNAs coding for RNA splicing and processing factors themselves may also be under regulatory control via intron retention.

The splicing of introns in Clk transcripts has been shown to be related to external cues such as heat shock, osmotic stress, and various physiological stress responses such as response to insulin[143,144]. This suggests regulators of post-transcriptional splicing – and therefore, detainment of intron retaining transcripts in the nucleus – influences the transcriptome in response to external stimuli. In ESCs, genes sets related to 1) cell cycle, cell division and histone methylation, 2) ion transport, and 3) neural development may be subject to regulation via nuclear retention (**Figure 39**). In TSCs, such gene sets include 1) wound healing, chemotaxis, and inflammatory response, 2) cell-matrix adhesion and cell-cell adhesion, 3) circulatory system process and blood circulation, 4) ion transport, and 5) extracellular matrix organization (**Figure 40**). In a pool of nuclear mRNAs alone, in both ESCs and TSCs, mRNAs related to RNA splicing, RNA processing, as well as cell cycle and cell division – likely related to cell self-renewal and proliferation – predominate in terms of PIR, suggesting factors related to global cellular processes are also under regulatory control under intron retention and as well, that intron retention is a global phenomenon that affects a large portion of the transcriptome (**Figures 51** and **53**).

As such, differential expression of subcellular fractions using exon-intron junctions can reveal cytoplasmic-nuclear differences in intron retaining transcripts, whereas estimation of PIR using JQs in the two fractions reveal intron retaining transcripts that are prevalent in both compartments – which would be missed by differential expression.

3.3.12 Long noncoding RNAs are differentially expressed across subcellular fractions

The expression and subcellular localization of noncoding RNAs (ncRNAs) is another means of transcriptome regulation. For example, long-noncoding RNAs (lncRNAs) have been shown to exert regulatory control via multitude of ways, including: 1) regulation of chromatin state and histone modifications, 2) mediation of DNA-protein binding interactions, 3) sequestration of miRNAs, and 4) antisense interference of mRNAs[147]. As such, ncRNAs are involved in complex regulatory networks in both the cytoplasm and the nucleus. The differential expression profile of subcellular fractions then, reveals possible roles of ncRNAs in ESC and TSC maintenance.

Fico *et al*, in their extensive review of lncRNAs and ESCs, outlined the function of several lncRNAs in influencing pluripotency[148]; Meg3 is an imprinted lncRNA required for embryonic

development via interaction with the Polycomb repressive complex 2 (PRC2); Neat1 is a lncRNA required for formation of nuclear paraspeckles and organization remodeling; Malat1 has been shown to sequester miRNAs and its dysregulation leads to increased invasion and metastasis of multiple cancer cells; Lncenc1 has been shown to regulate the transcription of genes involved in glycolysis required for ESC self-renewal; Tuna forms a RNA-multiprotein complex to regulate the transcription of Nanog and Sox2; Evx1as is a *cis*-acting lncRNA that regulates the expression of Evx1 gene in order to promote ESC differentiation; finally, lncRNA Bvht directly binds to PRC2 during ESC differentiation towards cardiac cells.

LncRNAs such as Snhg3 and Gas5 has also been documented in literature in relation to mouse ESC self-renewal, via regulation of ESC markers such as Oct4, Nanog, and Sox2[98-101]. The localization of these lncRNAs remain unclear, as both lncRNAs have been detected in either compartment – suggesting there may exist both cytoplasmic and nuclear mode of action.

On the other hand, dysregulation of lncRNA H19 has been linked with impaired TGF- β signaling pathway and decreased trophoblast cell invasion and migration. Induced H19 expression has also been shown to induce trophoblast fate in mouse ESCs via expression of Cdx2.

The differential expression of such lncRNAs in ESC and TSC fractions shows up-regulation in the cytoplasmic fraction, despite known nuclear function of lncRNAs such as Neat1, Meg3, Bvht, and Lncenc1 (**Figure 17**). Despite literature evidence that lncRNA function is reflective of their subcellular localization, recent findings suggest the localization behaviour of lncRNAs is a dynamic process; for example, lncRNAs SNHG1 and NORAD in human colon cancer cells show both cytoplasmic and nuclear enrichment, but upon DNA damage, they are retained in the nucleus[149]. This suggests lncRNAs such as Bvht and Neat1, which have been shown to be associated with cell differentiation, may show cytoplasmic localization prior to cell fate commitment.

The cytoplasmic enrichment of Snhg3 and Gas5 suggest their role in maintenance of ESC self-renewal and regulatory networks with Oct4, Nanog, and Sox2 may primarily function in the cytoplasm. Interestingly, Snhg3 and Gas5 also show up-regulation in the cytoplasmic fraction in

TSCs, suggesting multiple roles of such lncRNAs. Indeed, SNHG3 in human has been shown to be involved in invasion and migration of various cancer cells[150]. Furthermore, GAS5 has been shown to regulate miR-21 expression, which in turn regulates invasion, migration, and proliferation of trophoblast cell lines in human[151].

An interesting finding is the cytoplasmic enrichment of Lncenc1 in both ESCs and TSCs, which has been shown to preserve ESC self-renewal in mouse via interaction with RNA-binding proteins to regulate transcription of genes related to glycolysis. Most widely studied function of Lncenc1 in literature is its involvement in the nucleus to regulate transcription, but a qRT-PCR panel of Lncenc1 in mouse ESCs by Sun *et al* found Lncenc1 localization in both the cytoplasm and the nucleus, with higher levels in the cytoplasm[100]. The cytoplasmic function of Lncenc1 remains unclear, yet evidence from the differential expression profile (**Figure 17**) suggest its function may not be confined to just ESCs.

Interestingly, Malat1 is up-regulated in the cytoplasmic fraction in ESCs, but does not show subcellular differential expression in TSCs; Malat1 has been shown to sequester the expression of miR-34a and as expected, when ESC and TSC samples are directly compared in differential expression, miR-34a was found to be up-regulated in TSCs (**Section 4.2.5**). MiR-34a has been shown to target mRNAs involved in both Ras and Rap1 signaling pathway, which is involved in formation of cell-cell adhesions and junctions, as well as in cell migrations[152]; this suggests the expression of Malat1 is under regulatory control, which in turn regulates key miRNA-mRNA processes in TSCs. The roles of miRNAs in ESCs and TSCs will be discussed further in **Chapter 4.**

3.3.13 Spliced ncRNAs are significantly up-regulated in the cytoplasmic fraction

The comparison between subcellular differential expression using exon-intron junction reads (**Section 3.2.11**) and exon-exon junction reads (**Section 3.2.13**) show stark differences in ncRNA data. The differential expression profile in **Figure 38** and **43** suggest unspliced ncRNAs are up-regulated in the nucleus and spliced ncRNAs in the cytoplasm. While intuitive, this behaviour was not observed in mRNA data, where both unspliced and spliced reads showed up-regulation in the nuclear fraction. This low number of spliced mRNAs in the cytoplasmic fraction was attributed to

hyperactive transcription and turnover of mRNAs in the cytoplasm - due to the high biosynthetic demand required for cell self-renewal and proliferation - as well as the possibility of mature mRNA accumulation in the nucleus due to the rate of nuclear export (**Section 3.3.2, 3.3.3,** and **3.3.4**). The low number of intron-retaining transcripts was attributed to gene regulatory control via nonsense-mediated decay (**Section 3.3.10**). On the other hand, the high number of exon-intron junctions in the nuclear fraction may be due to nuclear detained introns (**Section 3.3.11**), as well as detection of pre-mRNA due to the caveat that the differential expression profile encompasses non-polyadenylated RNA as well.

This result suggests multiexonic ncRNAs are subject to an alternative set of kinetic processes which govern their subcellular localization. The enrichment of spliced junction reads in ncRNAs in the cytoplasmic fraction suggest the significant population of processed ncRNAs persist in the cytoplasm, rather than being subject to a high rate of turnover. This observation is in agreement with findings in literature, where in particular, long non-coding RNAs (lncRNAs) were found to be more abundant in the cytoplasm. The nuclear role of lncRNAs have been well-documented, however, ranging from mediating protein-DNA interactions to directly regulating chromatin modifications[147]. This suggest that despite known nuclear functions, processed lncRNAs may yet be more abundant in the cytoplasm in absolute number of transcripts. This was shown in the differential expression of select lncRNAs associated with pluripotency, where lncRNAs with nuclear function were found to be up-regulated in the cytoplasmic fraction in both ESCs and TSCs when using exon count data (**Figure 17**; discussed in **Section 3.3.12**). This localization behaviour of spliced lncRNAs with nuclear role also suggest the possibility of subcellular shuttling between the two compartments. Indeed, despite the abundance of lncRNAs in the cytoplasm, a large-scale study of lncRNA localization using RNA *in situ* fluorescence hybridization found that the presence of lncRNAs within the cell is ubiquitous[153].

Classes of small ncRNAs that do not have splice variants – such as snoRNAs and miRNAs – were profiled using exon-intron count proportions (**Section 3.2.10**) in **Figure 31** and **32**. SnoRNAs and miRNAs were detected in both the cytoplasmic and nuclear fractions in ESCs and TSCs, reiterating the ubiquitous nature of ncRNAs within the cell.

Chapter 4

4 Differential Expression Profile of microRNAs and Identification of mRNA Targets

4.1 Methods

4.1.1 *Measurement of miRNA concentration for normalization*

For cytoplasmic-nuclear count balance correction as per RNA-seq data in **Section 3.1.1**, concentration of RNA content across the fractionated samples was calculated prior to sequencing. In the case of small RNA-seq, miRNA concentration was measured using the Qubit Fluorometer in an analogous setup to **Section 2.1.4**. All measurements, as before, were obtained in four technical replicates and averaged.

4.1.2 Generation of miRNA count data

Bowtie and cutadapt was used in a Linux environment running Ubuntu 18.04.

Generated single-end small-RNA-sequencing reads from the Illumina sequencing run were subject to quality control with FastQC and adapter trimming with cutadapt as per **Section 2.1.5.** All trimmed reads below 10 nucleotides in length were discarded prior to alignment (`-m 10`). Trimmed reads were then aligned to the mm10 genome using Bowtie with default parameters and mm10 indices from Bowtie's database. Bowtie was used in alignment of small RNA-seq reads instead of HISAT2 or Bowtie2 due to the absence of expected gapped reads in the data.

4.1.3 Normalization of miRNA counts for cytoplasmic-nuclear fractions

featureCounts was used in a Linux environment running Ubuntu 18.04.

As per **Section 2.1.6**, featureCounts output for cytoplasmic and nuclear counts for miRNAs were corrected using a scalar factor based on measured miRNA concentrations from **Section 4.1.1**. As before, significant figures for scalar factors were carried over from the fluorometer measurements.

$$\alpha x + \beta y = \gamma z$$

$$x + 1.13y = 0.490z$$

(ESCs)

$$x + 0.647y = 0.177z$$

(TSCs)

4.1.4 Differential expression analysis of miRNAs

After alignment with Bowtie and adapter trimming with cutadapt, mm10 miRNA annotations were downloaded from miRbase and count tables were generated using featureCounts. Only the reads that corresponded to the mature, single-stranded form of miRNAs were considered in generation of the count data. Count data were subject to quality control with NOISeq as before. Resulting count tables then underwent edgeR-voom-limma differential expression workflow using the same design matrices as **Section 3.1.1**. Lists of differentially expressed miRNAs across cytoplasmic-nuclear and ESC-TSC pairwise comparisons were generated.

4.1.5 miRNA-target network analysis with MEINTURNET

MEINTURNET was accessed and used on Firefox Browser 76.0.1 on macOS Mojave.

A web-based interactive tool MEINTURNET was used to generate network properties of provided miRNAs and their target RNAs. TargetScan database was used to generate lists of validated target RNAs. Differentially expressed miRNAs in each cell fraction (i.e., in ESC cytoplasmic vs. nuclear and TSC cytoplasmic vs. nuclear pairwise comparisons) from **Section 4.1.4** were used as input for MEINTURNET. Default parameters were used in generation of network maps and KEGG database was used for functional pathway enrichment analysis of target RNAs.

4.1.6 miRNA-target analysis in cell to cell comparisons

miRNA-target network and functional enrichment analysis with MEINTURNET was repeated with differentially expressed miRNAs from ESC to TSC pairwise comparisons instead (**Section 4.1.4**). As before, default parameters were used with TargetScan and KEGG databases.

4.1.7 miRNA in situ hybridization assay for miR-15b

Wild type mouse ESCs and TSCs were thawed and maintained on respective media and MEFs (**Section 2.1.1**) for two subsequent passages before being split onto gelatinized plastic (ESCs) and MEF-conditioned media (TSCs). At the next passage day, both cell lines were moved to 4-well plates and maintained to achieve 80% confluency. Cells were fixed with 4% paraformaldehyde (diluted in PBS). Each 4-well plate consisted of a single well assigned to be a negative control (i.e., no hybridization probe) and three other wells to be hybridized. ViewRNA miRNA ISH Cell Assay Kit (Affymetrix) was used with a Type 1 MicroRNA Probe Set (Affymetrix) for miR-15b for *in situ* hybridization and signal amplification in fixed ESCs and TSCs. Fast Red Tablets (Sigma) were used for the immunohistochemical staining and samples were subsequently stained with VECTASHIELD-DAPI mounting medium (Vector Labs) on cover slips. Cell imaging was done on Zeiss Spinning Disk Confocal Microscope at excitation wavelengths of 530 nm (Fast Red) and 360 nm (DAPI). Processing of confocal images was performed with Zeiss ZEN imaging processing software running on Windows 10 and ImageJ running under macOS Mojave.

4.1.8 miRNA in situ hybridization assay for miR-6240

Wild type mouse TSCs were thawed, maintained, split onto 4-well plates, and fixed as before. Using the same plate setup and the same hybridization kit as **Section 4.1.2**, fixed TSCs were hybridized with Type 1 MicroRNA Probe Set (Affymetrix) for miR-6240. Cells were stained as before and imaged on a spinning disk confocal microscope.

4.1.9 Generation of validated target list of differentially expressed miRNAs

Lists of differentially expressed miRNAs from **Section 4.1.1** were imported to R Studio and used as input for following analysis. Using R package MultiMiR and its built-in function `get_multimir()`, all validated target RNAs from curated databases such as miRTarBase,

Tarbase, and miRecords were generated. Resulting list of target RNAs were then checked for overlap with the mRNA differential expression profile from **Section 3.1.1**.

4.2 Results

4.2.1 Measurement of miRNA concentration

Accurate measurement of miRNA concentration in all sample groups was necessary for downstream count correction using concentration derived scalar factors. Qubit Fluorometer measurements were used due to its lower limit of detection, low amount of sample needed, as well as high specificity for RNA species of interest (i.e., in this case, miRNAs) due to the nature of fluorescent molecule binding mechanism. Overall, miRNA concentration was measured to be higher in the cytoplasmic fraction ESCs and higher in the nuclear fraction in TSCs; this behaviour was shown in total RNA measurements as well (**Section 2.2.2**).

Table 7: **Measured miRNA concentrations in fractionated cell lysates**

	TSC samples (ng/µL)	ESC samples (ng/µL)
Cytoplasmic fraction	20.2	57.8
Nuclear fraction	31.2	51.2
Whole cell lysate	114	118

4.2.2 Processing of miRNA sequencing data

Prior to differential expression, as with RNA-seq data in **Section 3.2.2**, unsupervised clustering of featureCounts output was performed to ensure predictable clustering and separation of sample groups. As before, sample groups are expected to cluster well within the experimental condition of interest – in this case, by cell type (i.e., ESC and TSC) and cell fraction (i.e., cytoplasmic and nuclear).

NOISeq generated PCA plots show similar results from **Section 3.2.2**, such that samples cluster by both cell type and fraction, where first dimension separates ESCs and TSCs and second dimension separates cytoplasmic and nuclear fraction samples (**Figure 55, A & B**). This result

suggests, as expected and as seen before, the largest source of variance in data is related to the differences in phenotype such that samples co-vary by their cell type. Furthermore, the clustering of cytoplasmic and nuclear fraction count data suggest the count correction using RNA-concentration derived scalar factor (**4.1.3**) does not introduce potential biases or artifacts that affect co-variance in the current dataset.

Boxplots in the NOISeq quality control report visualize the distribution and shape of the count data. In all sample groups in the current analysis, the read count distribution show skewness towards higher values with tailing – such that variance increases with increasing value (**Figure 55**, **C & D**). This behaviour is adjusted for in downstream analysis with voom as in the case of RNA-seq data.

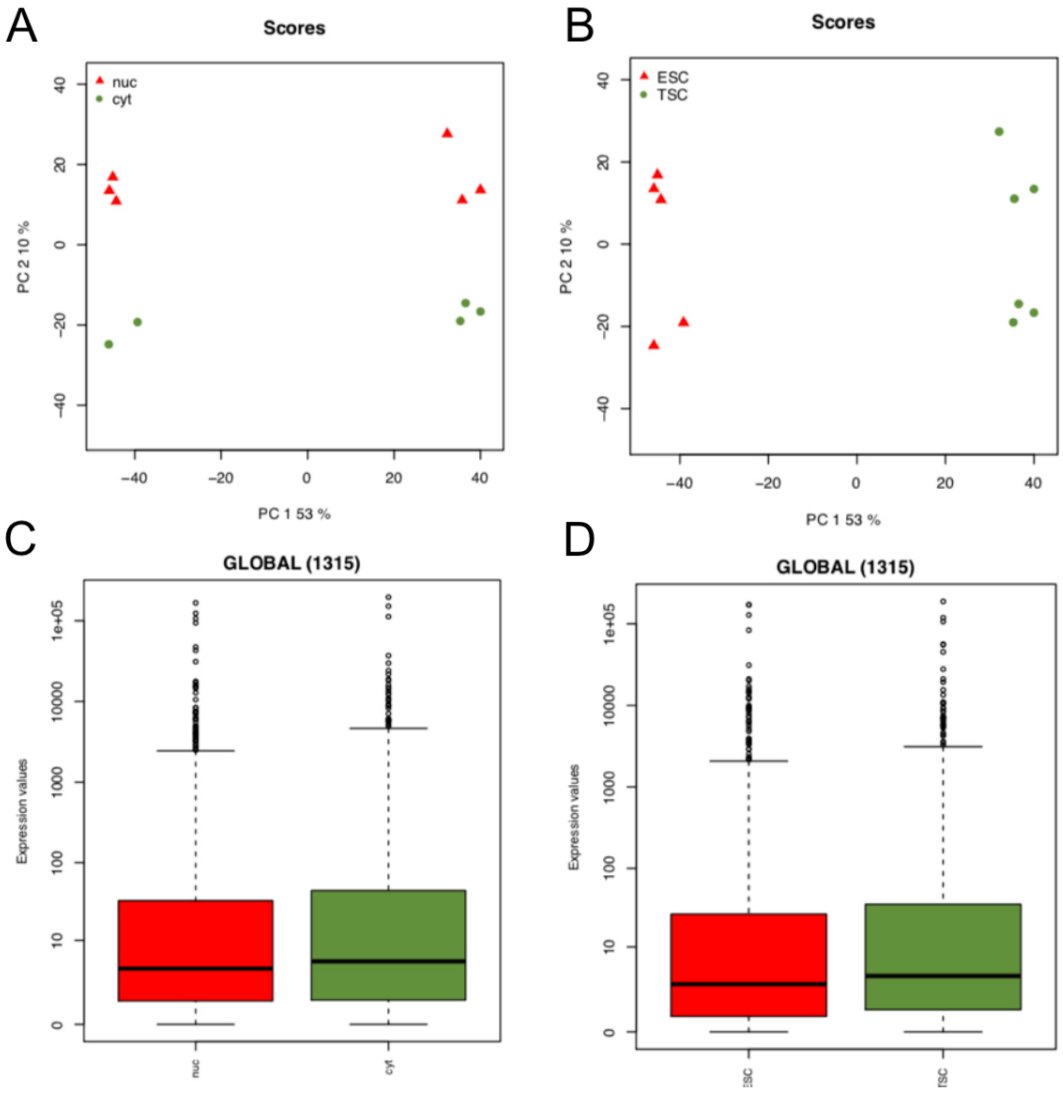


Figure 55: **NOIseq quality control plots of fractionated small RNA sequencing count data**

A) Similarly to NOISeq quality control with RNA-seq counts, PCA plots show clustering of sample groups by cell fraction in the second dimension; **B)** sample groups are clustered by cell type in the first dimension; **C)** boxplots show the read count data distribution show skewness towards high values as expected in both cell fraction sample groups; **D)** similar results show when sample groups are separated by cell types

4.2.3 miRNA differential expression profile

edgeR-voom-limma pipeline illustrated in **Section 3.2.3** with RNA-seq data was used to generate a differential expression profile of single-stranded miRNAs (limma's `decideTests()` with `adjust.method = "fdr"` and `p.value = 0.05`), which revealed 72 differentially

expressed miRNAs in ESCs and 28 miRNAs in TSCs – in pairwise comparisons between the two subcellular fractions (**Figure 56**).

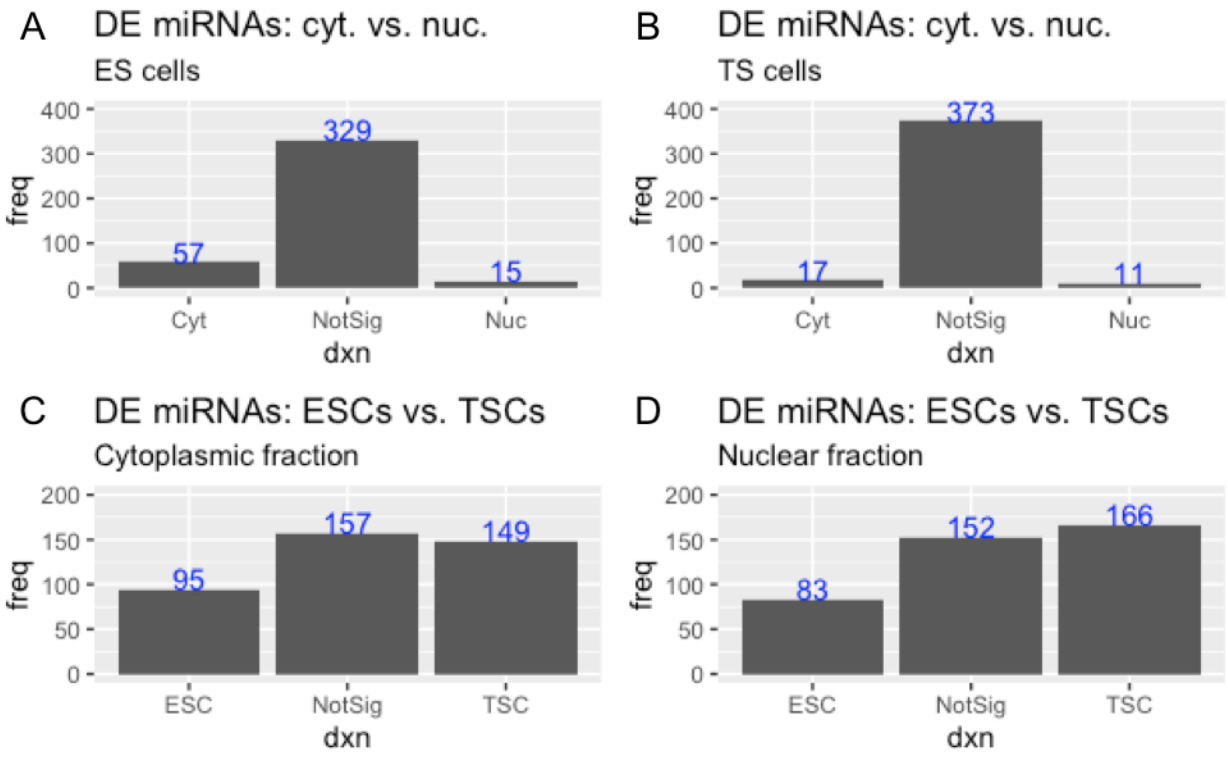


Figure 56: **Differential expression profile of miRNAs**

A) Differential expression analysis reveals a total of 72 miRNAs differentially expressed in cytoplasmic-nuclear fraction comparison in ESCs; **B)** total of 28 differentially expressed miRNAs are found in TSC cytoplasmic-nuclear comparison; **C)** cell-to-cell comparisons show larger population of miRNAs up-regulated in TSCs versus ESCs in cytoplasmic fraction samples; **D)** similar results is shown in nuclear fraction samples.

An annotated volcano plot as presented below reveal the identity of the differentially expressed miRNAs in each pairwise comparison (**Figure 57 & 58**). The miRNAs passing both the adjusted p-value threshold ($p < 0.05$) and fold difference threshold (log-fold-change > |2|) are colored in red and annotated. Using these miRNAs for downstream mRNA target network analysis provides insight into the regulatory mechanisms in cellular processes with subcellular specificity. Prioritizing differentially expressed miRNAs with a significant number of validated target mRNAs reveals biological processes under miRNA regulatory control. Such type of an analysis in ESCs

and TSCs then will reveal processes related to embryonic and trophoblast cell lineage under miRNA control, as well as pertinent miRNAs involved.

Furthermore, due to advances in *in situ* assays to target and bind specific miRNAs, it is possible to visualize the subcellular localization of differentially expressed miRNAs. Such assays will be useful in validating the presence of miRNAs enriched in cellular compartments.

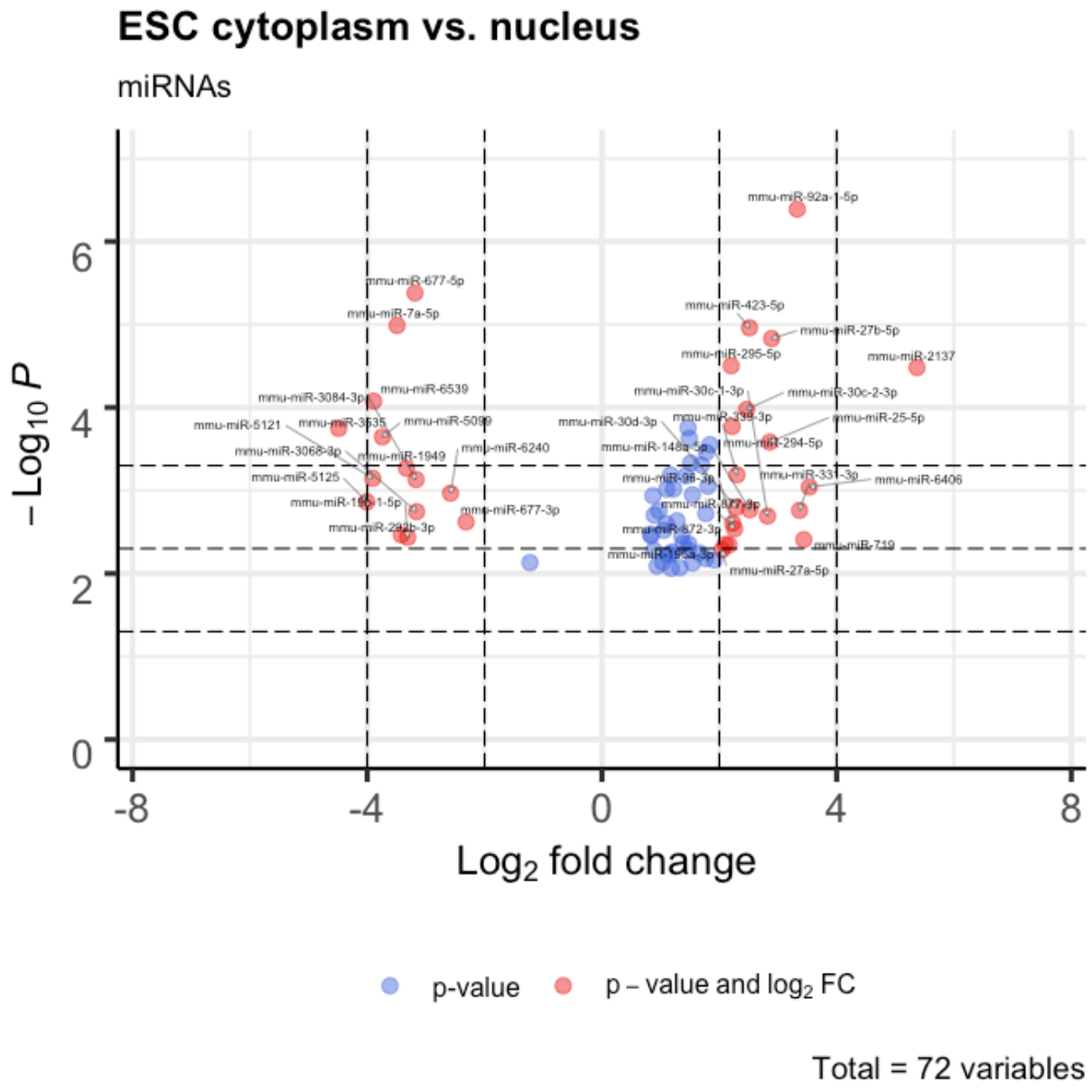


Figure 57: **Annotated volcano plot of miRNA differential expression profile in ESCs**

Volcano plot visualizes the 72 differentially expressed miRNAs in ESC cytoplasmic vs. nuclear comparison in context of log-fold-change and statistical significance; miRNAs which pass the threshold of adjusted p-value < 0.05 and log-fold-change greater than the absolute value of 2 are annotated and coloured in red.

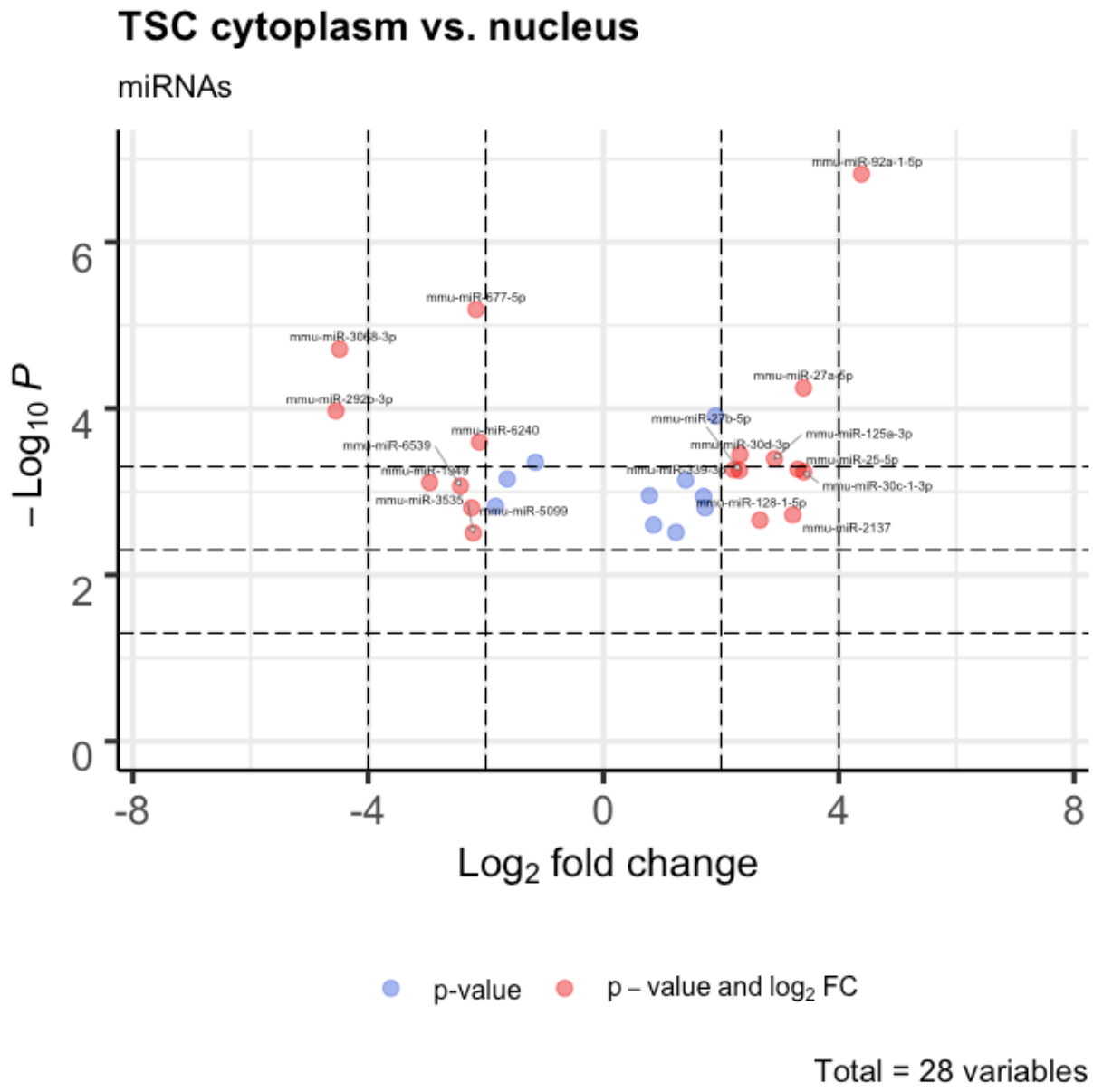


Figure 58: **Annotated volcano plot of miRNA differential expression profile in TSCs**

Volcano plot visualizes the 28 differentially expressed miRNAs in TSC cytoplasmic vs. nuclear comparison in context of log-fold-change and statistical significance; miRNAs which pass the threshold of adjusted p-value < 0.05 and log-fold-change greater than the absolute value of 2 are annotated and coloured in red.

4.2.4 miRNA enrichment network analysis using MIENTURNET

Using the lists of differentially expressed miRNAs in each cell fraction as input for MIENTURNT results in network graphs where vertices indicate miRNAs with edges connecting to mRNA targets from TargetScan database. The largest sample size belonged to up-regulated miRNAs in ESC cytoplasmic fraction (n = 57) which correspondingly led to the largest network graph. MiR-15a and miR-497a showed the largest number of mRNA targets (i.e., number of outward edges) in this sample group followed by miR-130 family. As miRNAs with large number of target transcripts have a greater influence in cellular outcomes, degree centrality in network graphs is a relevant metric. Furthermore, high inward degree centrality (i.e., number of inward edges) indicate mRNAs with higher number of miRNAs with modulatory control – this may suggest mRNAs related to collaborative miRNA mechanisms where multiple miRNAs regulate the expression of the same gene.

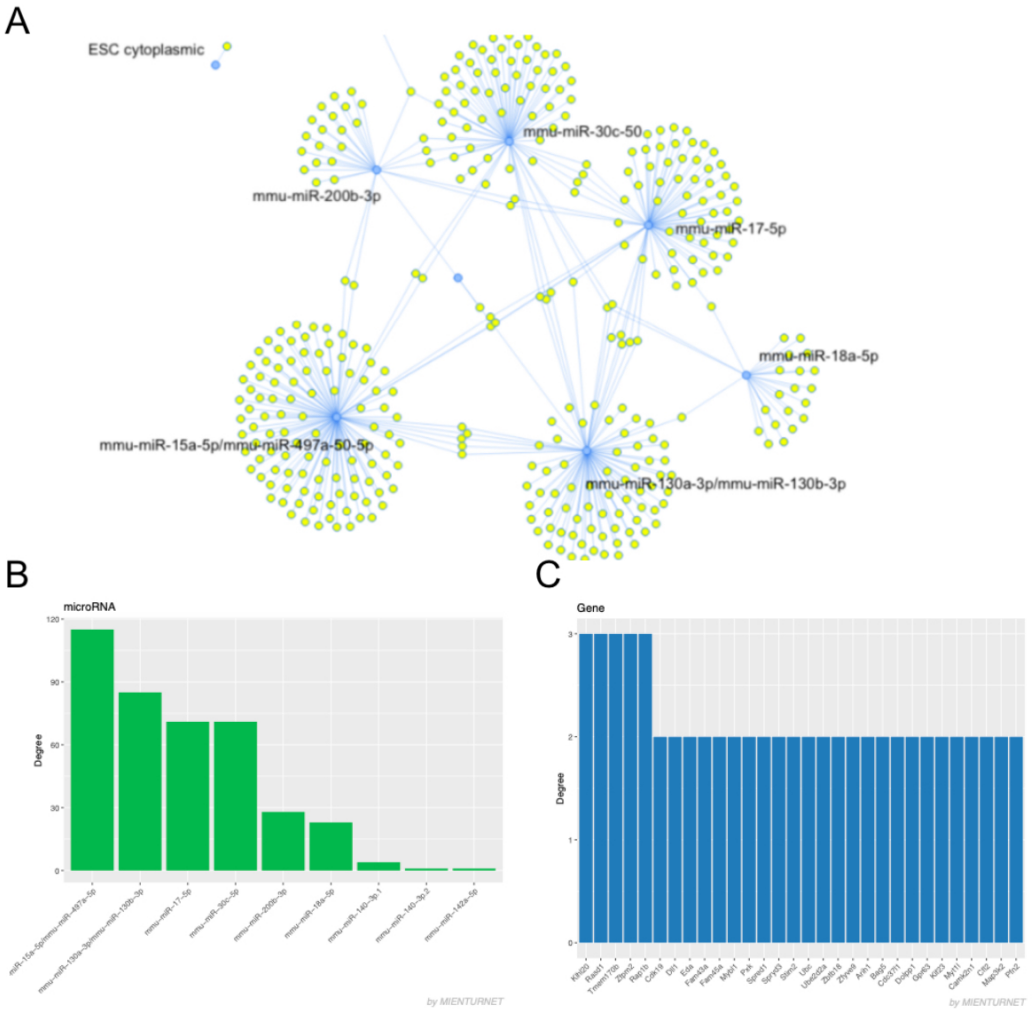


Figure 59: **Visualization of degree centrality of nodes in ESC cytoplasmic miRNA target network**

A) Network map of differentially expressed miRNAs in ESC cytoplasm and validated targets found by TargetScan; miRNAs with large number of outward edges are visualized with overlapping target genes; B) histogram showing the absolute number of mRNA targets of identified miRNAs; miRNAs with large outward centrality (i.e., high number of targets) may be more of biological significance; C) histogram showing the absolute number of miRNAs targeting each mRNA target; mRNA targets with large inward centrality (i.e., high number of interacting miRNAs) suggest a collaborative mechanism at which multiple miRNAs target a single gene.

KEGG analysis of validated target mRNAs from the network graph reveals cellular processes governed by miRNA regulatory control. In the data of cytoplasmic enriched miRNAs in ESCs, associated cellular processes largely include various signaling pathways – including terms such as VEGF signaling pathway, mTOR signaling pathway, TGF- β signaling pathway, and signaling pathways regulating pluripotency of stem cells. This suggests miRNAs present in the cytoplasm

of ESCs are largely intertwined with molecular signaling cascades involved in system development and cell growth and differentiation. Indeed, studies in neural stem cells support the influence of miRNAs in regulation of major components of mTOR and TGF- β signaling pathway in development. In the current analysis, this result suggest the possible influence of miRNAs such as miR-15a, miR-497a, and miR-130 family in developmental processes in mouse ESCs via interaction with molecular signaling pathways.

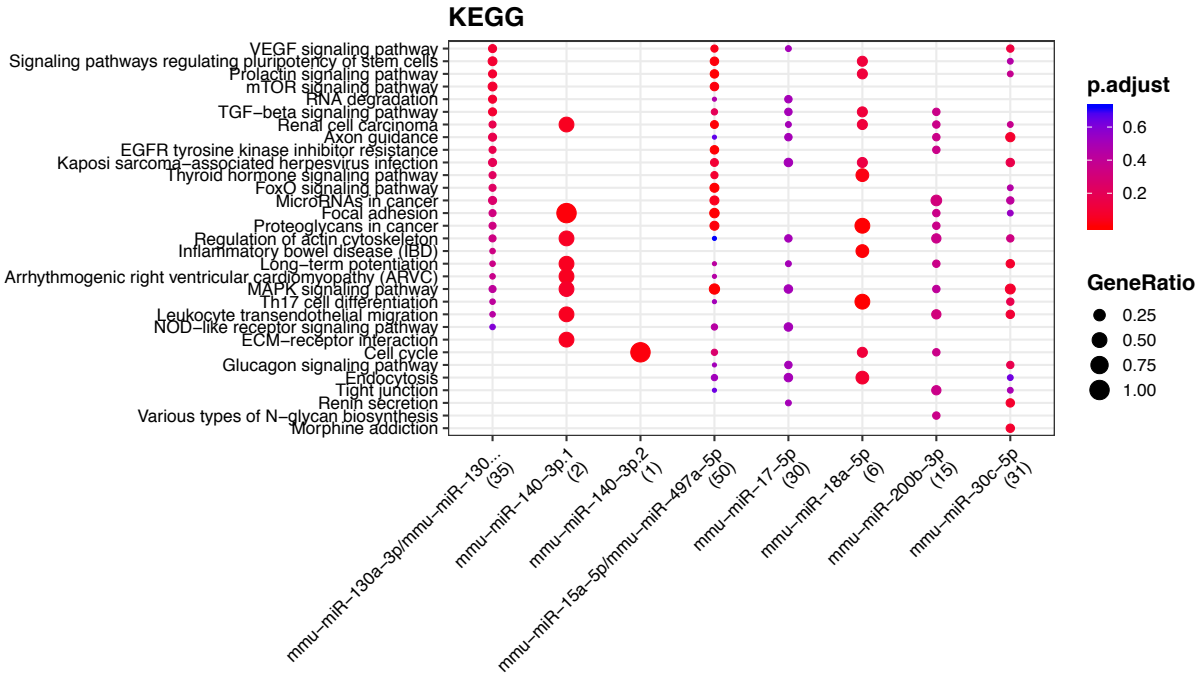


Figure 60: **KEGG pathway enrichment analysis result from ESC cytoplasmic miRNA targets**

KEGG pathway enrichment result suggests miRNAs up-regulated in the cytoplasmic fraction in ESCs play a role in functional pathways related to cell signaling pertinent in developmental processes.

The small sample size of differentially expressed miRNAs in the other three sample groups (i.e., ESC nuclear, TSC cytoplasmic and nuclear) did not allow for meaningful network analysis result with MEINTURNET. Over-representation test with either KEGG or GO database did not provide statistically significant results for the provided gene target list in all three sample groups. In terms of network maps, only two nuclear miRNAs appear in ESC and TSC samples with overlapping validated target mRNAs. In both sample groups, the identified miRNAs are miR-7a and miR-6539, suggesting such nuclear miRNAs may be involved in regulation of non-cell-specific processes.

Indeed, there is literature evidence that miR-7a in mouse targets components of the mTOR signaling pathway.

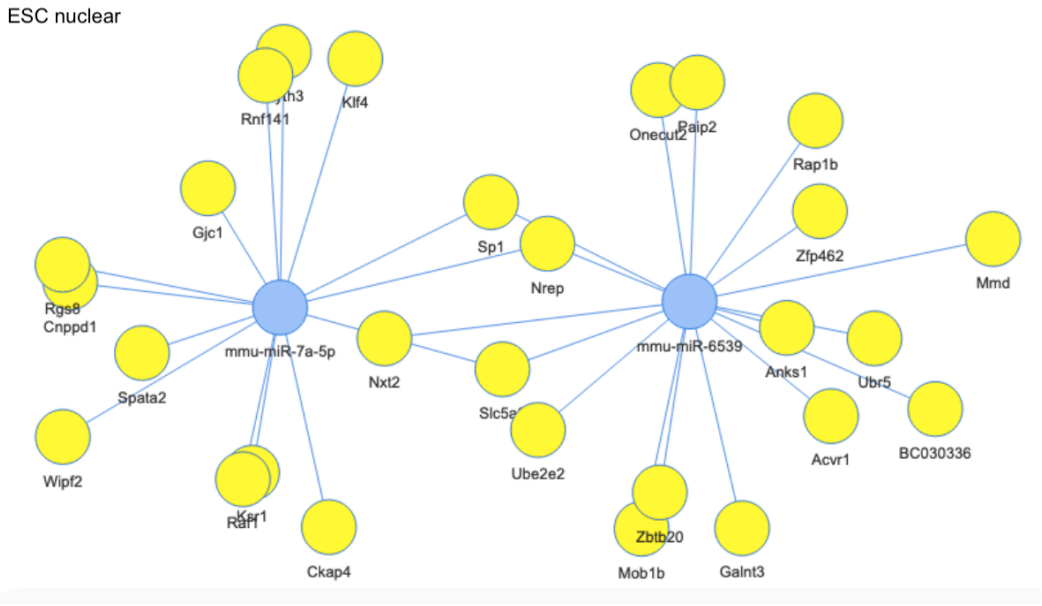


Figure 61: **Network map showing prioritization of ESC nuclear miRNAs and their targets**

Network analysis result on nuclear miRNAs in ESCs suggest only two miRNAs target multiple overlapping mRNAs according to TargetScan.

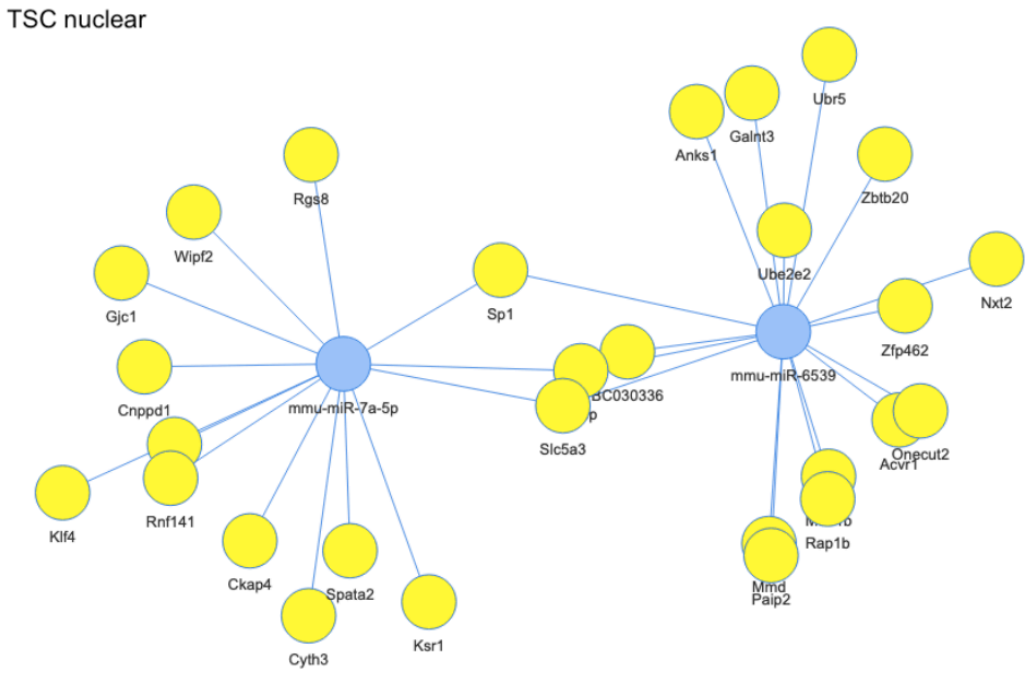


Figure 62: **Network map showing prioritization of ESC nuclear miRNAs and their targets**

Network analysis result on nuclear miRNAs in TSCs suggest only two miRNAs target multiple overlapping mRNAs according to TargetScan.

4.2.5 Functional gene set enrichment analysis of miR-7a and miR-677 targets

R package MultiMiR was used to generate target list of miRNAs differentially expressed in the nuclear fractions of ESCs and TSCs, without considerations in network similarity statistics employed by MIENTURNET. In this result, miR-7a and miR-677 showed the highest number of validated target mRNAs in both ESCs and TSCs (**Figure 63**).

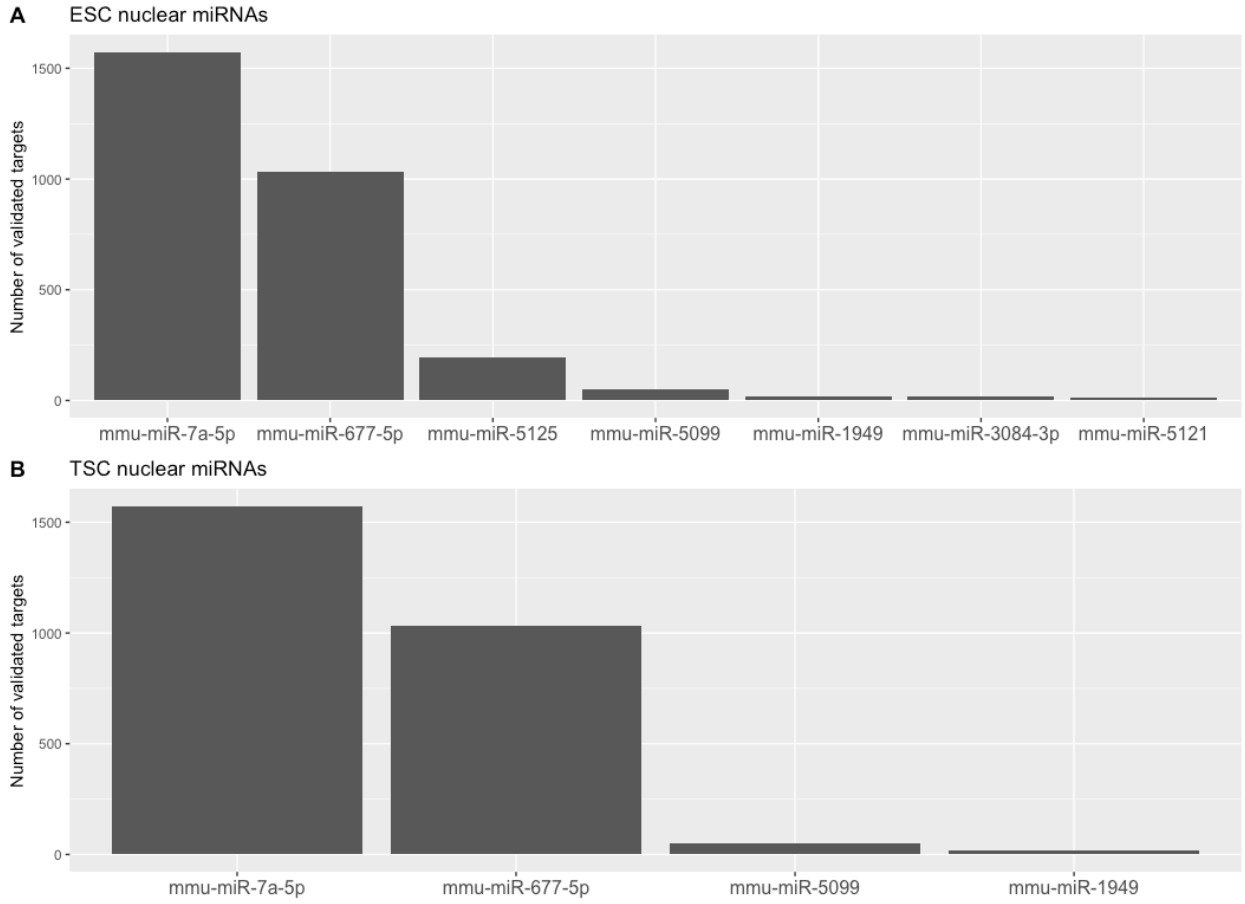


Figure 63: **MultiMiR result on up-regulated miRNAs in the nuclear fractions**

MultiMiR result on nuclear enriched miRNAs show miR-7a and miR-677 have the largest number of validated target mRNAs.

The target mRNA list of miR-7a and miR-677 generated by MultiMiR was used as input for gene set enrichment with GOrilla; the enrichment result shows that miR-7a is associated with target genes related to metabolic function (**Figure 64**), while miR-677 may regulate processes such as cell cycle phase transitions, signaling pathways, and metabolic processes (**Figure 65**). This suggests miR-7a and miR-677 may indeed play a role in maintaining pluripotency in ESCs and TSCs

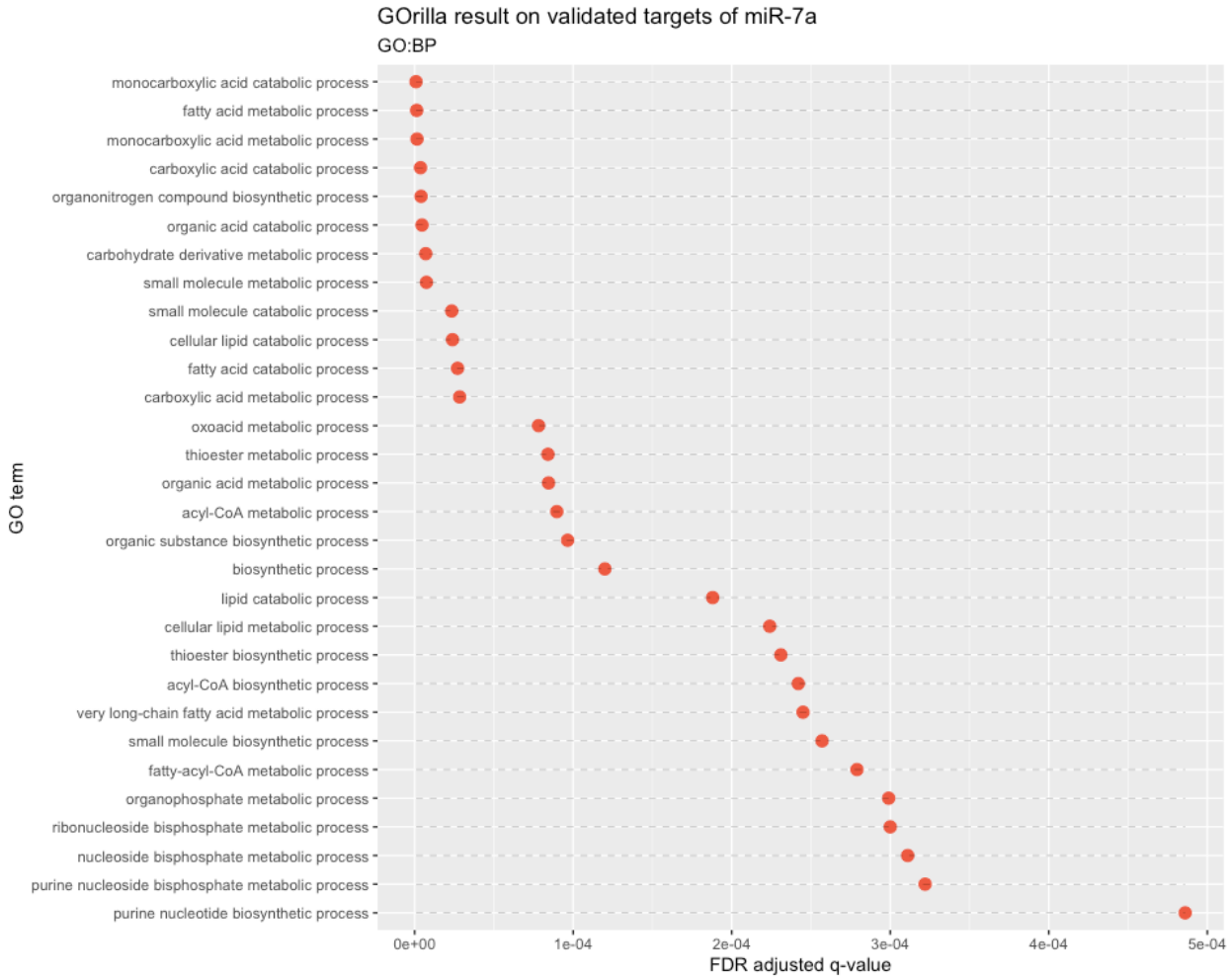


Figure 64: **GOrilla result on mRNA targets of miR-7a**

Gene set enrichment analysis result on mRNA targets of nuclear enriched miR-7a suggest miR-7a may regulate cellular processes related to metabolism.

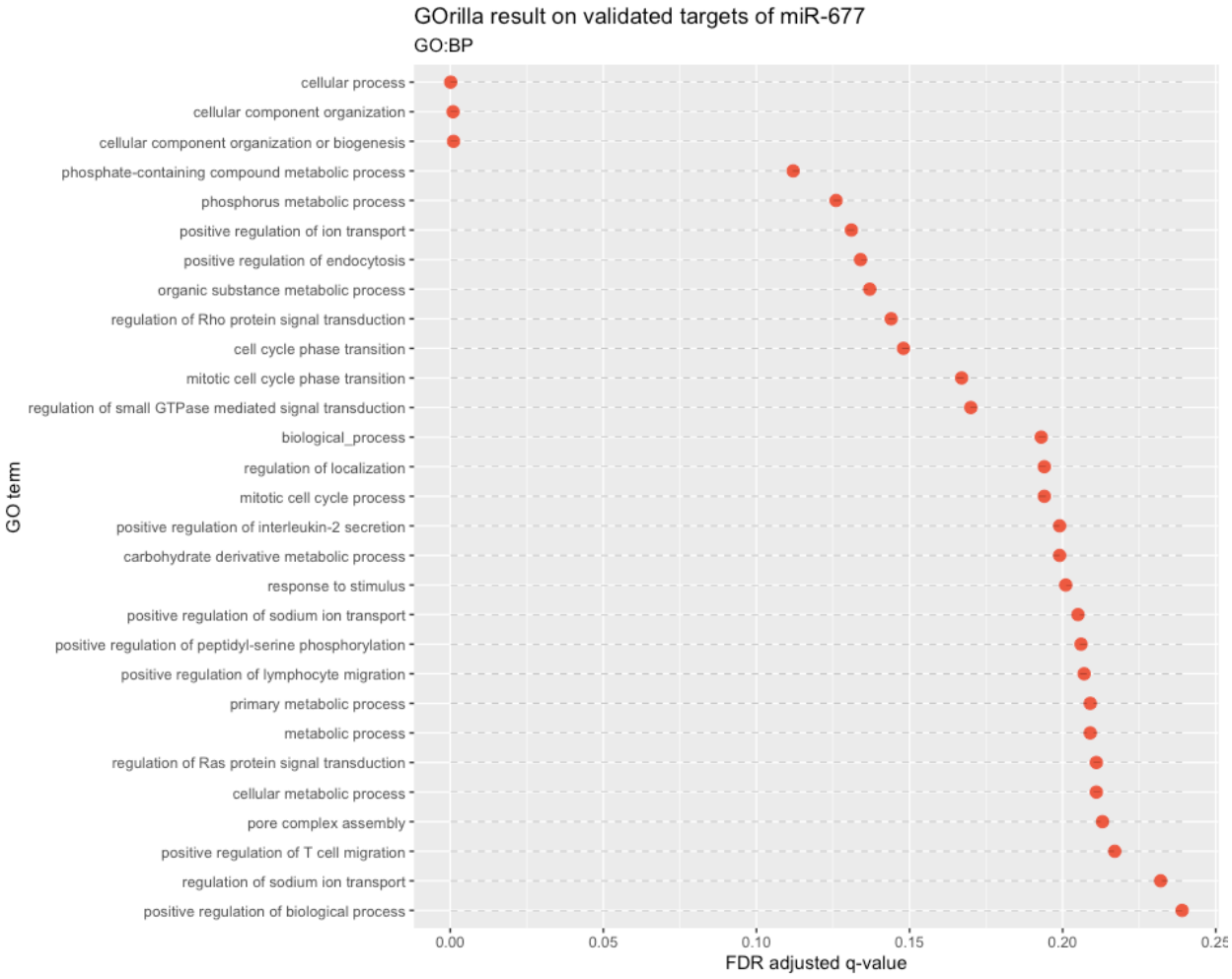


Figure 65: **GOrilla result on mRNA targets of miR-677**

Gene set enrichment analysis result on mRNA targets of nuclear enriched miR-677 suggest miR-677 may regulate cellular processes related to cell cycle phase transitions, signal transduction, and metabolic processes.

4.2.6 Network and functional enrichment analysis of cell specific miRNAs

Repeating the same type of analysis as **Section 4.2.4** but instead using differentially expressed miRNAs in cell-to-cell pairwise comparisons instead (i.e., ESC versus TSC in cytoplasmic fraction and ESC versus TSC in nuclear fraction) reveals identification of cell-specific miRNAs and associated processes under miRNA control. Segregating the cell-to-cell comparison by the cytoplasmic and nuclear fractions allows identification of potential differences in miRNA regulatory control by subcellular location. In the cytoplasmic samples, up-regulated miRNAs in ESCs with most number of validated targets are miR-497a, miR-124, miR-37b, miR-19a, and miR-130b. KEGG analysis for enriched functional pathways on the targets of these miRNAs show

similar results to analysis from **Section 4.2.4**; terms related to various signaling pathways such as mTOR, VEGF, and TGF- β pathways show enrichment. This result suggests miRNA control in such processes may be functionally more relevant in ESC lineage than in TSCs.

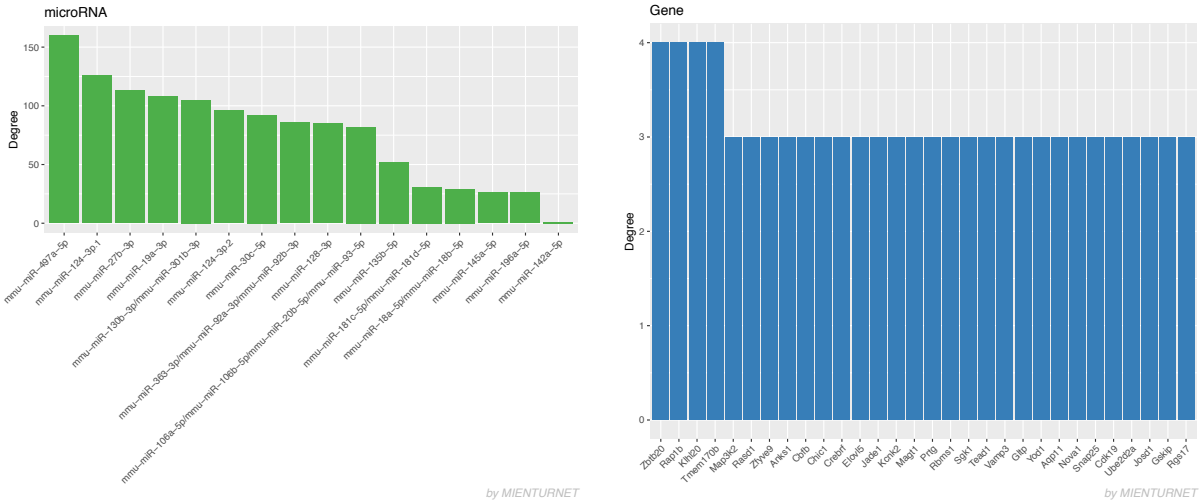


Figure 66: **Visualization of degree centrality in ESC cytoplasmic enriched miRNAs**

Left: histogram showing miRNAs up-regulated in ESCs (vs. TSCs) in the cytoplasmic fraction and the number of mRNA targets validated by TargetScan; high value of degree centrality indicates large number of mRNA targets per miRNA; such miRNAs include miR-497a, miR-124, and miR-27b; **right:** histogram showing target mRNAs and the number of interacting miRNAs; high value for degree centrality indicate multiple miRNAs interacting with a single gene.

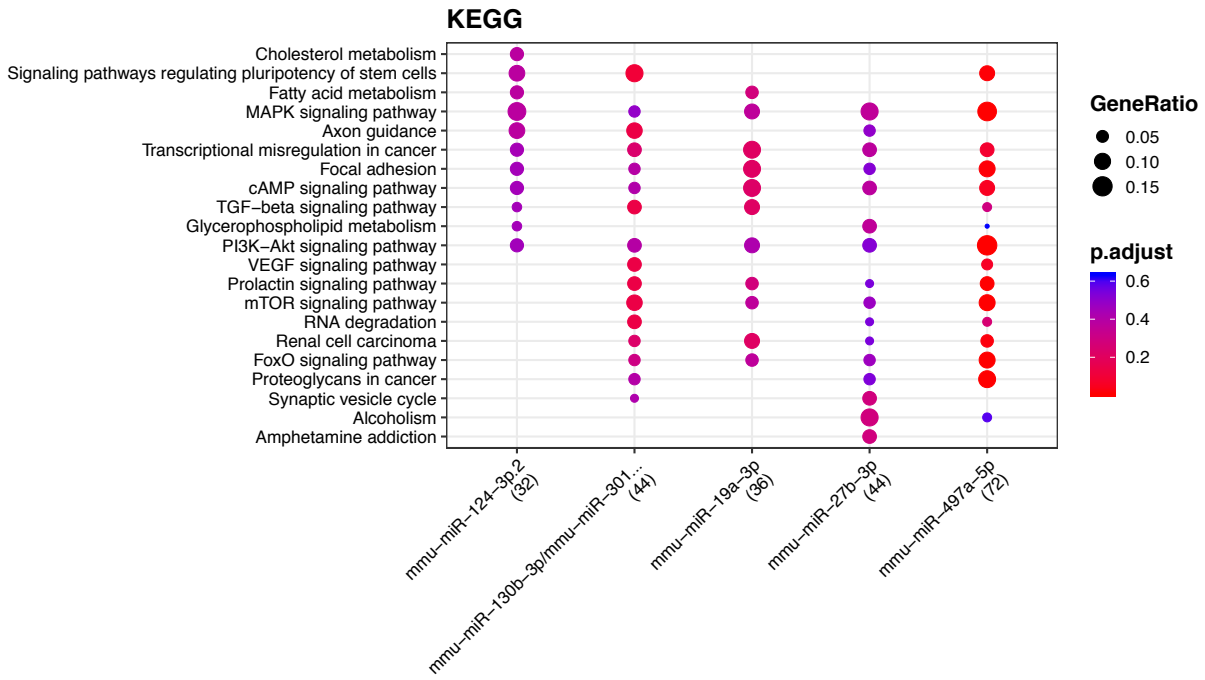


Figure 67: **KEGG pathway enrichment analysis result from ESC miRNA targets**

KEGG result reveal up-regulated miRNAs in ESCs (vs. TSCs) in the cytoplasmic fraction play a role in various signaling pathways involved in developmental processes.

On the contrary, network analysis of differentially expressed miRNAs in the cytoplasmic fraction of TSCs (versus cytoplasmic fraction of ESCs) identify miR-322, let-7, miR-29 family, miR-34a, and miR-148a as miRNAs with most number of validated targets. KEGG analysis on the target of these miRNAs reveal enrichment of Rap1 signaling pathway and Ras signaling pathway, which have shown to play a role in formation of cell adhesions and junctions, as well as in cell migration. Other enriched functional pathways include extracellular matrix receptor interaction and regulation of actin cytoskeleton – which also suggest miRNAs up-regulated in TSCs play a regulatory role in maintenance of cell structure and integrity.

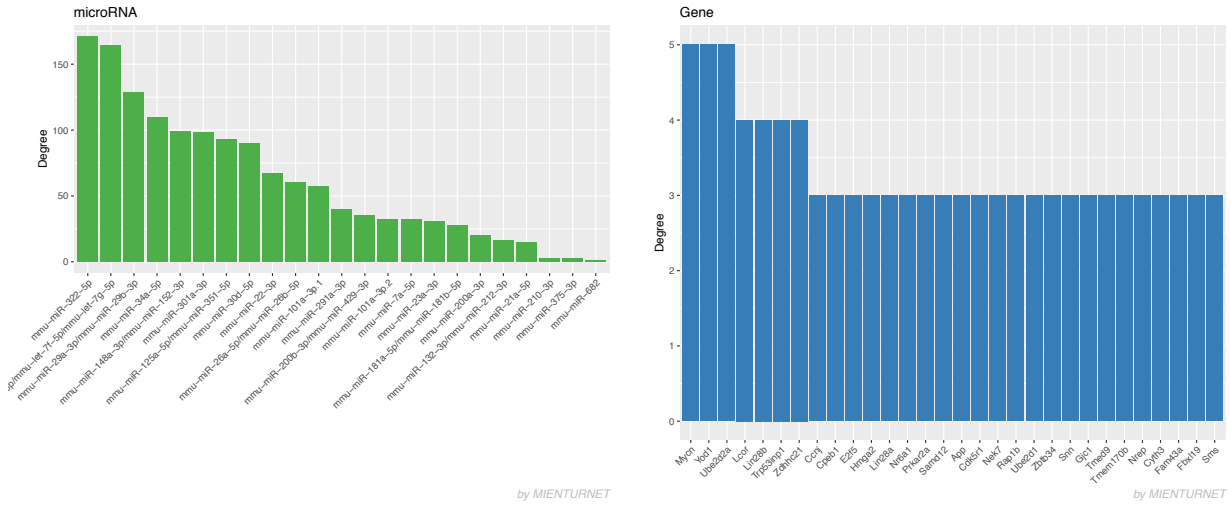


Figure 68: **Visualization of degree centrality in TSC cytoplasmic enriched miRNAs**

Left: histogram showing miRNAs up-regulated in TSCs (vs. ESCs) in the cytoplasmic fraction and the number of mRNA targets validated by TargetScan; miRNAs such as miR-322, the let-7 family, and miR-29 family show high number of mRNA targets; **right:** histogram showing target mRNAs and the number of interacting miRNAs; high value for degree centrality indicate multiple miRNAs interacting with a single gene.

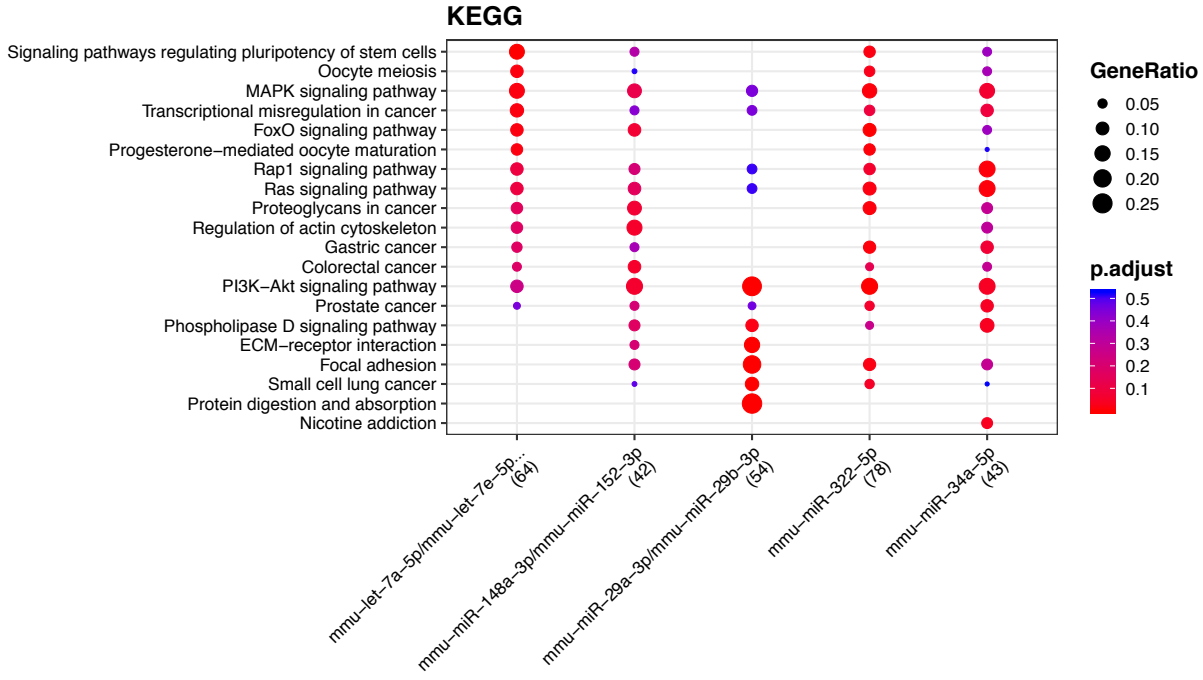


Figure 69: **KEGG pathway enrichment analysis result from TSC miRNA targets**

KEGG result reveal up-regulated miRNAs in TSCs (vs. ESCs) in the cytoplasmic fraction play a role in various signaling pathways involved in regulation of cell structure.

Network analysis result on differentially expressed miRNAs in cell-to-cell comparisons in nuclear fractions show a significant overlap in listed miRNAs as in the cytoplasmic fraction comparisons. As in the cytoplasmic fractions, up-regulated miRNAs in ESC nucleus (versus TSC nucleus) with most validated targets are miR-27b, miR-124, miR-19 family, miR-128, and miR-363. The top ten miRNAs listed in terms of the number of targets show 100% overlap with the results from cytoplasmic fraction. In miRNAs enriched in nuclear fraction of TSCs as well, there is significant overlap of miRNAs. Therefore, subsequent KEGG analysis on mRNA targets will lead to nearly identical results as above. This behaviour suggests that in cell-to-cell differential expression setup, cell-specific miRNAs may predominate in enrichment in both cytoplasmic and nuclear fractions such that subcellular differences cannot be elucidated. This result is due to the nonzero detection of pertinent miRNAs in both the cytoplasmic and nuclear samples.

4.2.7 *miRNA-FISH for miR-15b*

As a validation tool for differential subcellular localization of miRNAs, a well-studied miRNA in miR-15b – a miRNA involved in determination of trophoblast fate – was subject to miR-FISH assay. As seen, red fluorescence signal corresponding to miR-15b detection can be seen within the cell periphery in TSC colonies – suggesting an enrichment of miR-15b in TSCs, as expected (**Figure 70**). Furthermore, as indicated by merged image with DAPI staining, miR-15b may be localized to the nuclei of TSCs.

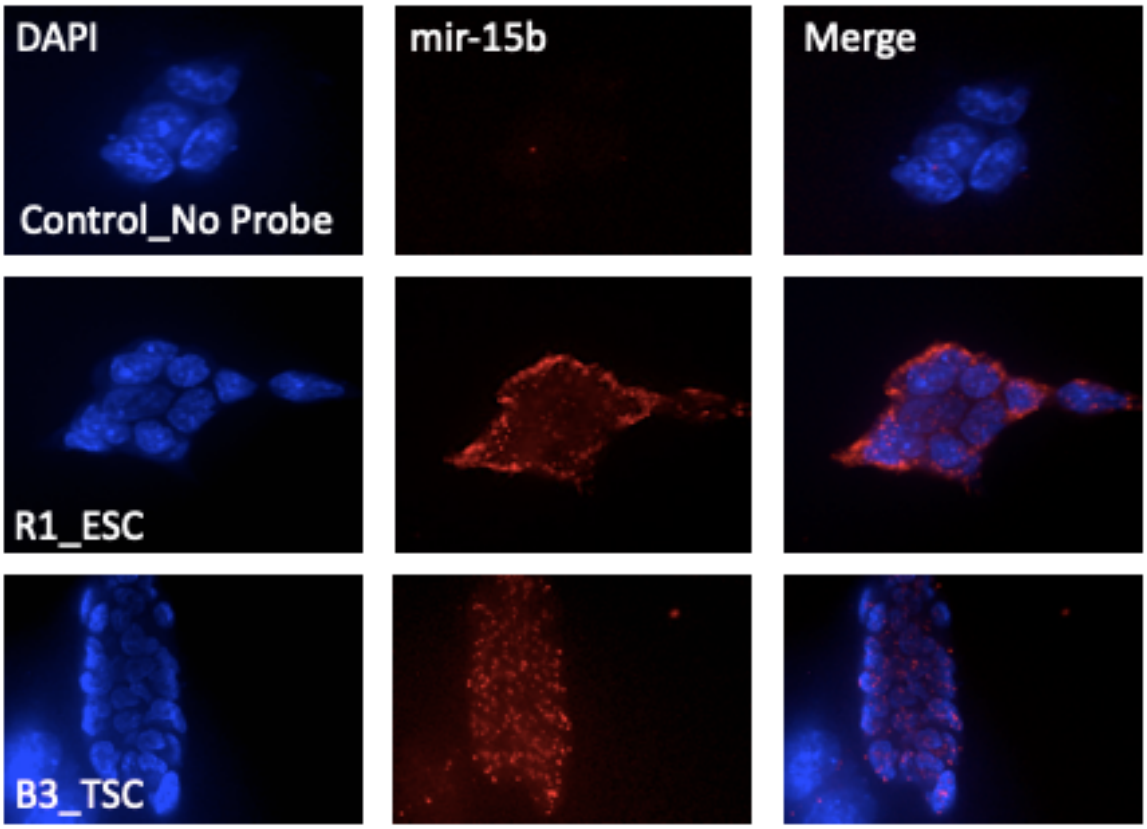


Figure 70: **Visualization of miR-15b using miR-FISH**

Top row: in a control sample with no hybridization probe, no significant fluorescence signal is detected in the red channel as expected – suggesting low amount of autofluorescence; **middle row**: in ESC sample hybridized with miR-15b FISH probe, fluorescence signal is detected but show localization around cell colony periphery, suggesting detectable levels of miR-15b expression or a possible artifact in cell staining; **bottom row**: in TSC sample hybridized with miR-15b FISH probe, fluorescence signal is detected within the DAPI stained nuclei, suggesting nuclear presence of miR-15b.

4.2.8 miRNA-FISH for miR-6240 nuclear detection

Detection of miR-15b using miR-FISH in TSC colonies suggest the suitability of such assay as a visualization tool for miRNA nuclear localization. As such, miR-FISH was performed on a nuclear miRNA as identified by differential expression analysis in TSC data (**4.2.3**) to ensure the validity of the experimental design. Obtained confocal images on the FISH assay for miR-6240 show detection of fluorescence signal within the TSC colonies as well as within the nuclei (**Figure 71**).

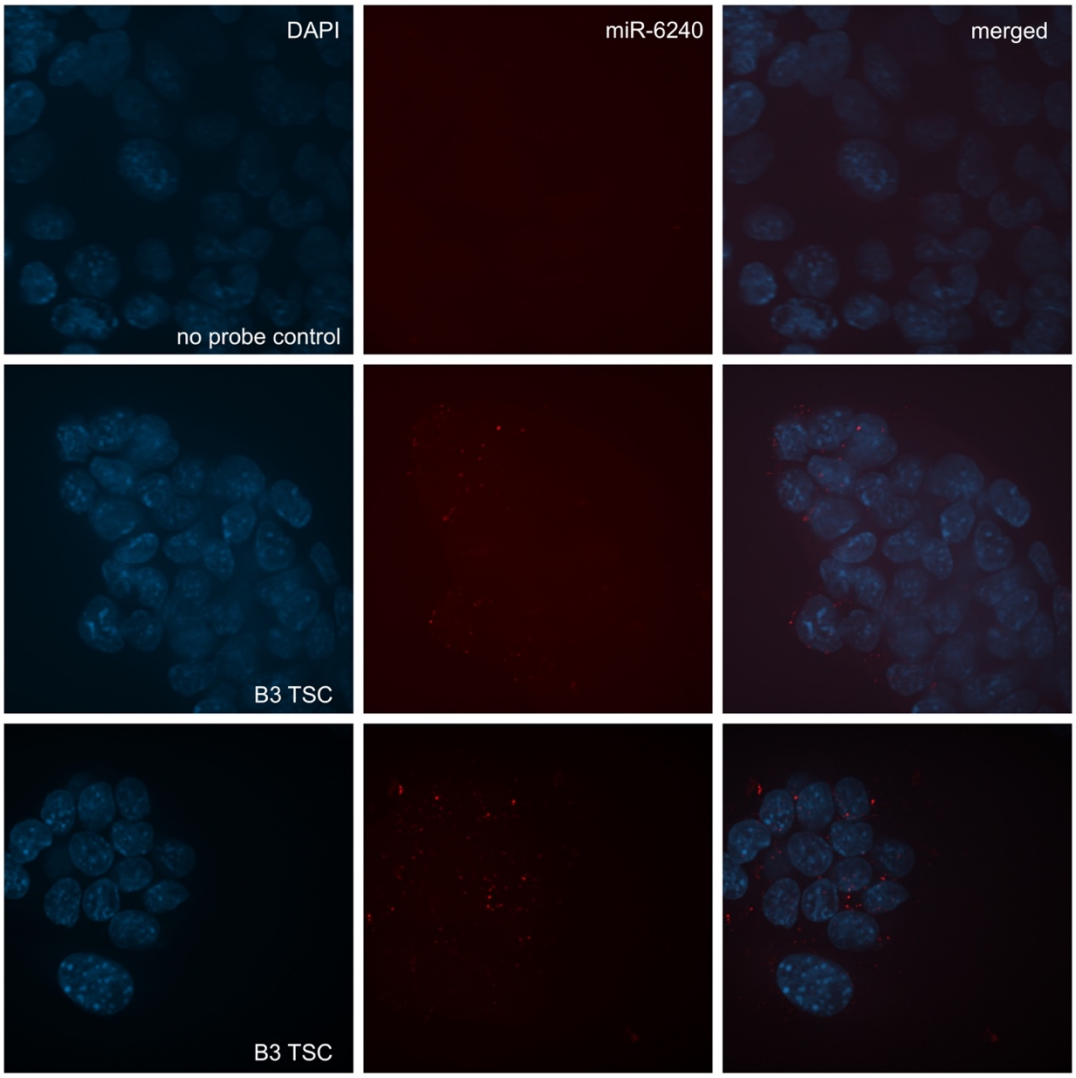


Figure 71: **Visualization of miR-6240 using miR-FISH**

Top row: in a control group in absence of hybridization probe, no fluorescent signal is detected in the red channel; **middle row**: in TSC sample with probe for miR-6240, some fluorescence signal can be seen within the cell colony; **bottom row**: fluorescence signal can be seen again in and around individual nuclei in another cell colony of a TSC sample.

4.3 Discussion

4.3.1 miRNAs related to signaling networks are up-regulated in ESCs' cytoplasmic fraction

The differential expression profile between the two subcellular fractions in ESCs suggest the presence of mature, single-stranded miRNAs in both cytoplasmic and nuclear compartments (**Figure 56**). Analysis of mRNA localization in **Chapter 3** showed that transcripts related

metabolism, translation, and chromatin modifications tend to be up-regulated in the cytoplasm – a result indicative of ESCs' hyperactive transcriptome, high biosynthetic demand, and epigenetic regulatory mechanisms for self-renewal and maintenance of pluripotency. Furthermore, ncRNA differential expression showed an enrichment of ncRNAs related to cell response to stress and chemical stimuli. As the ESCs' self-renewal and proliferation is regulated by signaling networks from intrinsic factors as well as external stimuli (i.e., cytokines, growth factors), regulatory ncRNAs such as miRNAs must be key factors in governing such signaling pathways.

Indeed, validated target mRNAs of cytoplasmic up-regulated miRNAs in ESCs from differential expression (**Figure 56**) from TargetScan suggest their role in regulating various signaling pathways; KEGG enrichment result shows miRNA-mRNA interactions in the ESC cytoplasm may largely be associated with regulation of signaling pathways related to differentiation and self-renewal (**Figure 60**). In particular, miR-15a/miR-497a, miR-130a and miR-130b, miR-17, and miR-30c had the highest number of validated mRNA targets, suggesting these miRNAs may play a significant role in regulatory network in ESCs.

KEGG enrichment result (**Figure 60**) suggest the signaling pathways governed by such miRNAs include the vascular endothelial growth factor (VEGF) signaling pathway, the mammalian target of rapamycin (mTOR) signaling pathway, the transforming growth factor (TGF)-β signaling pathway, and forkhead box transcription factors class O (FOXO) signaling pathway. All of the above signaling networks has been associated with ESC self-renewal and differentiation in literature[154-156]. Blocking VEGF signaling in mouse ESCs has been shown to maintain ground state pluripotency, as VEGF secretion has been associated with the differentiation of mouse ESCs towards meso-endoderm lineages *in vitro*[154]. Leukemia inhibitory factor (LIF) signaling pathway contributes to ESC maintenance by suppressing mTOR, and an *in vitro* depletion of LIF in mouse ESCs showed down-regulation of pluripotency markers Oct4, Sox2, and Nanog and an up-regulation of post-implantation epiblast markers[157]. TGF-β family signaling has been shown to maintain pluripotency in ESCs, in particular due to induction of Smad2 localization to the nucleus, as well as via Nodal signaling[157,158]. Finally, a CHIP-Seq study found that FOXO proteins and their orthologues in both human and mouse ESCs were shown to be essential for maintenance of pluripotency via interaction with regulatory sequences of SOX2 and OCT4 genes[156]. The

association of up-regulated miRNAs in the cytoplasmic fraction of ESCs with such signaling networks then suggests miR-15a/miR-497a, miR-130a and miR-130b, miR-17, and miR-30c as candidates to regulate ESC maintenance and pluripotency.

The role of miRNAs in relation to signaling pathways and pluripotency have been documented in literature. Yin *et al* found that an over-expression of miR-15a in mouse hindlimb led to a suppression of angiogenesis via inhibition of endogenous VEGF function[159]. MiR-17, as part of the miR-17-92 cluster, has been associated with essential roles in cell cycling progression and cell proliferation[160]. Result in **Figure 60** suggest both miR-15a and miR-17 are also abundant in the cytoplasm of mouse ESCs, suggesting possible role of these miRNAs in maintenance of ESCs as well.

4.3.2 ESC-TSC differential expression suggest miRNA role in cell fate specification

The role of miRNAs in cell lineage determination is facilitated by Watson-Crick base-pairing between the miRNA and the target transcript. This miRNA-mRNA interaction influences transcription factor networks, which in turn modulates the transcription of genes at the DNA level. For example, both miR-15a cluster and let-7 family has been shown to directly target and down-regulate the expression of CDK6, which consequently prevents cells from entering the S phase of the cell cycle[161,162]. Similarly, an induction of miR-15a, miR-322, and miR-467g in mouse ESCs was sufficient in causing the up-regulation of trophoblast marker Cdx2 and Gata3, as well as down-regulation of ESC marker Oct4[79]. Profiling the differential expression of miRNAs between ESCs and TSCs then, can reveal miRNAs that may play a pertinent role in the maintenance of each cell lineage. In comparison of the cytoplasmic fractions from ESCs and TSCs, candidates for such miRNAs were shown in **Figures 66** and **68**. In ESCs, miRNAs with the most number of validated mRNA targets (as per TargetScan database) were: miR-497a, miR-124, miR-27b, miR-19a, and miR-130b/miR-301b with > 100 known targets. In TSCs, miR-322, let-7 family, miR-29a/29b, and miR-34a had the most known target mRNAs.

KEGG enrichment analysis on the target mRNAs of the candidate miRNAs in ESCs found that these miRNAs are likely associated with signaling pathways that govern differentiation and lineage specification (**Figure 67**). Up-regulated miRNAs in ESCs show association with signaling

pathways associated with differentiation, such as TGF-β and VEGF pathways, as well as those associated with ESC maintenance such as FOXO and mTOR (described in Section **4.3.1**). Furthermore, signal cascade pathways such as MAPK and cAMP pathways are also found enriched, suggesting miRNA influence in signal transduction and gene expression regulation. Other enriched pathways include terms associated with metabolism and cancer, which suggest miRNA role in both regulation of metabolic processes and cell cycling. As both processes are pertinent in maintenance of pluripotency and self-renewal, this suggest miRNAs function as key regulators in ESCs.

Similarly, in TSCs, candidate miRNAs were associated with signaling pathways related to regulation of pluripotency (**Figure 69**). MAPK and FOXO signaling pathways were enriched, as well as Ras1 and Rap signaling pathways. In literature, both Ras1 and Rap signaling pathways have been associated with formations of cell adhesions and junctions, as well as in cell migration[163,164]. In particular, a study of mouse ESCs found that ectopic induction of Ras-MAPK pathway was sufficient in inducing extraembryonic trophoectoderm fate, with associated increase in Cdx2 and decrease in Nanog expression[162]. Other enriched terms include regulation of actin cytoskeleton and extracellular matrix receptor interaction, which suggest miRNAs enriched in TSCs also play a role in cell structure organization and cell adhesion.

Therefore, profiling of miRNAs in two distinct cell lineage representatives allows for a comparison of miRNA influence in maintenance of each identity. As observed above, by identifying miRNAs with a large number of mRNA targets and associated functional networks, it can be elucidated that ESCs and TSCs both harness miRNA-mRNA interactions in order to maintain pluripotency and cell identity.

4.3.3 miRNAs can be localized to the nucleus

Findings in literature that support both cytoplasmic and nuclear roles of miRNAs, as well as detection of RNA-induced silencing complex (RISC) components in the nucleus suggest miRNAs may shuttle between the two cellular compartments[39-41]. Indeed, the differential expression profile between the two subcellular fractions was able to identify miRNAs up-regulated in the nuclear

fraction in both ESCs and TSCs (**Figures 57** and **58**). Using *in situ* hybridization assay, the nuclear presence of nuclear miRNAs was also visualized in fixed cell culture (**Figures 70** and **71**).

The function of miRNAs found to be up-regulated in the nuclear fraction is unclear. MIENTURNET using TargetScan showed that only two miRNAs: miR-7a and miR-6539 showed a significant amount of overlapping validated targets (**Figures 61** and **62**). The function of these two miRNAs in relation to pluripotency, however, remains largely unknown in literature; one particular study of mouse adult pancreatic islets found that levels of miR-7a was up-regulated and targets five components of the mTOR signaling pathway, such that the inhibition of miR-7a led to the activation of mTOR signaling and proliferation of adult β-cells in primary islets[165].

Alternatively, R package MultiMiR was used to find validated target mRNAs for miRNAs up-regulated in the nuclear fractions, without considerations in network similarities as per MIENTURNET. Using MultiMiR, miR-7a and miR-677 showed significantly higher number of validated targets than other nuclear enriched miRNAs in both ESCs and TSCs (**Figure 63)**. This result reiterates that miR-7a and miR-677 not only show localization to the nucleus in both cells, but also potentially influence a significant portion of the transcriptome. Gene set enrichment analysis using GOrilla on the list of miR-7a and miR-677 targets from MultiMiR shows that miR-7a may be associated with the regulation of metabolic processes (**Figure 64**), while miR-677 may regulate metabolic processes, cell cycle phase transitions, and signal transduction (**Figure 65**). This result suggest nuclear detected miRNAs such as miR-7a and miR-677 may indeed play a role in cell proliferation and self-renewal in both ESCs and TSCs.

Chapter 5

5 Conclusions

It is evident from the differential expression analysis of the two subcellular fractions that both protein-coding mRNAs and non-coding RNAs (such as lncRNAs, snoRNAs, and miRNAs) are pervasive in both the cytoplasm and the nucleus. Furthermore, profiling exon-exon and exon-intron junction data showed that intron retaining behaviour is a phenomenon occurring in both subcellular compartments as well. As a proxy of cellular function, the functional signature of up-regulated mRNAs and their subcellular localization gave insight into the processes governing maintenance of ESCs and TSCs and by extension, the inner cell mass and the extraembryonic trophoectoderm in the pre-implantation blastocyst.

Both ESCs and TSCs, as evident in literature, rely on an organized network of molecular signaling to self-renew and maintain their cell identity. This led to both similarities and dissimilarities in their gene expression behaviour. The similarities in up-regulated genes and their associated biological function was prevalent throughout the differential expression analyses - the reliance on metabolic processes to support the biosynthetic demands of self-renewal and proliferation led to the accumulation of genes with metabolic function at the site of protein synthesis; the hyper-activation of the transcriptome associated with proliferating cells led to an overall shift of RNA population towards the nuclear fraction, which indicate an accumulation of nascent mRNA; differential expression profile of miRNAs suggest favored cytoplasmic enrichment, suggesting their role in mRNA regulation at the level of translation; finally, mRNAs related to cell division and cell cycle accumulated in the nuclear fraction, which suggest genes related to developmental timing may be actively retained in the nucleus. Furthermore, analysis with junction quotients for estimation of percent intron retention showed that in both ESCs and TSCs, genes related to cell division, cell cycle, as well as RNA processing and splicing – factors that relate to intron retention themselves – show considerable levels of intron retaining behaviour, which suggest intron retention as a conserved regulatory mechanism.

Dissimilarities in the data between ESCs and TSCs reveal divergent cellular processes required for each cell lineage and how they are regulated. This is evident by the expression of intron

retaining mRNAs related to immune and circulatory function in TSCs, which suggest TSCs may fine-tune the expression of TE related genes using alternative splicing events. The accumulation of cell-adhesion-molecules and genes related to cell signaling in the nuclear fraction of TSCs suggest nuclear retained transcripts may be associated with signaling cues and regulation via nuclear localization. ESCs showed an enrichment of intron retaining mRNAs related to DNA repair and response to stress, which reiterate the role of intron retention in signal response. Lineage specific functional gene sets were also observed, such that intron retaining mRNAs related to neural development were up-regulated in ESCs' nuclear fraction and mRNAs related to cell-cell adhesion were up-regulated in TSCs' nuclear fraction. This showed that intron retention, while a conserved mechanism that regulate processes such as cell division and cell cycle, also show regulatory control on lineage-specific processes.

These findings support the hypothesis that RNA-seq analysis of two subcellular fractions in ESCs and TSCs reveal cellular processes regulated by subcellular localization. The subcellular localization profile is then reflective of how the ICM and TE fate is established by gene expression regulation programs. Up-regulated transcripts in the cytoplasmic fraction may be indicative of cellular processes associated with high rate of protein synthesis. The identity of these transcripts was found to be associated with processes related to the ability to self-renew (i.e., ribosomal assembly, metabolic processes) as well as lineage-specific processes (i.e., neural development, cell-cell adhesion). Up-regulated transcripts in the nuclear fraction may be indicative of processes related with temporal regulation of gene expression, or an accumulation of nascent mRNA. These transcripts were found to be related to cell division, cell cycle, and response to stimuli.

The hypothesis that transcripts related to developmental timing and response to stimuli may also show intron-retaining behaviour is supported in the differential expression profile using exon-intron junction counts. The advantage of comparing exon-intron junction data in the two fractions is that it contextualizes intron retention, as the fate of intron-retaining transcript varies with its subcellular localization. As such, exon-intron junction differential expression profile showed that transcripts related to cell cycle, cell division, chromatin modifications, ion transport, and neural development may be subject to intentional nuclear detainment in ESCs, whereas transcripts related

to wound healing, inflammatory response, cell-matrix and cell-cell adhesion, blood circulation, and extracellular matrix organization are subject to nuclear detainment in TSCs.

The disadvantage in using differential expression profile to infer intron retention is that such analysis disregards transcripts that are not differentially expressed between the two fractions. Using junction quotients per gene in each fraction to estimate intron retention allows identification of intron retaining transcripts prevalent in both fractions. This analysis showed that genes that may be related to intron retention and alternative splicing events themselves also show intron retaining behaviour and is expressed in both the cytoplasm and the nucleus.

It is important to reiterate that in an experiment of RNA-seq data at a single time-point, regulatory processes such as intron retention triggered non-sense mediated decay (IR-NMD) cannot be inferred. As IR-NMD has been shown to be a pervasive mechanism in maintenance of cell self-renewal, a future experiment with an incorporated time component in data collection will be useful to investigate changes in gene expression in relation to intron retaining behaviour. In the current data, accumulated intron retaining transcripts in the cytoplasm show a positive correlation with the extent of intron retention, which suggest transcripts which evade IR-NMD may indeed be stabilized at the site of protein synthesis. This finding shows that in both the cytoplasm and the nucleus, transcripts that retain their introns persist in detectable amounts – an observation conserved in both ESCs and TSCs.

Future experiments should aim to investigate the consequence of intron retention in relation to the location of the retained intron. The limitation of an experiment forgoing the location of intron retention within the transcript is that it assumes detectable intron retaining transcript is likely to be stabilized. Literature evidence shows that the fate of an intron retaining transcript (i.e., whether it is likely to be subject to degradation or stabilization) is related to the location of the retained intron relative to the open reading frame. Incorporating this factor will lead to a better understanding of how intron retention modulates gene expression and predict whether detected transcript is subject to IR-NMD.

Profiling miRNAs provided added insight into how ESCs and TSCs maintain their cell identity. Differential expression of miRNAs showed miRNAs are closely related to cell signaling pathways that govern their cellular function. As such, it is evident that ESC- and TSC-enriched miRNAs associate with key signaling pathways such as TGF- β, VEGF, and mTOR as well as Ras1 and Rap pathways in ESCs and TSCs, respectively. In terms of subcellular localization, miR-7a and miR-677 showed nuclear localization in both ESCs and TSCs and have shown to associate with regulation of metabolic processes, cell cycle, and cell signaling, which suggest their involvement in cellular processes pertinent in maintenance of self-renewal.

In order to directly verify the miRNA role in mRNA regulation, a pull-down assay of differentially expressed miRNAs and their associated targets can be used to identify miRNA-mRNA interactions directly from the source. Such immunoprecipitation assays include HITS-CLIP and PAR-CLIP, which rely on crosslinking between RNA-binding proteins (such as the Argonaute protein involved in miRNA-mRNA interaction) and protein binding sites to map interactions. Future experiments should aim to couple such assays in ESCs and TSCs to RNA-seq in order to identify miRNA-mRNA interactions involved in each cell lineage specification.

Incorporating pull-down assays and genomic location specific analysis of intron retention will also allow inference on whether intron retention is related to the likelihood of gene regulation via miRNA targeting. Literature evidence shows intron retaining transcripts are likely to harbor longer 3’ untranslated regions, which in turn harbors potential binding sites for miRNAs. Therefore, future analysis should aim to investigate whether intron retention and miRNAs function in conjunction to modulate overall gene expression.

Overall, the ability to quantify RNAs at a single nucleotide resolution allowed the analysis to not just account for exons, but introns and junction boundaries as well. Harnessing this ability, it was possible to infer variability in not only overall gene expression, but the splicing behaviour of individual genes as well. Profiling mRNAs without poly-adenylation selection as well as non-coding RNAs allowed for a more comprehensive analysis of the subcellular dynamics of RNA population in the cytoplasm and the nucleus, accounting for accumulation nascent mRNA, nuclear detainment of intron retaining transcripts, and cytoplasmic enrichment for protein synthesis. For

biological systems at developmental crossroads such as ESCs and TSCs, the regulatory control on gene expression remains crucial and understanding the transcriptome continues to shed further light on the mechanisms at which developmental potential is maintained.

Chapter 6

6 Implementation of Bioinformatics Analysis

cutadapt, HISAT2, samtools, BEDOPS, featureCounts, and Bowtie were executed on the UNIX command line interface running on Ubuntu 18.04. CirGO was used with dependencies from Python 2.7.4 on the command line. R packages edgeR, limma, voom, NOISeq, topGO, and clusterProfiler were used on R version 3.6.3. R sessionInfo() output is shown below.

```
> sessionInfo()

R version 3.6.3 (2020-02-29)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Mojave 10.14.5

Matrix products: default
BLAS:
/System/Library/Frameworks/Accelerate.framework/Versions/A/Frame
works/vecLib.framework/Versions/A/libBLAS.dylib
LAPACK:
/Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRl
apack.dylib

locale:
[1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8

attached base packages:
[1] parallel   stats4     stats       graphics   grDevices utils
datasets
[8] methods   base

other attached packages:
 [1] topGO_2.38.1           SparseM_1.78           GO.db_3.10.0
 [4] graph_1.64.0           forcats_0.5.0          stringr_1.4.0
 [7] dplyr_1.0.0            purrr_0.3.4            readr_1.3.1
[10] tidyr_1.1.0            tibble_3.0.1           ggplot2_3.3.1
[13]   tidyverse_1.3.0                      org.Mm.eg.db_3.10.0
AnnotationDbi_1.48.0
[16] IRanges_2.20.2         S4Vectors_0.24.4       Biobase_2.46.0
[19] BiocGenerics_0.32.0    clusterProfiler_3.14.3 edgeR_3.28.1
[22] limma_3.42.2

loaded via a namespace (and not attached):
 [1] fgsea_1.12.0        colorspace_1.4-1    ellipsis_0.3.1
```

```
 [4] ggridges_0.5.2      qvalue_2.18.0       fs_1.4.1
 [7] rstudioapi_0.11     farver_2.0.3        urltools_1.7.3
[10] graphlayouts_0.7.0  ggrepel_0.8.2       bit64_0.9-7
[13] fansi_0.4.1         lubridate_1.7.8     xml2_1.3.2
[16] splines_3.6.3       GOSemSim_2.12.1     polyclip_1.10-0
[19] jsonlite_1.6.1      broom_0.5.6         dbplyr_1.4.4
[22] ggforce_0.3.1       BiocManager_1.30.10 compiler_3.6.3
[25] httr_1.4.1          rvcheck_0.1.8       backports_1.1.7
[28] assertthat_0.2.1    Matrix_1.2-18       cli_2.0.2
[31] tweenr_1.0.1        prettyunits_1.1.1   tools_3.6.3
[34] igraph_1.2.5        gtable_0.3.0        glue_1.4.1
[37] reshape2_1.4.4      DO.db_2.9           fastmatch_1.1-0
[40] Rcpp_1.0.4.6        enrichplot_1.6.1    cellranger_1.1.0
[43] vctrs_0.3.1         nlme_3.1-148        ggraph_2.0.3
[46] rvest_0.3.5         lifecycle_0.2.0     DOSE_3.12.0
[49] europepmc_0.4       MASS_7.3-51.6       scales_1.1.1
[52] tidygraph_1.2.0     hms_0.5.3           RColorBrewer_1.1-2
[55] yaml_2.2.1          memoise_1.1.0       gridExtra_2.3
[58] triebeard_0.3.0     stringi_1.4.6       RSQLite_2.2.0
[61] BiocParallel_1.20.1 rlang_0.4.6         pkgconfig_2.0.3
[64] matrixStats_0.56.0  lattice_0.20-41     cowplot_1.0.0
[67] bit_1.1-15.2        tidyselect_1.1.0    plyr_1.8.6
[70] magrittr_1.5        R6_2.4.1            generics_0.0.2
[73] DBI_1.1.0           pillar_1.4.4        haven_2.3.0
[76] withr_2.2.0         modelr_0.1.8        crayon_1.3.4
[79] viridis_0.5.1       progress_1.2.2      locfit_1.5-9.4
[82] grid_3.6.3          readxl_1.3.1        data.table_1.12.8
[85] blob_1.2.1          reprex_0.3.0        digest_0.6.25
[88] gridGraphics_0.5-0  munsell_0.5.0       viridisLite_0.3.0
[91] ggplotify_0.0.5
```

Adapter trimming on raw RNA-seq FASTQ files:

```
$ cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -m 15 -o
trimmed.r1.fastq original.r1.fastq

$ cutadapt -a AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -m 15 -o
trimmed.r2.fastq original.r2.fastq
```

HISAT2 alignment & alignment processing

```
$ hisat2 -p 8 --rg PL:ILLUMINA -x hisat2.index --dta --rna-
strandness RF -1 trimmed.r1.fastq -2 trimmed.r2.fastq -S
alignment.sam
```

```
$ samtools view -bS alignment.sam alignment.bam

$ samtools sort alignment.bam alignment.sorted.bam

$ samtools index alignment.sorted.bam
```

Generation of intron annotations

Following R script was written by Devon Ryan and posted on the public web forum Biostars[166]: https://www.biostars.org/p/165226/.

```
> gtf <- makeTxDbFromGFF("UCSCgenes.annotation.gtf")
> exons <- exonsBy(gtf, by="gene")

> exons <- reduce(exons)
> exons <- exons[sapply(exons, length) > 1]

> introns <- lapply(exons, function(x) {

    gr = GRanges(seqnames=seqnames(x)[1],
ranges=IRanges(start=min(start(x)),
        end=max(end(x))),
        strand=strand(x)[1])
    db = disjoin(c(x, gr))
    ints = db[countOverlaps(db, x) == 0]

    if(as.character(strand(ints)[1]) == "-") {
        ints$exon_id = c(length(ints):1)
    } else {
        ints$exon_id = c(1:length(ints))
    }
    ints
})
> introns <- GRangesList(introns)
```

Generation of count tables

```
$ featureCounts -t exon -g gene_id -a UCSCgenes.annotation.gtf -
o counts.exons.txt alignment.sorted.bam
```

NOISeq QC report generation

```
> QCreport(counts, samples = NULL, norm = FALSE)
```

edgeR-voom-limma for differential expression

```
> y <- DGEList(counts = counts[,2:14], genes = counts$symbol,
group = class)

> design <- model.matrix(~0+class)
> contr.matrix<-makeContrasts(
  ESCcyt_vs_ESCnuc = ESC_cyt-ESC_nuc,
  TSCcyt_vs_TSCnuc = TSC_cyt-TSC_nuc,
  ESCcyt_vs_TSCcyt = ESC_cyt-TSC_cyt,
  ESCnuc_vs_TSCnuc = ESC_nuc-TSC_nuc,
  levels=colnames(design)
)

> v <- voom(y, design, plot = TRUE)
> vfit <- lmFit(v, design)
> vfit <- contrasts.fit(vfit, contrasts = contr.matrix)
> efit <- eBayes(vfit)


> result <- decideTests(efit, adjust.method = "fdr", p.value =
0.05)
```

topGO for gene ontology enrichment analysis

```
> allGO2genes <- annFUN.org(whichOnto = "BP", feasibleGenes =
NULL, mapping = "org.Mm.eg.db", ID = "symbol")

> GOdata <-  new("topGOdata", ontology = "BP", allGenes =
geneList, annot =a nnFUN.GO2genes, GO2genes = allGO2genes,
geneSel = topDiffGenes, nodeSize = 10)

> results.ks <- runTest(GOdata, algorithm = "classic", statistic
= "ks")
> goEnrichment <- GenTable(GOdata, KS = results.ks, orderBy =
"KS", topNodes = 200)
> goEnrichment <- goEnrichment[which(goEnrichment$Annotated <
500),]
```

Gene ontology enrichment visualization with CirGO

```
$ python CirGO.py -inputFile REVIGO.treemap.tsv -outputFile
graph.svg -fontSize 14 -numCat 40 -legend "GO:BP"
```

Selection of split reads from gene alignment

```
$ samtools view -h alignment.sorted.bam | awk '$0 ~ /^@/ || $5 ~
/N/' | $ samtools view -b > splitreads.bam
```

Generation of exon-intron junction data

The awk scripts used here was written by Alex Reynolds, originally posted on the public web forum Biostars[166] (https://www.biostars.org/p/315680/) and accessed from his GitHub pages: https://gist.github.com/alexpreynolds.

```
$ awk '($3 == "exon")' UCSCgenes.annotation.gtf | gtf2bed | cut
-f1-6 > exons.bed

$ awk -f transcripts2mergedExons.awk exons.bed >
merged.exons.bed

$ awk -f mergedExons2exonIntronList.awk merged.exons.bed >
exons.introns.bed

$ awk -f exonIntronList2JunctionList.awk exons.introns.bed >
exons.introns.junctions.bed

$ bedops --everything --range 5 exons.introns.junctions.bed >
exons.introns.junctions.pad5.bed

$ bedmap --echo --count --delim '\t'
exons.introns.junctions.pad5bed <(bam2bed <
original.sorted.bam) > exon.intron.junction.counts.bed
```

Generation of exon-exon junction data

```
$ bedmap --echo --count --delim '\t' exons.bed <(bam2bed <
splitreads.bam) > exon.exon.junction.counts.bed
```

Functional gene set enrichment analysis with clusterProfiler

```
> goObject <- enrichGO(gene = geneList, OrgDb = org.Mm.eg.db,
keyType = "SYMBOL", ont = "BP")
```

Adapter trimming in small RNA-seq FASTQ files

```
$ cutadapt -a TGGAATTCTCGGGTGCCAAGG -m 15 -o trimmed.small.fastq
small.fastq
```

Alignment of small RNA-seq files with Bowtie

```
$ bowtie -S bowtie.indices.ebwt trimmed.small.fastq >
alignment.small.sam
```

7 Citations

1. Martin, Jeffrey A., and Zhong Wang. "Next-generation transcriptome assembly." *Nature Reviews Genetics* 12.10 (2011): 671-682.
2. Esteller, Manel. "Non-coding RNAs in human disease." *Nature reviews genetics* 12.12 (2011): 861-874.
3. Zhang, Peijing, et al. "Non-Coding RNAs and their Integrated Networks." *Journal of Integrative Bioinformatics* 16.3 (2019).
4. Mattick, John S., and Igor V. Makunin. "Non-coding RNA." *Human molecular genetics* 15.suppl_1 (2006): R17-R29.
5. Sanchez Calle, Anna, et al. "Emerging roles of long non-coding RNA in cancer." *Cancer science* 109.7 (2018): 2093-2100.
6. Raveh, Eli, et al. "The H19 Long non-coding RNA in cancer initiation, progression and metastasis–a proposed unifying theory." *Molecular cancer* 14.1 (2015): 184.
7. Holdt, Lesca M., et al. "Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks." *PLoS Genet* 9.7 (2013): e1003588.
8. Han, Siew Ping, et al. "Differential subcellular distributions and trafficking functions of hnRNP A2/B1 spliceoforms." *Traffic* 11.7 (2010): 886-898.
9. Dori, Dov, and Mordechai Choder. "Conceptual modeling in systems biology fosters empirical findings: the mRNA lifecycle." *PloS one* 2.9 (2007): e872.
10. Hedlund, Eva, and Qiaolin Deng. "Single-cell RNA sequencing: technical advancements and biological applications." *Molecular aspects of medicine* 59 (2018): 36-46.
11. Mutz, Kai-Oliver, et al. "Transcriptome analysis using next-generation sequencing." *Current opinion in biotechnology* 24.1 (2013): 22-30.
12. Quail, Michael A., et al. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." *BMC genomics* 13.1 (2012): 1-13.
13. Pease, Jim, and Roy Sooknanan. "A rapid, directional RNA-seq library preparation workflow for Illumina® sequencing." *Nature methods* 9.3 (2012): i-ii.
14. Kloc, Malgorzata, N. Ruth Zearfoss, and Laurence D. Etkin. "Mechanisms of subcellular mRNA localization." *Cell* 108.4 (2002): 533-544.
15. Fujii, Ritsuko, et al. "The RNA binding protein TLS is translocated to dendritic spines by mGluR5 activation and regulates spine morphology." *Current Biology* 15.6 (2005): 587-593.
16. Tiruchinapalli, Dhanrajan M., et al. "Activity-dependent trafficking and dynamic localization of zipcode binding protein 1 and β-actin mRNA in dendrites and spines of hippocampal neurons." *Journal of Neuroscience* 23.8 (2003): 3251-3261.
17. Fu, Xiang-Dong, and Manuel Ares Jr. "Context-dependent control of alternative splicing by RNA-binding proteins." *Nature Reviews Genetics* 15.10 (2014): 689-701.
18. Glisovic, Tina, et al. "RNA-binding proteins and post-transcriptional gene regulation." *FEBS letters* 582.14 (2008): 1977-1986.
19. Jacobsen, Anders, et al. "Signatures of RNA binding proteins globally coupled to effective microRNA target sites." *Genome research* 20.8 (2010): 1010-1019.

20. Xiao, Xiaoxiong, et al. "LncRNA MALAT1 sponges miR-204 to promote osteoblast differentiation of human aortic valve interstitial cells through up-regulating Smad4." *International journal of cardiology* 243 (2017): 404-412.
21. Tao, Fangfang, et al. "miR-211 sponges lncRNA MALAT1 to suppress tumor growth and progression through inhibiting PHF19 in ovarian carcinoma." *The FASEB Journal* 32.11 (2018): 6330-6343.
22. Li, Qiulian, et al. "Disrupting MALAT1/miR-200c sponge decreases invasion and migration in endometrioid endometrial carcinoma." *Cancer letters* 383.1 (2016): 28-40.
23. Sun, Lei, et al. "Long noncoding RNA MALAT1 promotes uveal melanoma cell growth and invasion by silencing of miR-140." *American journal of translational research* 8.9 (2016): 3939.
24. Krol, Jacek, Inga Loedige, and Witold Filipowicz. "The widespread regulation of microRNA biogenesis, function and decay." *Nature Reviews Genetics* 11.9 (2010): 597-610.
25. Winter, Julia, et al. "Many roads to maturity: microRNA biogenesis pathways and their regulation." *Nature cell biology* 11.3 (2009): 228-234.
26. Okamura, Katsutomo, et al. "The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila." *Cell* 130.1 (2007): 89-100.
27. Breving, Kimberly, and Aurora Esquela-Kerscher. "The complexities of microRNA regulation: mirandering around the rules." *The international journal of biochemistry & cell biology* 42.8 (2010): 1316-1329.
28. Miyoshi, Keita, et al. "Characterization of the miRNA-RISC loading complex and miRNA-RISC formed in the Drosophila miRNA pathway." *Rna* 15.7 (2009): 1282-1291.
29. Wang, Xiaowei. "Composition of seed sequence is a major determinant of microRNA targeting patterns." *Bioinformatics* 30.10 (2014): 1377-1383.
30. Laurent, Louise C., et al. "Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence." *Stem cells* 26.6 (2008): 1506-1516.
31. Li, Q., et al. "Downregulation of miR-140 promotes cancer stem cell formation in basal-like early stage breast cancer." *Oncogene* 33.20 (2014): 2589-2600.
32. Wolfson, Benjamin, Gabriel Eades, and Qun Zhou. "Roles of microRNA-140 in stem cell-associated early stage breast cancer." *World Journal of Stem Cells* 6.5 (2014): 591.
33. Loeb, Gabriel B., et al. "Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting." *Molecular cell* 48.5 (2012): 760-770.
34. Jopling, C. L., K. L. Norman, and P. Sarnow. "Positive and negative modulation of viral and cellular mRNAs by liver-specific microRNA miR-122." *Cold Spring Harbor symposia on quantitative biology*. Vol. 71. Cold Spring Harbor Laboratory Press, 2006.
35. Blahna, Matthew T., and Akiko Hata. "Regulation of miRNA biogenesis as an integrated component of growth factor signaling." *Current opinion in cell biology* 25.2 (2013): 233-240.
36. Leuschner, Philipp JF, and Javier Martinez. "In vitro analysis of microRNA processing using recombinant Dicer and cytoplasmic extracts of HeLa cells." *Methods* 43.2 (2007): 105-109.
37. Ganesan, Gayatri, and Satyanarayana MR Rao. "A novel noncoding RNA processed by Drosha is restricted to nucleus in mouse." *Rna* 14.7 (2008): 1399-1410.

38. Hwang, Hun-Way, Erik A. Wentzel, and Joshua T. Mendell. "A hexanucleotide element directs microRNA nuclear import." *Science* 315.5808 (2007): 97-100.
39. Wei, Yao, et al. "Importin 8 regulates the transport of mature microRNAs into the cell nucleus." *Journal of Biological Chemistry* 289.15 (2014): 10270-10275.
40. Huang, Vera, and Long-Cheng Li. "miRNA goes nuclear." *RNA biology* 9.3 (2012): 269-273.
41. Roberts, Thomas C. "The microRNA biology of the mammalian nucleus." *Molecular Therapy-Nucleic Acids* 3 (2014): e188
42. Roy, Bishakha, Larisa M Haupt, and Lyn R Griffiths. "Alternative splicing (AS) of genes as an approach for generating protein complexity." *Current genomics* 14.3 (2013): 182-194.
43. Kalsotra, Auinash, and Thomas A. Cooper. "Functional consequences of developmentally regulated alternative splicing." *Nature Reviews Genetics* 12.10 (2011): 715-729.
44. Poulos, Michael G., et al. "Developments in RNA splicing and disease." *Cold Spring Harbor perspectives in biology* 3.1 (2011): a000778.
45. Wong, Justin J-L., et al. "Orchestrated intron retention regulates normal granulocyte differentiation." *Cell* 154.3 (2013): 583-595.
46. Boutz, Paul L., Arjun Bhutkar, and Phillip A. Sharp. "Detained introns are a novel, widespread class of post-transcriptionally spliced introns." *Genes & development* 29.1 (2015): 63-80.
47. Mauger, Oriane, Frédéric Lemoine, and Peter Scheiffele. "Targeted intron retention and excision for rapid gene regulation in response to neuronal activity." *Neuron* 92.6 (2016): 1266-1278.
48. Wong, Justin J-L., et al. "Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology." *Bioessays* 38.1 (2016): 41-49.
49. Wong, Justin J-L., et al. "Orchestrated intron retention regulates normal granulocyte differentiation." *Cell* 154.3 (2013): 583-595.
50. Jacob, Aishwarya G., and Christopher WJ Smith. "Intron retention as a component of regulated gene expression programs." *Human genetics* 136.9 (2017): 1043-1057.
51. Dvinge, Heidi, and Robert K. Bradley. "Widespread intron retention diversifies most cancer transcriptomes." *Genome medicine* 7.1 (2015): 45.
52. Flodrops, Marion, et al. "TIMP1 intron 3 retention is a marker of colon cancer progression controlled by hnRNPA1." *Molecular Biology Reports* (2020): 1-10.
53. Fujimura, Atsushi, et al. "Expression of a constitutively active calcineurin encoded by an intron-retaining mRNA in follicular keratinocytes." *PLoS One* 6.3 (2011): e17685.
54. Vanichkina, Darya P., et al. "Challenges in defining the role of intron retention in normal biology and disease." *Seminars in cell & developmental biology*. Vol. 75. Academic Press, 2018.
55. Schmitz, Ulf, et al. "Intron retention enhances gene regulatory complexity in vertebrates." *Genome biology* 18.1 (2017): 1-15.
56. Braunschweig, Ulrich, et al. "Widespread intron retention in mammals functionally tunes transcriptomes." *Genome research* 24.11 (2014): 1774-1786.
57. Mayer, Andreas, and L. Stirling Churchman. "A Detailed Protocol for Subcellular RNA Sequencing (subRNA-seq)." *Current protocols in molecular biology* 120.1 (2017): 4-29.

58. Bouvrette, Louis Philip Benoit, et al. "CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells." *Rna* 24.1 (2018): 98-113.
59. Lefebvre, Fabio Alexis, et al. "CeFra-seq: systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing." *Methods* 126 (2017): 138-148.
60. Galante, Pedro Alexandre Favoretto, et al. "Detection and evaluation of intron retention events in the human transcriptome." *Rna* 10.5 (2004): 757-765.
61. Touat-Todeschini, Leila, et al. "Selective termination of lnc RNA transcription promotes heterochromatin silencing and cell differentiation." *The EMBO journal* 36.17 (2017): 2626-2641.
62. Peshkin, Leonid, et al. "On the relationship of protein and mRNA dynamics in vertebrate embryonic development." *Developmental cell* 35.3 (2015): 383-394.
63. Okano, Hideyuki, et al. "Steps toward safe cell therapy using induced pluripotent stem cells." *Circulation research* 112.3 (2013): 523-533.
64. Strumpf, Dan, et al. "Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst." *Development* 132.9 (2005): 2093-2102.
65. Marikawa, Yusuke, and Vernadeth B. Alarcón. "Establishment of trophectoderm and inner cell mass lineages in the mouse embryo." *Molecular Reproduction and Development: Incorporating Gamete Research* 76.11 (2009): 1019-1032.
66. Sasaki, Hiroshi. "Mechanisms of trophectoderm fate specification in preimplantation mouse development." *Development, growth & differentiation* 52.3 (2010): 263-273.
67. Yao, Chunmeng, Wenhao Zhang, and Ling Shuai. "The first cell fate decision in pre-implantation mouse embryos." *Cell Regeneration* 8.2 (2019): 51-57.
68. Rizzino, Angie. "Concise review: The Sox2-Oct4 connection: Critical players in a much larger interdependent network integrated at multiple levels." *Stem cells* 31.6 (2013): 1033-1039.
69. Avilion, Ariel A., et al. "Multipotent cell lineages in early mouse development depend on SOX2 function." *Genes & development* 17.1 (2003): 126-140.
70. Tapia, Natalia, et al. "Dissecting the role of distinct OCT4-SOX2 heterodimer configurations in pluripotency." *Scientific reports* 5 (2015): 13533.
71. Johnson, Martin H., and Josie ML McConnell. "Lineage allocation and cell polarity during mouse embryogenesis." *Seminars in cell & developmental biology*. Vol. 15. No. 5. Academic Press, 2004.
72. Jedrusik, Agnieszka, et al. "Role of Cdx2 and cell polarity in cell allocation and specification of trophectoderm and inner cell mass in the mouse embryo." *Genes & development* 22.19 (2008): 2692-2706.
73. Niwa, Hitoshi, et al. "Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation." *Cell* 123.5 (2005): 917-929.
74. Dietrich, Jens-Erik, and Takashi Hiiragi. "Stochastic patterning in the mouse pre-implantation embryo." *Development* 134.23 (2007): 4219-4231.
75. Wang, Yangming, et al. "DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal." *Nature genetics* 39.3 (2007): 380-385.
76. Singh, Sanjay K., et al. "REST–miR-21–SOX2 axis maintains pluripotency in E14Tg2a. 4 embryonic stem cells." *Stem cell research* 15.2 (2015): 305-311.

77. Trohatou, Ourania, et al. "Sox2 suppression by miR-21 governs human mesenchymal stem cell properties." *Stem cells translational medicine* 3.1 (2014): 54-68.
78. Tay, Yvonne, et al. "MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation." *Nature* 455.7216 (2008): 1124-1128.
79. Nosi, Ursula, et al. "Overexpression of trophoblast stem cell-enriched microRNAs promotes trophoblast fate in embryonic stem cells." *Cell Reports* 19.6 (2017): 1101-1109.
80. Yu, Lili, et al. "Core pluripotency factors promote glycolysis of human embryonic stem cells by activating GLUT1 enhancer." *Protein & cell* 10.9 (2019): 668-680.
81. Dahan, Perrine, et al. "Metabolism in pluripotency: Both driver and passenger?." *Journal of Biological Chemistry* 294.14 (2019): 5420-5429.
82. Ng, Ray Kit, et al. "Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5." *Nature cell biology* 10.11 (2008): 1280-1290.
83. Chen, Ying, et al. "Roles of CDX2 and EOMES in human induced trophoblast progenitor cells." *Biochemical and biophysical research communications* 431.2 (2013): 197-202.
84. Tamm, Christoffer, Sara Pijuan Galitó, and Cecilia Annerén. "A comparative study of protocols for mouse embryonic stem cell culturing." *PloS one* 8.12 (2013): e81156.
85. Ohtsuka, Satoshi, Yoko Nakai-Futatsugi, and Hitoshi Niwa. "LIF signal in mouse embryonic stem cells." *Jak-stat* 4.2 (2015): 1-9.
86. Grigor'eva, Elena V., et al. "FGF4 independent derivation of trophoblast stem cells from the common vole." *PLoS One* 4.9 (2009): e7161.
87. Tanaka, Satoshi, et al. "Promotion of trophoblast stem cell proliferation by FGF4." *Science* 282.5396 (1998): 2072-2075.
88. Ohinata, Yasuhide, and Tomoyuki Tsukiyama. "Establishment of trophoblast stem cells under defined culture conditions in mice." *PloS one* 9.9 (2014): e107308.
89. Oron, Efrat, and Natalia Ivanova. "Cell fate regulation in early mammalian development." *Physical biology* 9.4 (2012): 045002.
90. Nagy, Andras, et al. "Derivation of completely cell culture-derived mice from early-passage embryonic stem cells." *Proceedings of the National Academy of Sciences* 90.18 (1993): 8424-8428.
91. Tanaka, Satoshi, et al. "Promotion of trophoblast stem cell proliferation by FGF4." *Science* 282.5396 (1998): 2072-2075
92. Zaghlool, Ammar, et al. "Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues." *BMC biotechnology* 13.1 (2013): 99.
93. Ben-Yishay, Rakefet, and Yaron Shav-Tal. "The dynamic lifecycle of mRNA in the nucleus." *Current opinion in cell biology* 58 (2019): 69-75.
94. Efroni, Sol, et al. "Global transcription in pluripotent embryonic stem cells." *Cell stem cell* 2.5 (2008): 437-447.
95. Chen, Tao, and Bas van Steensel. "Comprehensive analysis of nucleocytoplasmic dynamics of mRNA in Drosophila cells." *PLoS Genetics* 13.8 (2017): e1006929.
96. Gauthier, Daniel J., and Claude Lazure. "Complementary methods to assist subcellular fractionation in organellar proteomics." *Expert review of proteomics* 5.4 (2008): 603-617.
97. Rapaport, Franck, et al. "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data." *Genome biology* 14.9 (2013): 1-13.

98. Cui, Yi, et al. "LncRNA Neat1 mediates miR-124-induced activation of Wnt/β-catenin signaling in spinal cord neural progenitor cells." *Stem cell research & therapy* 10.1 (2019): 1-11.
99. Tu, Jiajie, et al. "Gas5 is an essential lncRNA regulator for self-renewal and pluripotency of mouse embryonic stem cells and induced pluripotent stem cells." *Stem cell research & therapy* 9.1 (2018): 71.
100. Sun, Zihao, et al. "The long noncoding RNA Lncenc1 maintains naive states of mouse ESCs by promoting the glycolysis pathway." *Stem cell reports* 11.3 (2018): 741-755.
101. Pickard, Mark R., and Gwyn T. Williams. "Molecular and cellular mechanisms of action of tumour suppressor GAS5 LncRNA." *Genes* 6.3 (2015): 484-499.
102. Yu, Qiangfeng, et al. "MALAT1 functions as a competing endogenous RNA to regulate SMAD5 expression by acting as a sponge for miR-142-3p in hepatocellular carcinoma." *Cell & bioscience* 9.1 (2019): 39.
103. Abell, Amy N., et al. "Trophoblast stem cell maintenance by fibroblast growth factor 4 requires MEKK4 activation of Jun N-terminal kinase." *Molecular and cellular biology* 29.10 (2009): 2748-2761.
104. Raghu, Deepthi, et al. "GALNT3 maintains the epithelial state in trophoblast stem cells." *Cell reports* 26.13 (2019): 3684-3697.
105. Yang, Li. "Splicing noncoding RNAs from the inside out." *Wiley Interdisciplinary Reviews: RNA* 6.6 (2015): 651-660.
106. Evans, Ciaran, Johanna Hardin, and Daniel M. Stoebel. "Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions." *Briefings in bioinformatics* 19.5 (2018): 776-792.
107. Zhao, Shanrong, Zhan Ye, and Robert Stanton. "Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols." *RNA* (2020): rna-074922.
108. Li, Xiaohong, et al. "Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies." *BMC genomics* 21.1 (2020): 75.
109. Risso, Davide, et al. "The role of spike-in standards in the normalization of RNA-seq." *Statistical Analysis of Next Generation Sequencing Data*. Springer, Cham, 2014. 169-190.
110. Risso, Davide, et al. "Normalization of RNA-seq data using factor analysis of control genes or samples." *Nature biotechnology* 32.9 (2014): 896-902.
111. Vallejos, Catalina A., et al. "Normalizing single-cell RNA sequencing data: challenges and opportunities." *Nature methods* 14.6 (2017): 565.
112. Conesa, Ana, et al. "A survey of best practices for RNA-seq data analysis." *Genome biology* 17.1 (2016): 13.
113. Law, Charity W., et al. "RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR." *F1000Research* 5 (2016).
114. Williams, Alexander G., et al. "RNA-seq data: challenges in and recommendations for experimental design and analysis." *Current protocols in human genetics* 83.1 (2014): 11-13.
115. Sun, Shiquan, et al. "Differential expression analysis for RNAseq using Poisson mixed models." *Nucleic acids research* 45.11 (2017): e106-e106.

116. Trask, Heidi W., et al. "Microarray analysis of cytoplasmic versus whole cell RNA reveals a considerable number of missed and false positive mRNAs." *RNA* 15.10 (2009): 1917-1928.
117. Kilchert, Cornelia, Sina Wittmann, and Lidia Vasiljeva. "The regulation and functions of the nuclear RNA exosome complex." *Nature Reviews Molecular Cell Biology* 17.4 (2016): 227.
118. Solnestam, Beata Werne, et al. "Comparison of total and cytoplasmic mRNA reveals global regulation by nuclear retention and miRNAs." *BMC genomics* 13.1 (2012): 1-9.
119. Halpern, Keren Bahar, et al. "Nuclear retention of mRNA in mammalian tissues." *Cell reports* 13.12 (2015): 2653-2662.
120. Percharde, Michelle, Priscilla Wong, and Miguel Ramalho-Santos. "Global hypertranscription in the mouse embryonic germline." *Cell reports* 19.10 (2017): 1987-1996.
121. Gaspar-Maia, Alexandre, et al. "Open chromatin in pluripotency and reprogramming." *Nature reviews Molecular cell biology* 12.1 (2011): 36-47.
122. Tsang, Jason CH, et al. "Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells." *Genome biology* 16.1 (2015): 1-16.
123. Kolodziejczyk, Aleksandra A., et al. "Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation." *Cell stem cell* 17.4 (2015): 471-485.
124. Tsogtbaatar, Enkhtuul, et al. "Energy Metabolism Regulates Stem Cell Pluripotency." *Frontiers in Cell and Developmental Biology* 8 (2020).
125. Juan, Aster H., et al. "Roles of H3K27me2 and H3K27me3 examined during fate specification of embryonic stem cells." *Cell reports* 17.5 (2016): 1369-1382.
126. Kohan-Ghadr, Hamid-Reza, et al. "Potential role of epigenetic mechanisms in regulation of trophoblast differentiation, migration, and invasion in the human placenta." *Cell adhesion & migration* 10.1-2 (2016): 126-135.
127. Mentch, Samantha J., and Jason W. Locasale. "One carbon metabolism and epigenetics: understanding the specificity." *Annals of the New York Academy of Sciences* 1363.1 (2016): 91.
128. Sharma, Shiv K., et al. "(Bis) urea and (bis) thiourea inhibitors of lysine-specific demethylase 1 as epigenetic modulators." *Journal of medicinal chemistry* 53.14 (2010): 5197-5212.
129. Folmes, Clifford DL, et al. "Somatic oxidative bioenergetics transitions into pluripotency-dependent glycolysis to facilitate nuclear reprogramming." *Cell metabolism* 14.2 (2011): 264-271.
130. Wang, Jian, et al. "Dependence of mouse embryonic stem cells on threonine catabolism." *Science* 325.5939 (2009): 435-439.
131. Senner, Claire E., et al. "TET1 and 5-Hydroxymethylation Preserve the Stem Cell State of Mouse Trophoblast." *Stem Cell Reports* (2020).
132. Ueno, Masaya, et al. "c-Met-dependent multipotent labyrinth trophoblast progenitors establish placental exchange interface." *Developmental cell* 27.4 (2013): 373-386.

133. Zita, Matteo Moretto, et al. "Gene expression profiling reveals a novel regulatory role for Sox21 protein in mouse trophoblast stem cell differentiation." *Journal of Biological Chemistry* 290.50 (2015): 30152-30162.
134. Gheldof, Alexander, and Geert Berx. "Cadherins and epithelial-to-mesenchymal transition." *Progress in molecular biology and translational science*. Vol. 116. Academic Press, 2013. 317-336.
135. Latos, Paulina Anna, and Myriam Hemberger. "From the stem of the placental tree: trophoblast stem cells and their progeny." *Development* 143.20 (2016): 3650-3660.
136. Fendereski, Mona, et al. "Mouse Trophoblasts Can Provide Antiviral Protection to Embryonic Stem Cells." *The FASEB Journal* 34.S1 (2020): 1-1.
137. Aikawa, Hiroaki, et al. "Innate immunity in an in vitro murine blastocyst model using embryonic and trophoblast stem cells." *Journal of bioscience and bioengineering* 117.3 (2014): 358-365.
138. Ullrich, Sebastian, and Roderic Guigó. "Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development." *Nucleic acids research* 48.3 (2020): 1327-1340.
139. Edwards, Christopher R., et al. "A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages." *Blood, The Journal of the American Society of Hematology* 127.17 (2016): e24-e34.
140. Biamonti, Giuseppe, et al. "The alternative splicing side of cancer." *Seminars in cell & developmental biology*. Vol. 32. Academic Press, 2014.
141. Kouyama, Yuta, et al. "Oncogenic splicing abnormalities induced by DEAD-Box Helicase 56 amplification in colorectal cancer." *Cancer science* 110.10 (2019): 3132.
142. Naro, Chiara, and Claudio Sette. "Timely-regulated intron retention as device to fine-tune protein expression." *Cell Cycle* 16.14 (2017): 1321.
143. Araki, Shinsuke, et al. "Inhibitors of CLK protein kinases suppress cell growth and induce apoptosis by modulating pre-mRNA splicing." *PloS one* 10.1 (2015): e0116929.
144. Uzor, Simon, et al. "Autoregulation of the human splice factor kinase CLK1 through exon skipping and intron retention." *Gene* 670 (2018): 46-54.
145. Shi, Yongsheng, and James L. Manley. "A complex signaling pathway regulates SRp38 phosphorylation and pre-mRNA splicing in response to heat shock." *Molecular cell* 28.1 (2007): 79-90.
146. Alfonso, Julieta, et al. "Identification of genes regulated by chronic psychosocial stress and antidepressant treatment in the hippocampus." *European Journal of Neuroscience* 19.3 (2004): 659-666.
147. Quinn, Jeffrey J., and Howard Y. Chang. "Unique features of long non-coding RNA biogenesis and function." *Nature Reviews Genetics* 17.1 (2016): 47.
148. Fico, Annalisa, et al. "Long non-coding RNA in stem cell pluripotency and lineage commitment: functions and evolutionary conservation." *Cellular and Molecular Life Sciences* 76.8 (2019): 1459-1471.
149. Zhang, Jie, et al. "LncRNA NORAD contributes to colorectal cancer progression by inhibition of miR-202-5p." *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* 26.9 (2018): 1411-1418.

150. Xuan, Yi, and Yanong Wang. "Long non-coding RNA SNHG3 promotes progression of gastric cancer by regulating neighboring MED18 gene methylation." *Cell death & disease* 10.10 (2019): 1-12.
151. Hu, Litian, et al. "Long noncoding RNA GAS5 suppresses the migration and invasion of hepatocellular carcinoma cells via miR-21." *Tumor Biology* 37.2 (2016): 2691-2702.
152. Chen, Wei-Yu, et al. "MicroRNA-34a regulates WNT/TCF7 signaling and inhibits bone metastasis in Ras-activated prostate cancer." *Oncotarget* 6.1 (2015): 441.
153. Soares, Ricardo J., et al. "Evaluation of fluorescence in situ hybridization techniques to study long non-coding RNA expression in cultured cells." *Nucleic acids research* 46.1 (2018): e4-e4.
154. Bekhite, Mohamed M., et al. "VEGF-mediated PI3K class IA and PKC signaling in cardiomyogenesis and vasculogenesis of mouse embryonic stem cells." *Journal of cell science* 124.11 (2011): 1819-1830.
155. Cherepkova, Maria Y., Galina S. Sineva, and Valery A. Pospelov. "Leukemia inhibitory factor (LIF) withdrawal activates mTOR signaling pathway in mouse embryonic stem cells through the MEK/ERK/TSC2 pathway." *Cell death & disease* 7.1 (2016): e2050-e2050.
156. Zhang, Xin, et al. "FOXO1 is an essential regulator of pluripotency in human embryonic stem cells." *Nature cell biology* 13.9 (2011): 1092-1099.
157. Watabe, Tetsuro, and Kohei Miyazono. "Roles of TGF-β family signaling in stem cell renewal and differentiation." *Cell research* 19.1 (2009): 103-115
158. Park, Kyung-Soon. "Tgf-Beta family signaling in embryonic stem cells." *International journal of stem cells* 4.1 (2011): 18.
159. Yin, Ke-Jie, et al. "Vascular endothelial cell-specific microRNA-15a inhibits angiogenesis in hindlimb ischemia." *Journal of Biological Chemistry* 287.32 (2012): 27055-27064.
160. Mogilyansky, Elena, and Isidore Rigoutsos. "The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease." *Cell Death & Differentiation* 20.12 (2013): 1603-1614.
161. Johnson, Charles D., et al. "The let-7 microRNA represses cell proliferation pathways in human cells." *Cancer research* 67.16 (2007): 7713-7722.
162. Hydbring, Per, et al. "Identification of cell cycle-targeting microRNAs through genome-wide screens." *Cell Cycle* 16.23 (2017): 2241-2248.
163. Lu, Chi-Wei, et al. "Ras-MAPK signaling promotes trophectoderm formation from embryonic stem cells and mouse embryos." *Nature genetics* 40.7 (2008): 921-926.
164. Zhang, Yi-Lei, et al. "Roles of Rap1 signaling in tumor cell migration and invasion." *Cancer biology & medicine* 14.1 (2017): 90.
165. Latreille, Mathieu, et al. "MicroRNA-7a regulates pancreatic β cell function." *The Journal of clinical investigation* 124.6 (2014): 2722-2735.
166. Parnell, Laurence D., et al. "BioStar: an online question & answer resource for the bioinformatics community." *PLoS Comput Biol* 7.10 (2011): e1002216.