CHARACTERISATION OF ANTISENSE NONCODING RNAS PRESENT DURING EARLY ZEBRAFISH DEVELOPMENT

Kumar Sanjana Pillay

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Biosciences
College of Life & Environmental Sciences
University of Birmingham

December 2019

ABSTRACT

Long non-coding RNA’s are a broad class of non-protein coding RNAs >200 nucleotide in length. Although the importance of lncRNAs in various cellular functions is increasingly becoming clear, their role in regulating vertebrate development still remains unexplored. A recent study on gene expression during zebrafish development identified 1133 lncRNAs, a majority of these lncRNAs (566) were antisense ncRNA (Pauli et al., 2012). Antisense lncRNAs, a class of lncRNAs, have been shown to act either as positive or negative regulators of overlapping protein-coding mRNAs.

To understand the role of antisense RNAs in early vertebrate development, we took a data mining approach. A computational analysis of RNA sequencing data from zebrafish development indicates that a number of antisense RNAs are present in embryos even before the zygotic gene transcription is initiated indicating that they are probably of maternal origin. In addition, our results hint at the role of the antisense RNAs in regulating transcription of developmental genes. Additionally, implying their role in zygotic genome activation. To further investigate this hypothesis, we employed transcriptomics methods to examine RNAs from the nuclear and cytosolic fractions from zebrafish embryos at different stages of development. The RNA-seq and CAGE-seq revealed spatial and temporal regulation of lncRNA expression during zebrafish development. It further revealed differences in promoter width and TSS usage between the nuclear and cytosolic fractions.

ACKNOWLEDGMENT

A lot of people are involved in the successful completion of this thesis. Firstly, I would like to thank my supervisor, Dr. Aditi Kanhere, for her immense support and guidance throughout my PhD. She was very helpful and supported enormously in my overall growth as a researcher. I would also like to thank my second supervisor, Prof. Ferenc Mueller, who helped me immensely by sharing his vast knowledge in the field of zebrafish and for providing the fishes without which my research could not have been possible. I would also like to take this opportunity to greet all my ZENCODE fellow researchers and friends for the knowledge and laughs we exchanged in the past 3 years. My warmest thanks to all the people of the 8th floor who were always eager to lend freights of reagents. The lunch and tea times chat with cakes and chocolates that we shared together. To my friend and lab mates, Sue, Diaa and Rhian, a big thank you for all the support, encouragement and the exchange of knowledge. Thanks to all my dear friends Elis, Zuhal, Justyna, Varvara, Emily, Sandro and Tim for their incredible support and love ever since I moved to Birmingham.

I would also like to thank my parents and my sister for everything they have done for me, for their constant support and prayers. Thank you to my mum-in-law and father-in-law who have constantly wished for my success. Finally, and especially, thank you to my dearest husband Ravi. You have been very patient and supportive of everything. Thank you for always being there for me and believing in me.

Table of Contents

List of Figures

List of Tables

ABBREVIATIONS

AIRN	Antisense to Igf2r RNA
AP	Anterior-posterior
AS RNAS	Antisense RNAs
ATP	Adenosine triphosphate
BB	Balbiani body
BRB	Brom bones
BRE	TFIIB Recognition Element
BUC	Bucky ball
C-MYC	V-myc avian myelocytomatosis viral oncogene homolog
CAGE	Cap Analysis of Gene Expression
CDK-6	Cyclin dependent kinase-6
CDNA	Complementary deoxyribose nucleic acid
CEI	Cellular island
CHIP-SEQ	Chromatin ImmunoPrecipitation sequencing
CHIRP-SEQ	Chromatin isolation by RNA Purification sequencing
CHX	Cycloheximide
CRISPRI	Clustered Regularly Interspaced Short Palindromic Repeats interference
CTD	Carboxyl terminal repeat domain
DAVID	Database for Annotation, Visualization and Integrated Discovery
DCE	Downstream core element
DNMT1	DNA methyltransferase 1
DPE	Downstream Promoter Element
DPF	Days post fertilization
DSRNA	Double strand RNA
DV	Dorsal-ventral
EDEN-BP	Embryonic Deadenylation Element Binding Protein
ENCODE	Encyclopedia of DNA Elements
ERNAS	Enhancer RNAs
FPKM	Fragments Per Kilobase of transcript per Million
GO	Gene ontology
GTF’S	General transcription factors
hESCs	Human embryonic stem cells
H3K27AC	Histone H3 lysine 27 acetylation
H3K27ME3	Histone H3 lysine 27 trimethylation
H3K36ME3	Histone H3 lysine 36 trimethylation
H3K4ME3	Histone H3 lysine 4 trimethylation
H3K9ME	Histone H3 lysine 9 methylation
HOTAIR	Hox transcript antisense RNA
HOTTIP	Hox transcript
HPF	Hours post fertilisation
IGF2R	Insulin growth factor receptor 2
INR	Initiator element
IP3	Inositol 1,4,5-triphosphate
IPSC	Induced pluripotent stem cells
KCNQ1OT1	Kcnq1 overlapping transcript 1

LET-7	Lethal-7
LINCRNAS	Long intergenic RNAs
LINE	Long interspersed nucleotide element
LNCRNAS	Long non-coding RNAs
LTR	Long terminal repeat
MBT	Mid-blastula transition
mESCs	Mouse embryonic stem cells
MIRNAS	MicroRNAs
MLL1	Mixed-Lineage leukemia 1
MTE	Motif ten element
MZDICER	Maternal-zygotic mutant of dicer
MZT	Maternal-to-zygotic transition
N/C	Nucleocytoplasmic ratio
NCRNAS	Non-coding RNAs
PARS	Promoter-associated RNAs
PCR	Polymerase chain reaction
PIRNAS	PIWI-interacting RNAs
RNA POL II	RNA Polymerase II
PRC2	Polycomb repressive complex
PTMS	Post-translational modification
PU.1	Psutative oncogene Spi-1
RNA-SEQ	RNA-sequencing
RNP	Ribonucleoprotein
RRNA	Ribosomal RNA
SINE	Short interspersed nucleotide element
SIRNAS	Small interfering RNAs
SNORNAS	Small nucleolar RNAs
SNORNPS	Small nucleolar ribonucleoprotein
SNRNA	Small nuclear RNA
SOMs	Self organizing maps
STAR	Spliced transcripts alignment to a reference
TBP	TATA-binding protein
TCs	Tag clusters
TGF-βr	Transforming growth factor β receptor
TFS	Transcription factors
TRNA	Transfer RNA
TSS	Transcription start site
TSSA	Transcription start site associated
TUTASE4	Terminal uridylyl transferase 4
UASS	Upstream activating sequences
URSS	Upstream repressing sequences
VNTR	Variable nucleotide tandem repeats
XA	Active X-chromosome
XCPE1	X core promoter element 1
XIC	X-inactivation centre
YSL	Yolk syncytial layer
ZGA	Zygotic genome activation

CHAPTER 1: INTRODUCTION

1.1 Zebrafish as a model for vertebrate development

There have been many studies in the past demonstrating the resemblance of specific stages in all vertebrate embryos with development such as neurula stage, pharyngula stage (Richardson et al., 1997), or the tailbud stage (Slack et al., 1993). The Richardson et al. study showed a highly conserved developmental stage, late tailbud, the majority of vertebrate embryos possess somites, neural tube, notochord, optic anlagen and pharyngeal pouches (Richardson et al., 1997). Due to the conservation of developmental stages, studies on embryogenesis in other vertebrates like fish, amphibians and mouse are used to understand human embryonic development.

Zebrafish (*Danio rerio)* is a freshwater fish native to the streams of Himalayan region. Zebrafish derives its common name because of its physical appearance, the presence of black and white horizontal stripes. Zebrafish is one of the popular animal models which is routinely used to understand early development. Zebrafish has several advantages over other well-established vertebrate model organisms like mouse. They can reproduce throughout the year and zebrafish produces much larger number of offsprings per generation. Unlike in case of mouse, zebrafish embryos remain translucent throughout development, allowing us to observe them during embryogenesis. The embryos are relatively small and easy to maintain at higher densities. In comparison to other vertebrates, they are less expensive and easier to maintain. In addition, external fertilization process makes it easier to carry out genetic manipulation which is very useful in developmental studies. The embryos develop rapidly within a span of 3 days post fertilization (dpf) which is much time-saving as compared to other vertebrate model organisms such as frogs (21 dpf) and mouse (20-27 dpf).

1.1.1 Zebrafish embryogenesis & the stages of development

In Zebrafish, embryonic development starts by fertilization of externally laid eggs and spans across a period of 3 dpf. Initially, the embryo undergoes 10 rapid and asynchronous cell divisions which is followed by lengthening of the cell cycle. Zebrafish development is broadly categorised into seven periods - the zygote, cleavage, blastula, gastrula, segmentation, pharyngula and hatching (Kimmel et al., 1995) (Fig 1.1.1- 1.1.4). The focus of this study is on the early stages of development up to cell cycle 16 (gastrulation). Hence, for simplicity I will only be discussing in detail about the cleavage periods from zygote to gastrula period in this review of literature.

1.1.1.1 Zygote period (0-0.75 hours post fertilisation (hpf))

The formation of a zygote depends on the initial fusion of the sperm and egg, the respective male and female gametes. This process is termed as fertilization which is similar for most animals. In zebrafish, activation of the egg occurs by coming in contact with water which within a few seconds triggers the formation of an outer chorion layer to prevent polyspermy. The activation also prompts the completion of the second meiotic division in zebrafish oocytes. The sperm only enters the oocyte at a specific location in the animal pole called the micropyle. Once inside the egg, the sperm disassembles releasing the pronucleus which fuses with the egg pronucleus to form the zygotic nucleus triggering the first mitotic division. In zebrafish, the process takes about 40 minutes after fertilisation (zygote period) until the first cleavage occurs. During this time, the chorion swells, and the formation of a blastodisc from the cytoplasm at the animal pole occurs (Fig 1.1.1a) (Kimmel et al., 1995).

1.1.1.2 Cleavage period (0.75-2.25 hpf)

This period consists of six cell cycle divisions each occurring after an interval of 15 mins at regular orientations. Therefore, the embryos can be easily staged by the arrangement of their cells, or blastomeres (Kimmel et al., 1995). It includes six stages corresponding to each division cycle, i.e. the two-cell, four-cell, eight-cell, 16-cell, 32-cell and 64-cell stage. The first division cleaves the one cell embryo incompletely, leaving the yolk region intact, from the animal pole progressing to the vegetal pole into two blastomeres of equal length. Until cell cycle 5, all cell divisions are vertically oriented (Fig 1.1.1b).

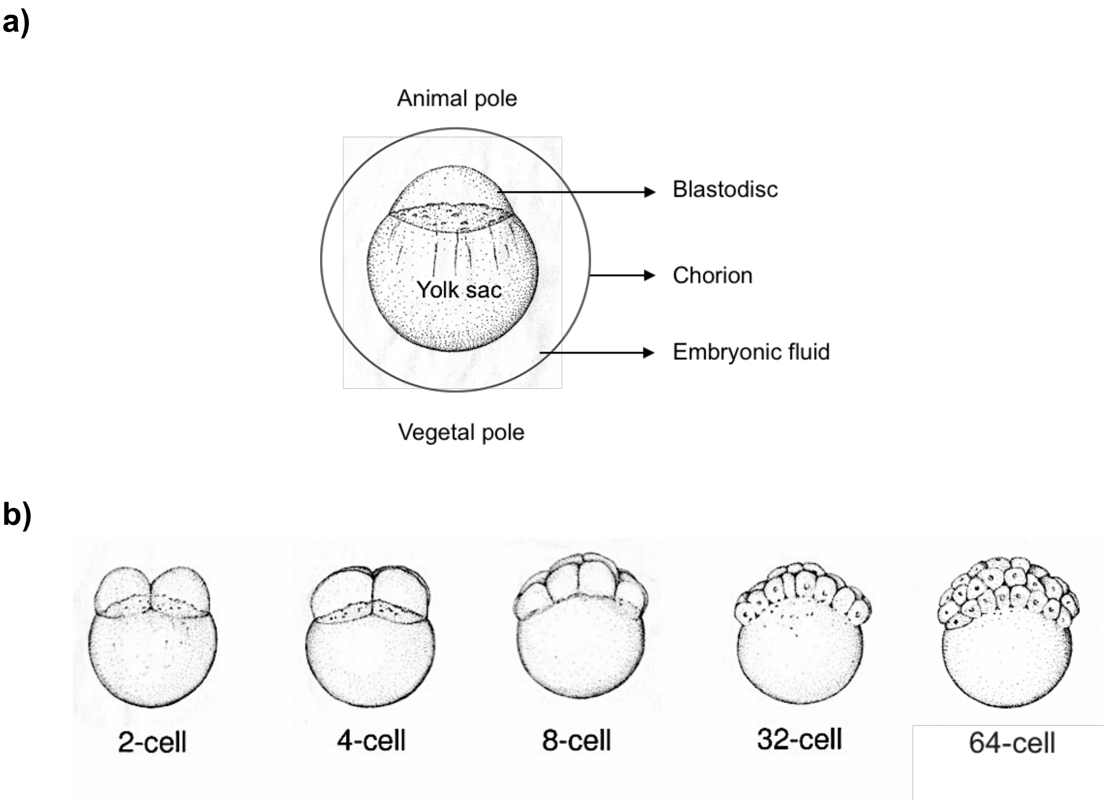


Figure 1.1.1 Early stages of development.
a) One-cell stage*. The zygote period involves the formation of a one-cell embryo by first mitotic division resulting in the movement of the cytoplasm towards the animal pole to form the blastodisc separating it from the yolk.* ***b) Cleavage period****. It involves six cleavages occurring at a regular interval and appear to be specialised. The specialisation includes the redistribution of cell fate determinants present in the egg. The figure shows five of the six stages in cleavage period. Figure taken from (Kimmel et al., 1995).*

1.1.1.3 Blastula period (2.25-5.25 hpf)

This period is so called because of the ball-like arrangement of cells, which starts with 128 cells and concludes at the fifteenth cell division at 30% epiboly (Fig 1.1.2) (Kimmel et al., 1995). The blastula period is characterised by several significant processes including the midblastula transition (MBT), formation of the yolk syncytial layer (YSL) and the movement of the cells to form the germ layers called epiboly. The stage of an embryo is determined by looking at the size rather than the position of the blastomeres as the positioning of the cleavage planes is asymmetrical during this period. The early divisions are “metasynchronous” with mitosis occurring at different times in different cells based on their orientation (Kimmel et al., 1995). Midblastula transition (MBT) occurs at the tenth cell cycle division and is distinguished by an increase in the timing of the cell cycle with asynchronous cell divisions. It also marks the formation of the yolk syncytial layer (YSL) which arises by the fusion of the cytoplasm from the marginal blastomeres that have shrunken with the yolk cell. All the YSL nuclei undergo rapid divisions at the same time for three cell cycles during MBT, without accompanying cytoplasmic divisions, after which the nuclei stop dividing and increase in size. The movement of the YSL and the blastodisc over the yolk in the late blastula is called epiboly which continues until the gastrulation period. The blastodisc undergoes considerable changes in shape due to the superficial movement of the blastomeres embedded inside the embryo. However, the marginal blastomeres intermingle to a much lesser extent, as compared to the central ones, owing to the fact that they later form the mesoderm and thus have cellular pattern or are fated (Kimmel et al., 1995).

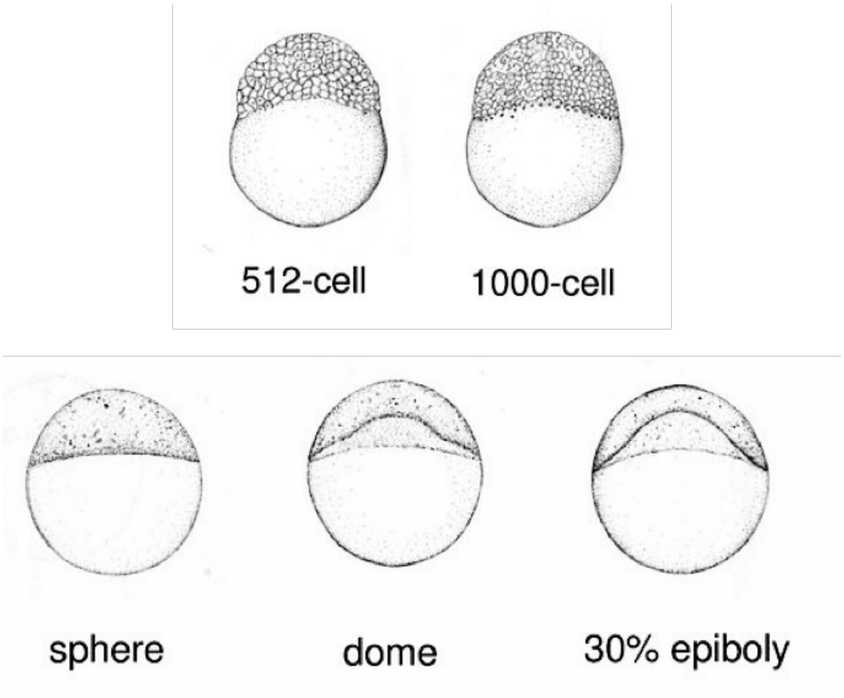


***Figure 1.1.2 Blastula period**.*
It consists of nine different stages namely, 128-cell, 256-cell, 512-cell, 1000-cell, high, oblong, sphere, dome and 30% epiboly stage. The figure above shows five of the nine stages in blastula period. 512-cell stage marks the beginning of the midblastula transition with dome showing the first signs of epiboly. Figure taken from (Kimmel et al., 1995).

1.1.1.4 Gastrula period (5.25-10 hpf)

Gastrula period is characterised by the formation of the primary germ layers and the dorsal-ventral axis of the embryo due to the continuation of epiboly (Fig 1.1.3). The beginning of involution at 50% epiboly leads to the formation of a germ ring all around the blastoderm rim by folding of the latter on itself (Kimmel et al., 1995). The accumulation of cells due to convergence movements on the germ ring leads to the formation of an embryonic shield which also marks the potential dorsal side of embryo. While the future dorsal side forms, epiboly is briefly impeded but resumes once the germ ring is created. The germ ring consists of two layers, the epiblast which forms the upper layer and feeds cells to the lower layer, called the hypoblast. The epiblast at the end of gastrulation gives rise to the ectoderm which forms the epidermis while the hypoblast produces the so called mesendoderm that gives rise to both the endoderm and the mesoderm. By mid-gastrulation, the cells lose their plasticity and become more committed, for example, if a cell is removed from the hypoblast and planted in the epiblast, it would try to return to the hypoblast and acquire hypoblast-derived tissues. Although in the beginning of gastrulation the cells behave more in a lineage-restricted manner they are not yet committed to a particular fate (Kimmel et al., 1995). The dorsal-ventral (DV) position of cells in the early gastrula relates to the anterior-posterior (AP) axis position of cells in the pharyngula period.

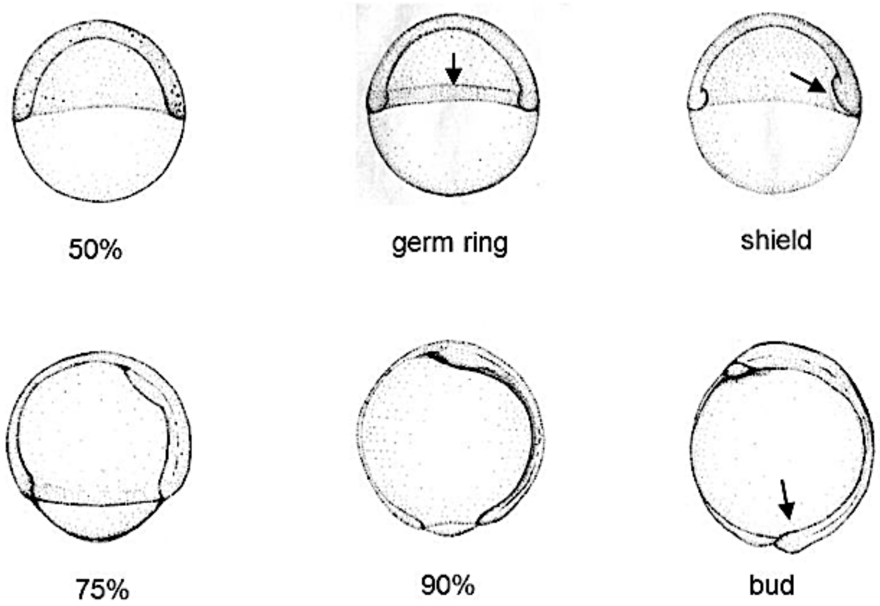


Figure 1.1.3 Gastrula period*.*
This period involves six stages of development, 50% epiboly, germ-ring, shield, 75% epiboly, 90% epiboly and the bud stage. The arrows mark the position of the germ ring, embryonic shield and the tail bud in the respective stages. Figure taken from (Kimmel et al., 1995).

1.1.1.5 Segmentation period (10-24 hpf)

This period is marked by the appearance of rudiments of primary organs, the formation of somites and the appearance of a tail at the caudal end which gives the period its name, “tail bud” (Fig 1.1.4). The mesodermal lining the neural tube on both sides forms the somites which in the future gives rise to the vertebrae and muscles. One can stage the embryos based on the number of somites present, with the first six emerging at three every hour and the later six at two per hour. As outlined in Fig 1.1.4, this period comprises of one-somite stage (10.25 hpf), five-somite stage (11.75 hpf), 14-somite stage (16 hpf), 20-somite stage (19 hpf) and 26-somite stage (22 hpf) (Kimmel et al., 1995).

1.1.1.6 Pharyngula period (24-48 hpf)

This period is marked by the presence of a complete notochord and a well-developed set of 30 somites. It consists of the prim-5 stage (24h), the prim-15 (30 hpf), prim-25 (36 hpf) and the high pec stage (42 hpf) (Fig 1.1.4). The term prim comes from the primordia of pharyngeal arches which are in the initial stages of development. The embryos can be staged based on the straightening of the head of the embryo which is accompanied by an extension of the tail. The fins start appearing with the median fin fold appearing in the prim-5 stage. There is also some level of pigment synthesis in the prim-5 stage along with the appearance of a cone shaped heart. The rudiments of the pectoral fin bud start appearing by the high pec stage with intensification of pigment synthesis (Kimmel et al., 1995)

1.1.1.7 Hatching period (48-72 hpf)

Different embryos synchronised by stage in a single developing clutch may or may not hatch at the same time. However, the embryos that do not hatch still continue to develop without any problem. The rudiments of all the major organs are now completely developed and staging of the embryo can be done by tracking the pectoral fin development. It includes the long-pec stage (48 hpf), the pec-fin stage (60 hpf) and the protruding-mouth stage (72 hpf) (Fig 1.1.4). The yolk sac starts depleting and the head increases in size such that they are approximately equal in size. The pigmentation in the retina becomes denser. The heart starts pumping by the end of the pec-fin stage. The protrusion of the mouth begins, and the circulation of blood is prominent (Kimmel et al., 1995).

1.1.1.8 Early larval period

The embryo, after 3dpf, is now called a larva and starts swimming actively. The mouth continues to protrude with the development of the swim bladder. The pectoral fins, eyes and the movement of the jaw is also active. All these changes help the larvae in looking for prey and start feeding on its own (Kimmel et al., 1995). There are specific maternal genes present at different timepoints (outlined in Fig 1.2.2) that are essential during early zebrafish embryogenesis.

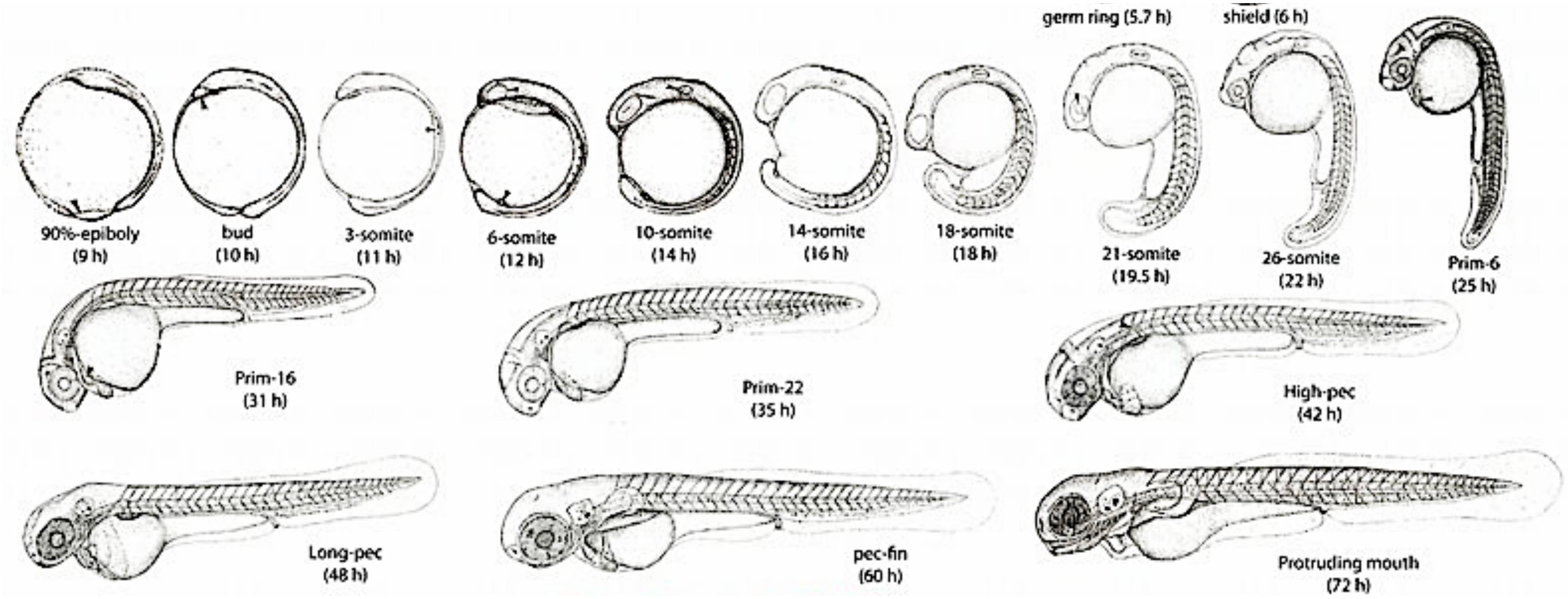


Figure 1.1.4 Stages in zebrafish development.

It shows the stages involved in zebrafish embryogenesis from the segmentation (9hpf) to larval (72hpf) period. The first row shows events occurring in the segmentation period (from 3-somite to 26-somite stages). The second row covers the pharyngula (prim6 to high pec) while the third row shows stages involved in hatching period (long pec to protruding mouth). Figure source http://www.uoneuro.uoregon.edu/k12/zfk12.html.

1.2 Maternal-to-Zygotic Transition – a key player in vertebrate development

During development the vertebrate embryo is in a transcriptionally inactive state for a few initial cell divisions. As result, during this period of inactive genome early development of the embryo is completely dependent on maternally provided products (Fig 1.2.1). As development progresses, the transcription of zygotic genome is triggered and simultaneous clearance of maternal RNAs and proteins leads to their replacement with newly synthesized zygotic RNAs. This process is called maternal-to-zygotic transition (MZT) (Lee et al., 2014). Main events leading to the activation of the zygotic genome is firstly, the clearance of maternal gene products, and secondly, transcription of genes (mostly transcription factors) responsible for the activation of zygotic genes needed for late development. The timing of the maternal-to-zygotic transition varies in different organisms as highlighted in Fig 1.2.1. In zebrafish, MZT coincides with mid-blastula transition (MBT) and occurs at cell cycle 10 or 3 hpf at the 1000-cell stage. In mouse, the maternal-to-zygotic transition occurs after the first cleavage which takes about 22 hpf.

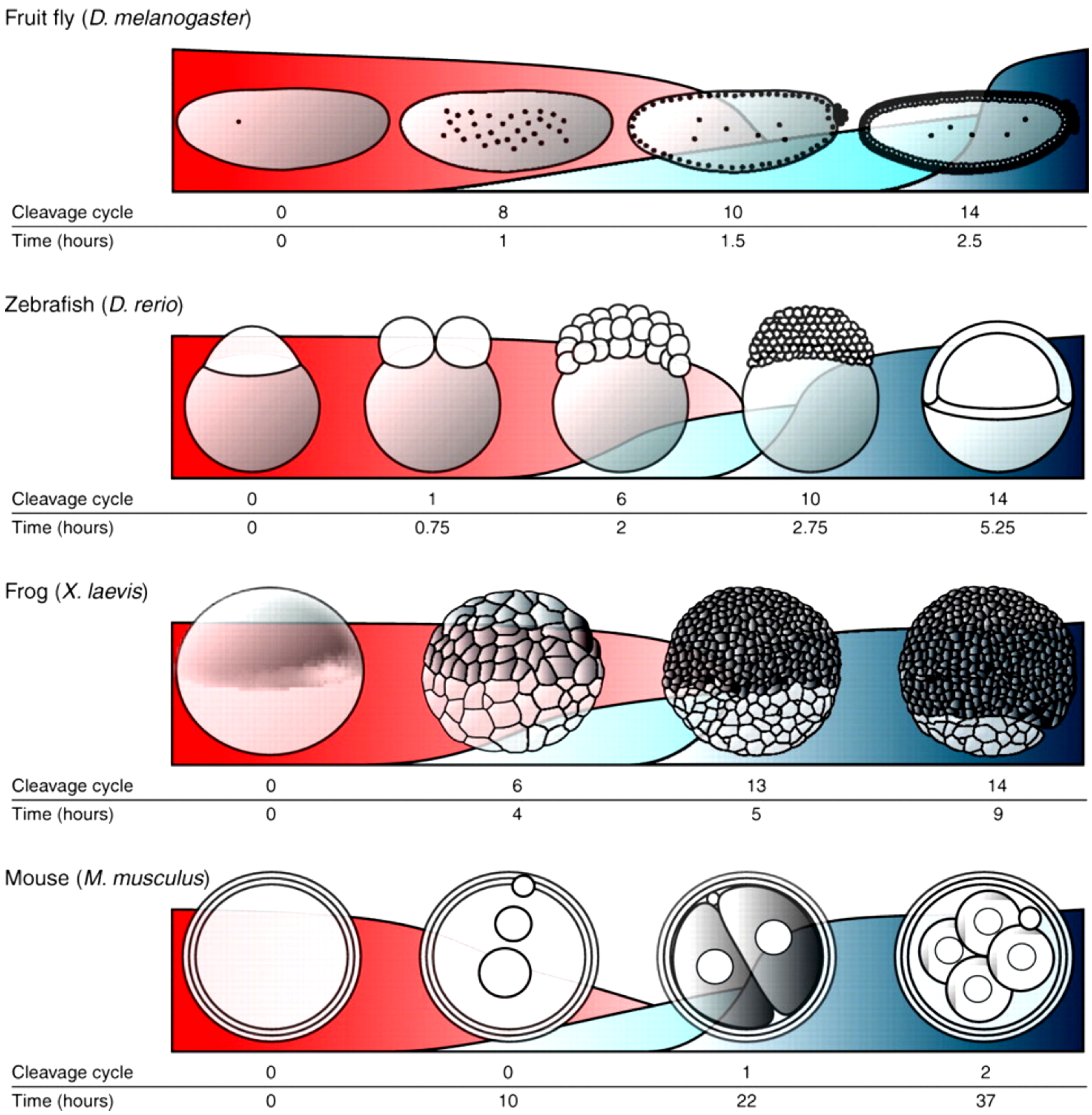
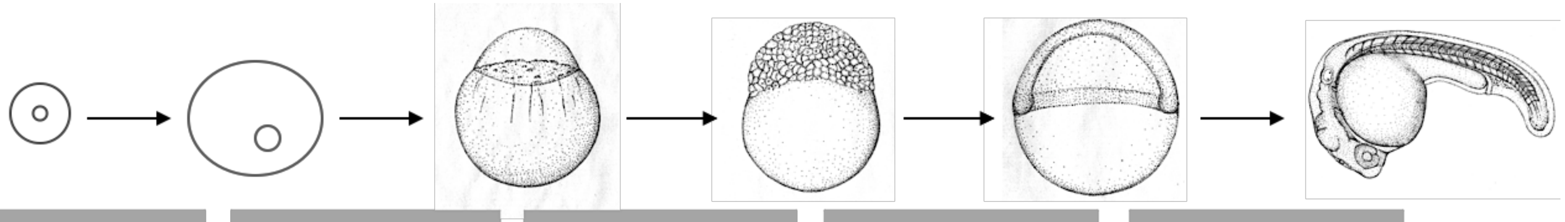


Figure 1.2.1 The maternal-to-zygotic transition in different model organisms. *The figure shows the RNA levels in different stages of development. the areas in red indicates maternal RNA levels and the light and dark blue areas represent the minor and major wave of zygotic RNA expression, respectively. The X-axis represents time in development (in hours). Figure taken from (Tadros and Lipshitz, 2009).*

1.2.1 Maternal-effect genetic screens

Zebrafish has emerged as a powerful model organism for the study of vertebrate development. This is due to the vast availability of forward genetic screens for mutations affecting early development in zebrafish (Driever et al., 1996, Hammerschmidt et al., 1996, Kane et al., 1996a, Kane et al., 1996b, Solnica-Krezel et al., 1996, Stemple et al., 1996). Since early development is primarily driven by maternal RNAs identifying mutations in these genes would help us in understanding their role in development. As suggested in Fig 1.2.2 these mutants have been divided into categories based on the development process that they affect. The process of oocyte development, for example, is driven by several maternal genes such as the *bucky ball* (*buc*). The mutant for this gene lacks a functional Balbiani body (Bb) which is a marker for the establishment of asymmetry (animal-vegetal axis). Therefore, the *buc* gene is essential in the formation of animal-vegetal axis during early oocyte development (Marlow and Mullins, 2008). Activation of the egg is defective in *brom bones* (*brb*) gene mutants which is essential for Ca^{2+} mediated egg activation via the inositol 1,4,5-triphosphate (IP3) signalling pathway (Mei et al., 2009). Mutations in *cellular island* (*cei*) gene showed defects in early cleavage process due the absence of Aurora B Kinase, synthesised by this gene and essential in cytokinesis (Yabe et al., 2009). The role of transcription factor *pou2/oct4* gene in epiboly and body patterning has been further discussed in the coming sections (1.2.3).



germ cell/oocyte development	egg activation	cleavage	epiboly	body plan
bucky ball nanos over easy ruehrei soufflé sunny side up zili	brom bones claustro emulsion jump start under repair	acytokinesis atmos aura barette bo peep cellular atoll cellular island cobblestone dicer futile cycle golden gate indivisible irreducible kawai neeble screeching halt waldo weeble	bedazzled betty boop poky slow pou2	blistered ichabod hecate mission impossible pollywog pou2 pug tokkaebi

Figure 1.2.2 Genes present during early zebrafish embryogenesis*.*
The figure above gives a list of genes that are expressed during the early stages of development. These genes are essential in the process of oocyte development, activation of the egg, cleavage, epiboly and body plan or segmentation. Figure reproduced from (Abrams and Mullins, 2009).

1.2.2 Maternal-to-zygotic transition is regulated by several factors

The first cleavage in a fertilized embryo in zebrafish takes about 40 mins to complete after which the cells divide rapidly approximately every 15 mins. These rapid cell divisions result in increasing the amount of DNA and number of nuclei; however, the overall cytosolic content of the embryo remains same. As these early cell divisions last for only about 15 mins, they ensure no transcription takes place as transcription of genes would be interrupted by the DNA replication machinery. The first few divisions are synchronous lacking a gap phase followed by asynchronous divisions with the introduction of a gap phase during mid-blastula transition (MBT) (Scharf and Gerhart, 1980). In most organisms, including zebrafish, mid-blastula transition coincides with the maternal-to-zygotic transition or, in other words, zygotic genome activation. There are several mechanisms suggested by which MZT is triggered, one of which is the dilution of transcription repressors as evidenced by a study on frog (*Xenopus laevis*) embryos. The transcriptional repressor identified is a DNMT1 (DNA methyltransferase I) protein whose depletion leads to activation of transcription in X. laevis (Ruzov et al., 2004). Similar studies in zebrafish and other organisms have proposed a model called the “excess repressor model” which suggests the dilution of these factors due to the increasing DNA and nucleocytoplasmic (N/C) ratio (Newport and Kirschner, 1982). Another model in Xenopus embryos suggests the removal of excess histones (repressors) through titration of chromatin modifying complexes by the transcription machinery such as the TATA-binding protein (Prioleau et al., 1994). Although the concept of transcriptional repressors keeping the zygotic genome inactive might be true it does not explain why some genes are still expressed during early cleavage and hence, are regulated in a gene specific manner.

1.2.3 Maternal transcript destabilization and their degradation

The proportion of maternal RNAs in an embryo varies in different animals ranging from 40% to 75% of the total protein-coding genes in these species (Lee et al., 2014). With the maternal-to-zygotic transition leading to expression of several zygotic RNAs there is simultaneous degradation of maternal RNA products as they might interfere with later development. A few maternal transcripts, however, are protected from degradation such as the *cey-1* gene in *C. elegans* (Seydoux and Fire, 1994), *nanos-1* gene in zebrafish (Koprunner et al., 2001) and *eIF5* gene in *D. melanogaster* (Lecuyer et al., 2007), in the germplasm to direct future cell fate decisions. In zebrafish, there are three different waves of maternal RNA degradation which take place prior to (2.5hpf), during (3hpf) or after MZT (4.5hpf) (illustrated in Fig 1.2.3) as suggested by a previous study (Mathavan et al., 2005). The clearance of maternal transcripts after zygotic genome activation is considered to be important during early embryonic development. One theory supporting this occurrence is the replacement of ubiquitously present maternal transcripts with these mRNAs (such as cdc5 phosphatase) being re-expressed from zygotic genome in a localised pattern in order to exert their influence (De Renzis et al., 2007). The maternal mRNAs based on their functional significance are categorised as unstable and stable transcripts, where the stable transcripts are mainly genes involved in translation or metabolism and the unstable transcripts have functions related to cell cycle (Tadros et al., 2007). Studies have proposed the existence of two mechanisms that ensure maternal transcript destabilisation and their subsequent degradation, maternal deadenylation of RNAs (destabilization) and zygotic activation of miRNAs (discussed in section 1.4.1) for degradation of maternal transcripts. Fertilization acts as a trigger for deadenylation of majority of the maternal transcripts via activation of an Embryonic Deadenylation Element Binding Protein (EDEN-BP) (Paillard et al., 1998) which brings deadenylation complexes to the AU rich cis elements of these RNAs (Duval et al., 1990). These deadenylated or destabilised maternal RNAs, however, requires zygotic genome activation to be eventually degraded. In *Drosophila* embryos, approximately 30% of the maternal RNAs become unstable but their expression levels remain the same as a third of these RNAs are zygotically transcribed (De Renzis et al., 2007). The second mechanism that ensures degradation of maternal RNAs, in zebrafish, is the activation of miRNA430 by Nanog, Pou5f3 and SoxB1 zygotic transcription factors (Lee et al., 2013). miRNA430 is the most highly transcribed early zygotic gene and is responsible for the clearance of majority of the maternal transcripts (Giraldez et al., 2006). The study found that the maternal-zygotic mutant of *dicer* (MZ*dicer*) embryos could not produce miRNA430 which lead to the stabilization of hundreds of maternal mRNAs that are otherwise degraded. miRNA430 is replaced by similar miRNAs in other model organisms such as miRNA-427 in *Xenopus laevis* (Lund et al., 2009) and *mir-309~6* in *Drosophila melanogaster* (Bushati et al., 2008).

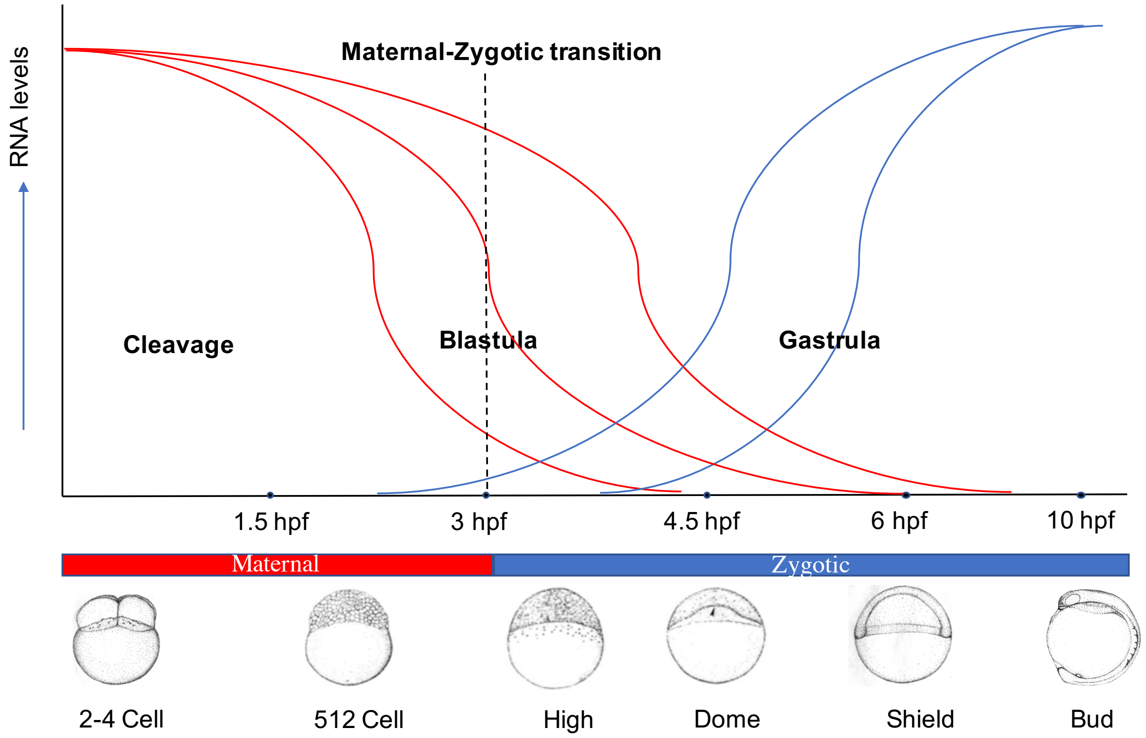


Figure 1.2.3 Different patterns of maternal RNA degradation in zebrafish. *The lines in red represent the maternal RNA levels while the blue lines represent the zygotic RNA levels. The X-axis explains the stages in development and on the Y-axis are the average RNA levels (Figure reproduced from (Mathavan et al., 2005)). The dotted line in the middle points to the zygotic genome activation.*

1.2.3.1 Zygotic Genome Activation

In all model organisms, there are two waves of zygotic genome activation (Fig 1.2.1), the minor wave which starts prior to mid-blastula transition (MBT) and the major wave, when majority of the zygotic genes have been activated (Mathavan et al., 2005). However, the timing of these major and minor waves is different in all these organisms, for example in mouse, both waves occur before the 2-cell cleavage stage (MZT). The minor wave just before the MBT is for the activation of genes, such as pluripotency-inducing transcription factors (Nanog, Pou5f3, Soxb2), responsible to equip the embryo for gastrulation and cell fate decisions. Mutations in these genes have been shown to have defects in late blastula stages with a complete failure to initiate the process of gastrulation in zebrafish embryos (Lee et al., 2013). In induced pluripotent stem cells (iPSC), Nanog, Pou5f3 and Soxb2 proteins bind to regions of repressed chromatin and bring chromatin remodelling complexes and other factors to these regions (Orkin and Hochedlinger, 2011). For zebrafish, a few zygotic transcripts are synthesized by 128 cell stage (cell cycle 7) and the rest of the zygotic genes are activated by 1000 cell stage (cell cycle 10) and are mostly both transcription factors and housekeeping genes (De Renzis et al., 2007). Majority of the zygotic RNAs are strictly zygotic and not present prior to ZGA, however, a few of them are initially deposited maternally but are re-expressed as part of the zygotic genome.

Zygotic genome activation is also important for the degradation of the maternal RNAs, by the zygotically transcribed miRNA430, and also for the embryo to complete gastrulation and enter epiboly. A previous study (Lee et al., 2013) found that blocking the splicing of zygotic RNAs (using morpholinos against the U1U2 components of splicing machinery) arrests them at gastrulation. Further, inhibition of translation by cycloheximide (CHX) in zebrafish embryos just before the start of MZT, ensuring translation inhibition of only zygotic mRNAs, prevented the embryo from entering epiboly.

1.3 Transcription and regulation of gene expression

Regulation of gene expression is necessary for early embryonic development. During early development, gene expression is very dynamic and is regulated in temporal as well spatial manner. Therefore, to fully understand the process of embryonic development, it is necessary to understand gene expression patterns and gene regulatory networks.

Transcription is a fundamental process that regulates gene expression levels. During transcription, genomic information is read by RNA polymerase II (Pol II) resulting in production of RNA molecules. The RNA molecules may or may not code for proteins. In a cell all the transcripts and their number at a specific developmental time or biological condition is defined as a transcriptome (Lee and Young, 2000). To better understand the changes to a transcriptome in a disease or during development several techniques have emerged in the past.

RNA-sequencing (RNA-seq) is one such technology that allows us in understanding the transcriptome as well as the changes in the abundance of transcripts in a cell. In the past decade RNA-seq has emerged as the preferred choice when it comes to cataloguing all transcripts, coding (mRNAs) and non-coding RNAs as well as their start (5'), end sites (3') and post-transcriptional processing or splice pattern. In general, all RNA-seq techniques include synthesis of libraries of cDNA fragments from a pool of RNAs (polyadenylated, polyA+ or non-polyadenylated, polyA- or both). This is followed by attachment of sequencing adapters either to the 5' end or both ends. The adapter ligated fragments are then PCR amplified and purified using gel or magnetic beads. These are then sequenced in short stretches called reads, typically 50-75 bp in length, from one end (single-ended) or from both ends (paired-end) (Wang et al., 2009).

1.3.1 RNA Polymerase II holoenzyme

The process of transcription is divided into three main parts, initiation, elongation and termination. During transcriptional initiation, Pol II binds upstream of a gene at the promoter sequence. The synthesis of RNA takes place during elongation phase and the synthesised RNA is released along with Pol II from DNA during termination phase (Kerppola and Kane, 1991). In humans, the transcription initiation complex consists of a 12-subunit (Rpb1-12) RNA polymerase II holoenzyme, general transcription factors (GTFs) and a coactivator or mediator complex (Buratowski, 1994). In yeast, the initiation apparatus, that is recruited onto the promoter of a gene, is made up of the Pol II enzyme, a few GTFs and the Srb/Mediator complex and is the best-defined RNA polymerase II holoenzyme (Fig 1.3.1) (Lee and Young, 2000).

RNA polymerase II CTD. RPB1 is the largest subunit of the Pol II holoenzyme which consists of a carboxyl terminal repeat domain (CTD) that is made up of repeats of a heptapeptide sequence (Y-S-P-T-S-P-S). The phosphorylation state of CTD is important for regulation of transcription. The unphosphorylated tail of Pol II is associated with the Srb/mediator complex, thus, favouring transcription initiation. While the heavily phosphorylated CTD of Pol II associates with elongation factors along with several RNA processing enzymes and thus favours the transition into the elongation phase (Corden and Patturajan, 1997, Corden, 1990).

General transcription factors. The binding of RNA Pol II to promoter is facilitated by TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. The preinitiation complex is formed by a stepwise binding of these factors at the promoter. First, the TFIID or TBP is bound by promoter elements, followed by TFIIA, TFIIB, then a complex of Pol II and TFIIF, TFIIE and in the end TFIIH (Conaway and Conaway, 1993).

Srb/Mediator complex. The Srb/Mediator complex is associated with the unphosphorylated CTD of Pol II. It is also believed to be responsible for the switch from transcription initiation to elongation by triggering the phosphorylation of the CTD via its interaction with the TFIIH GTF (Hengartner et al., 1998).

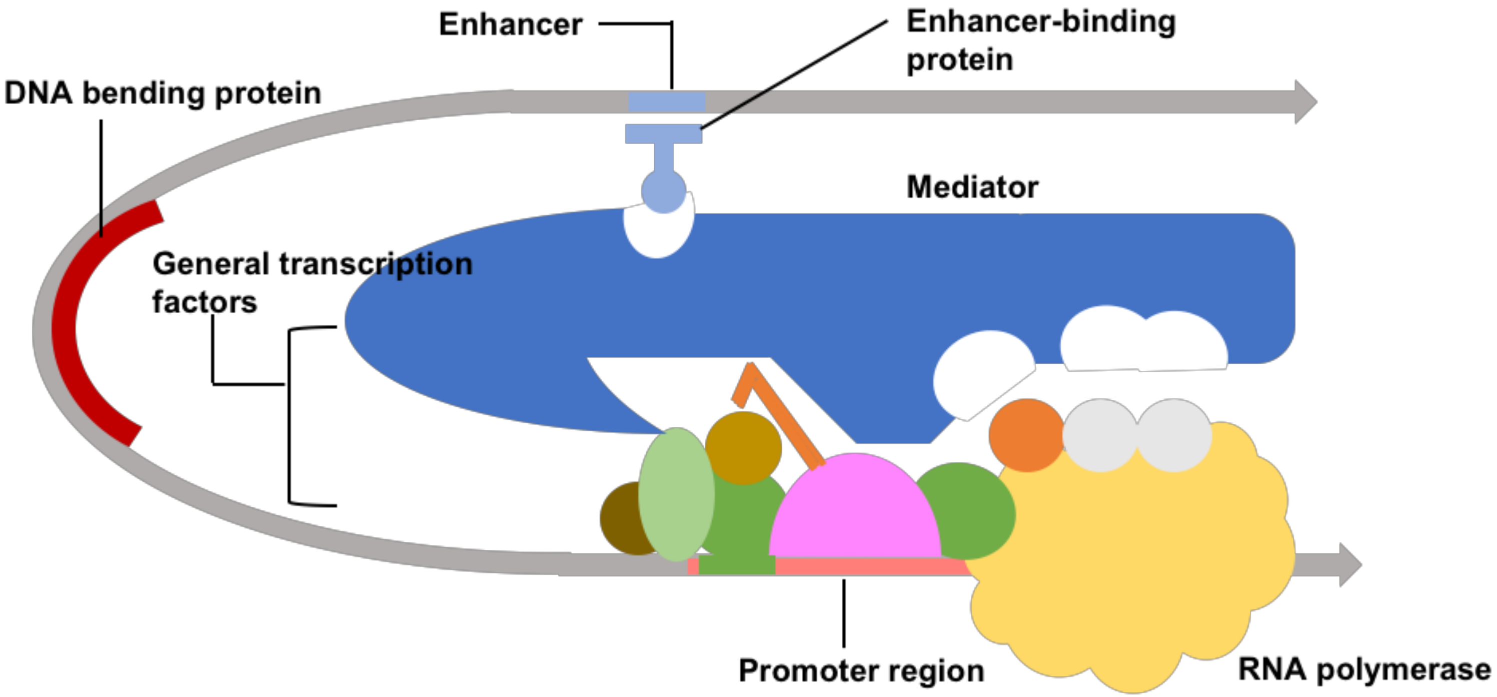


***Figure 1.3.1 Gene promoter organisation in Eukaryotes**.*
The figure above shows the arrangement of the general transcription factors (GTF's), meadiator complexes in blue (at enhancers) along with RNA Pol II (in yellow) at the promoter of a gene (in salmon colour). The arrows indicate the direction of transcription of gene. (The figure reproduced from web textbook ©CSLS/University of Tokyo).

1.3.2 Regulation of transcription

A typical eukaryotic protein-coding gene promoter consists of three basic elements, the transcription initiation site, RNA polymerase binding site and GTF binding sites (Blackwood and Kadonaga, 1998) (Fig 1.3.1).

Main sequence motifs of a eukaryotic promoter include a TATA box, an Initiator element (Inr), TFIIB Recognition Element (BRE) and Downstream Promoter Element (DPE). They together comprise around 100bp long region and are responsible for transcription initiation. The TATA box, rich in AT bases, is situated 25-30 bp upstream of the TSS (transcription start site) and provides the binding site for the TATA-binding protein (TBP). The composition of the promoters may vary with some TSS consisting of either an initiator element (Inr) or a TATA box or both and in some cases neither of them (Fig 1.3.2). Although they might not occur together in most of the core promoters. However, they both perform the same function of binding regulatory factors important for transcription initiation (Geng and Johnson, 1993). The more commonly occurring core promoter motifs, from Archaea to humans are the TATA box and BRE (Reeve, 2003). Other unconventional core promoter motifs include the DPE and motif ten element or MTE found to be conserved in both *Drosophila* and humans (Kadonaga, 2004, Lim et al., 2004), or the DCE (downstream core element) and XCPE1 (X core promoter element 1) that occur in selective human core promoters (Tokusumi et al., 2007, Lewis et al., 2000). In addition to proximal promoter elements there are distal regulatory elements which include the enhancers, the upstream activating sequences (UASs), the upstream repressing sequences (URSs) and the silencers. Transcriptional activators bind to UAS that promote the transcription of a gene both upstream and downstream. In contrast, URS elements represses the expression of a gene located downstream by preventing Pol II binding or interfering with the binding of activator complexes to the promoter (Cormack and Struhl, 1992), (Strubin and Struhl, 1992). Enhancers are able to activate transcription even when they are much further away by looping of the DNA. Thus, bringing the Pol II bound at the promoter in close proximity to activating factors or chromatin modifying complexes bound to the enhancers (Blackwood and Kadonaga, 1998). Much like enhancers, silencers inhibit transcription from very distal elements (Ogbourne and Antalis, 1998).

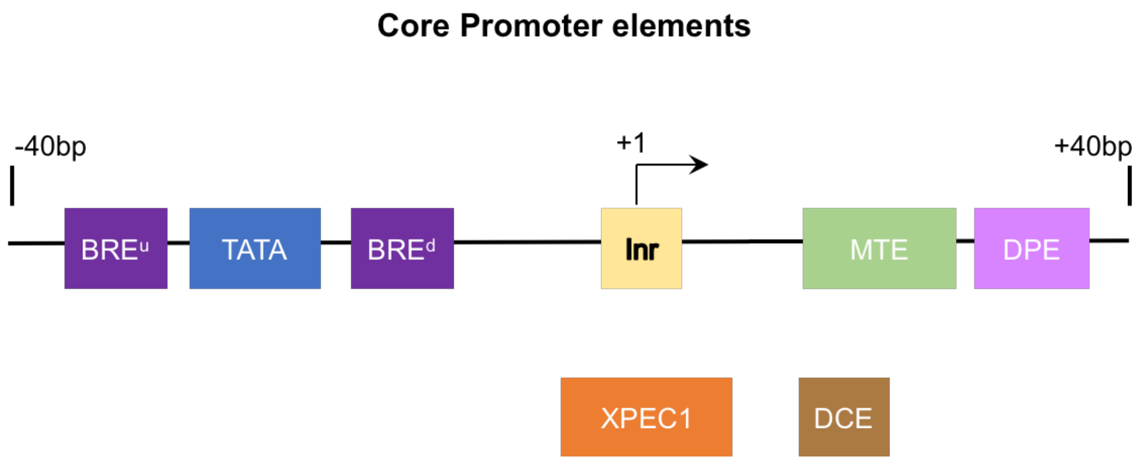


Figure 1.3.2 Core promoter elements.
The above schematics shows all the elements present in a core promoter. Although these do not occur together, some are only found in a fraction of promoters. The arrow indicates the direction of transcription and the motifs lie roughly 40bp upstream or downstream of TSS (+1). (Figure reproduced from (Juven-Gershon et al., 2008)).

1.3.3 Chromatin assembly & Nucleosome structure

Eukaryotic DNA is wound around histone protein complexes to form the nucleosomes. Chromatin is formed of repeating units of nucleosomes, which is composed of DNA wrapped around a core of 8 histones proteins connected to each other by a linker DNA and histone complex. The chromatin not only helps in the packaging of DNA into small nuclear space but also provides a mean of regulating DNA replication, repair and transcription (Fig 1.3.3). The nucleosomal core is formed of two molecules each of H3, H4, H2A and H2B histone variants while the linker histone is a H1 histone protein (Luger and Richmond, 1998, Luger et al., 1997). The DNA in this complex is held together by several interactions between its phosphate backbone and the histone protein. This highly dynamic interaction between the DNA and histone provides the basis for regulation of gene expression depending upon whether the DNA is accessible or not (Kornberg, 1974). This further leads to either a closed (heterochromatin) structure, when the DNA is not accessible, or an open chromatin conformation (euchromatin) associated with active post-translational modification mark. One mechanism is the exchange of histones with either newly synthesized ones or replacement with a different variant of histone (Venkatesh and Workman, 2015). Increase in the exchange of histones might, for example, increase the accessibility of DNA to the transcription machinery. On the other hand, replacement of a canonical histone protein with a variant might reduce the availability of DNA and thus, repress a gene. Several factors regulate this process of histone exchange during transcription and are termed as histone modifying complexes such as histone post-translational modification (PTMs), chromatin remodellers and histone chaperones (Belotserkovskaya and Berger, 1999, Cairns, 1998).

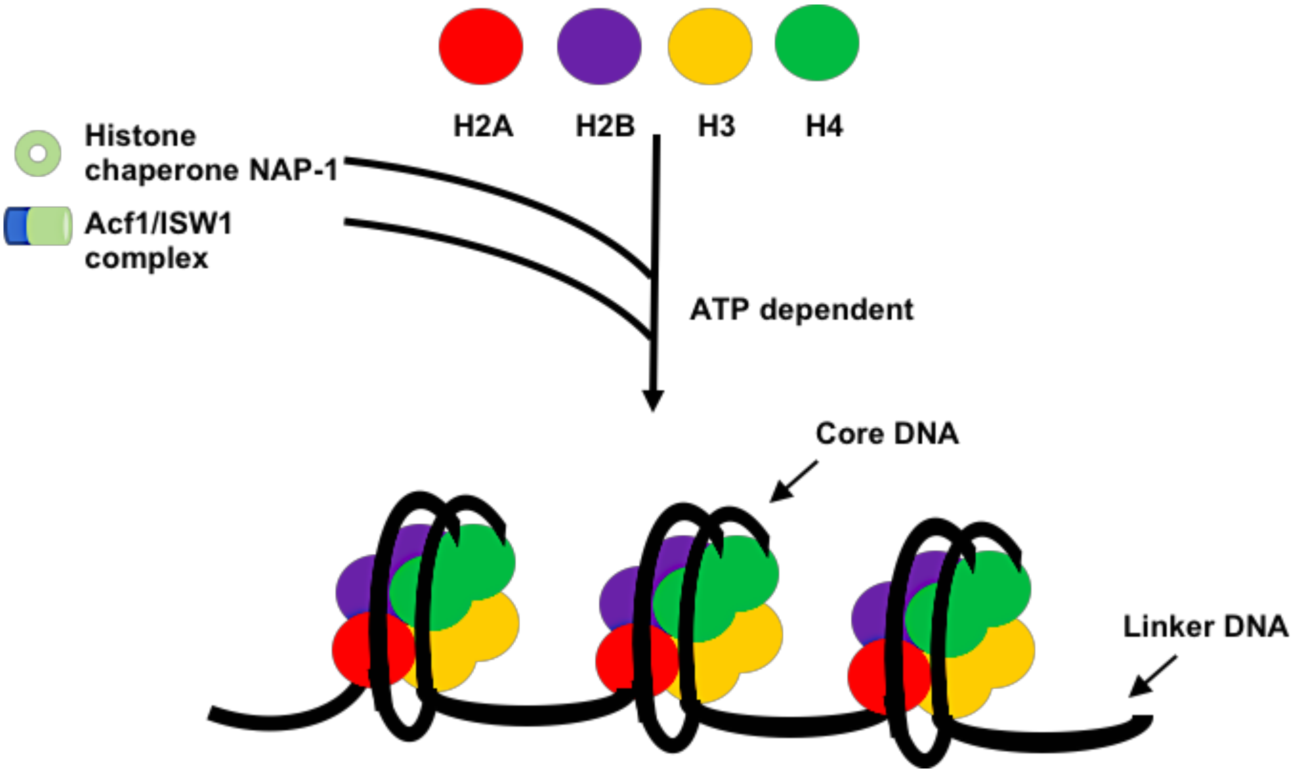


Figure 1.3.3 Nucleosome organisation.
The nucleosome is composed of 8 histone proteins, two molecules each of H2A (red), H2B (violet), H3 (yellow) and H4 (green) and are assembled with the help of histone chaperones and factors such as Acf1/ISW1/Swi. And DNA (in black) wrapped around the core of histone proteins (figure reproduced from DIAGENODE.COM).

1.3.4 Histone modifiers

Post-translational modification (PTM) of histones include methylation, ubiquitination, sumoylation, acetylation and phosphorylation of N-terminal ends of histones (Smolle and Workman, 2013). They are generally categorised as epigenetic changes as they affect the expression of a gene without changing the DNA sequence and are heritable. PTMs also result in a closed or an open chromatin structure, by either directly altering the interactions within a nucleosome or the interactions among adjacent nucleosomes. These modifications can also affect the process of transcription in their vicinity by acting as scaffold proteins, thus, bringing together several effector proteins (also known as ‘readers’). These modifications are reversible and enzymes for both addition (‘writers’) and removal (‘erasers’) of these marks exist in a cell, thus, controlling transcription in a time and space dependent manner (Venkatesh and Workman, 2015). Most modifications are associated primarily with histone H3 which can be acetylated at lysine 9, 14, 18, 23 and 27 (H3K9ac, H3K14ac, H3K18ac, H3K23ac and H3K27ac), methylated at lysine 4, 9, 27 and 36 (H3K4me3, H3K9me3, H3K27me3 and H3K36me3) and phosphorylated at serine 10 (H3S10P). The Fig 1.3.4 gives a schematic representation of a few histone modifications (that we will be further discussing in this thesis) and their relation to transcription (Bannister and Kouzarides, 2011). The H3K4me3 mark is associated with transcription initiation whereas the H3K27ac is a transcription activation mark (Spencer and Davie, 1999). The H3K36me3 is associated with gene body or exons during transcription elongation (Venkatesh et al., 2012). While the H3K27me3 mark is linked with repression of transcription (Cao et al., 2002).

Zygotic genome activation also leads to the establishment of histone 3 (H3) trimethylation (me3) marks at both 4th and 27th lysines (H3K4me3, H3K27me3) which are absent prior to MZT (Vastenhouw et al., 2010). The chromatin signature in zebrafish embryos is very similar to that of embryonic stem cells showing the presence of several genes (36%) marked with both H3K27me3 and H3K4me3 marks, indicating poised promoters. A lot of genes are also associated with only H3K4me3 (monovalent mark) and no active transcription believed to be required for recruiting transcription factors and RNAPol II on genome activation (Vastenhouw et al., 2010).

ChIP-sequencing or Chromatin ImmunoPrecipitation sequencing is a technique that allows us to study such modifications of chromatin and information about chromatin-binding proteins. It helps in identifying genome-wide presence of specific chromatin marks or binding sites for transcription factors (TFs) and other proteins involved in transcription regulation. The first step in ChIP-seq involves the crosslinking of DNA, within a cell, with bound proteins such as histones and TFs. DNA bound to the transcription machinery and other proteins is immunoprecipitated using an antibody specific for a histone mark or TFs. The bound DNA is separated and then purified. Before sequencing, the DNA is fragmented, end repaired, adapter ligated, PCR amplified and purified (Raha et al., 2010).

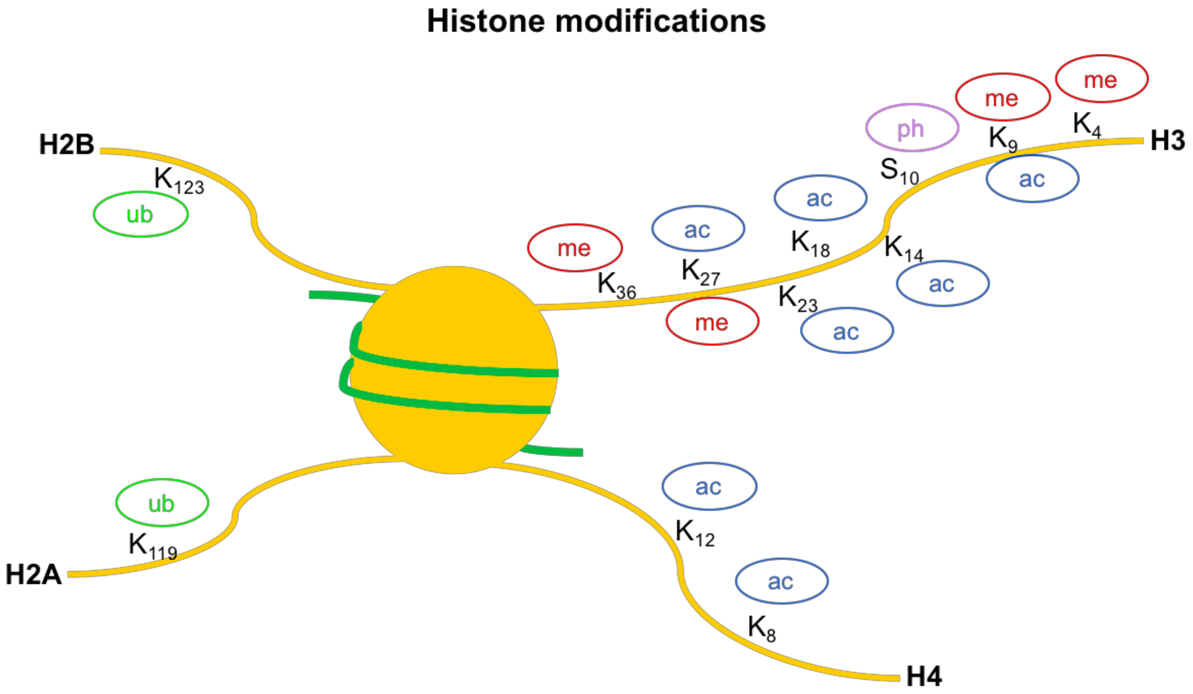


Figure 1.3.4 Regulation of transcription by histone PTMs.
The figure above shows the different types of histone modifications on H2A, H2B, H3 and H4. As the figure suggests majority of the PTMs are on H3. H2A and H2B only have ubiquitylation on K_{119} and K_{123}, respectively. The methylation marks are shown in red, acetylation in blue, phosphorylation in pink and ubiquitylation is in green. (Figure reproduced from (Bannister and Kouzarides, 2011)).

1.3.5 Chromatin remodellers.

Chromatin dynamics is also controlled by chromatin remodellers which influence nucleosome positioning and thus, promoting various states of active and inactive conformations (Kornberg and Lorch, 1999). As a result, they can have either positive or negative effect on transcription.

The chromatin remodelling is an energy intensive process which requires ATP (adenosine triphosphate) (Cairns, 1998). Different chromatin modellers have different specificity towards chromatin. For example, the SWI/SNF complex can be activated by both, nucleosomal or naked, DNA. It is recruited to the promoters of genes via its interaction with either the transcription machinery or DNA binding activator proteins (Belotserkovskaya and Berger, 1999). The SWR complex remodels chromatin by replacing the H2A histone with its variant, H2A.Z within a nucleosome. The IN080 complex, on the other hand, is involved in removal of the H2A.Z variant, preventing its mislocalisation (Watanabe and Peterson, 2010).

Histone chaperones. Another important class of proteins involved in histone dynamics during transcription are histone chaperones that interact with histones and regulate their transport and storage. Histone chaperones bind at specific regions on histones important for formation of nucleosomes, thus act by either stabilizing or destabilizing nucleosome structure. They also act as histone sinks or histone acceptors while associating with the chromatin remodellers (Kuryan et al., 2012). In addition to this, histone chaperones also associate with histone PTMs involved in transcription and regulate their function. For example, the histone chaperone Spt6 makes the conserved sites on nucleosomes accessible for Set2, a histone methyltransferase, to carry out the trimethylation of H3K36 (Suzuki et al., 2016). Histone chaperones thus promote histone PTM by making certain positions on nucleosomes accessible and therefore, assist in establishing activating as well as repressive marks.

1.3.6 Transcriptional Activators and Repressors

The co-existence of both positive (activators) and negative (repressors) regulators is essential for a controlled gene expression (Lee et al., 1998).

Activators have DNA binding motifs and for recruiting factors involved in regulating transcription. Multiple genes can be activated by the same activator and a single gene can be regulated by multiple activators (Ptashne and Gann, 1997). Multiple activators acting together even in small concentrations can activate transcription on a large scale (Carey, 1998). One good example of this are enhancers which act as scaffolds by bringing in several factors such as chromatin modifying complexes, transcription factors that together regulate transcription of a single gene.

Previous studies have shown that activators regulate transcription by recruiting chromatin-modifying complexes, such as Swi/Snf to gene promoters. For example, in yeast, transcriptional activator Swi5 recruits both Swi/Snf and the histone acetylase SAGA and then binds another co-activator, SBF, which in turn recruits the transcription initiation complex (Aasland et al., 1996).

Activators can also directly bind and recruit the RNA polymerase II complex. An activator either recruits the entire transcription initiation apparatus in one single step or recruits each transcription factor, in a step by step manner. Another theory suggests the recruitment of an initiation complex at the promoters that responds to specific regulators bound to enhancers (Greenblatt, 1997, Koleske et al., 1996).

Some activators regulate transcription by affecting the activity of RNA Pol II. Certain activators increase the processivity of Pol II while the others tend to stabilize the transcription apparatus after initiation, thus, promoting elongation process (Brown et al., 1998).

Repressors. General repressors usually regulate transcription by binding to the TATA binding protein or TBP at promoters (Hanna-Rose and Hansen, 1996, Hansen, 1996). For example, Mot1 via its interaction with the TBP prevents it from binding DNA. While Nc2 binds to TBP and inhibits the formation of the RNA Pol II holoenzyme and its assembly into an initiation complex (Auble et al., 1997).

Gene specific repressors function by either binding to an activator or competing with the activator for its binding site. Examples also include Ssn6-Tup1 that represses via interactions with the transcription machinery while bound to the DNA on specific promoters (Carlson, 1997). Hsp90, a heat shock repressor protein, binds to the activator Hsf1 thus preventing the formation of the Hsf1 trimer essential for binding of the heat shock DNA element.

Histone deacetylases also act as gene-specific repressors by modifying the chromatin around a gene. These deacetylases can be recruited to the site of inhibition either by other co-repressors (N-Cor, SMRT, Rb and Groucho) or by DNA binding proteins or in some cases by methylated DNA binding proteins (MeCP2) (Ayer, 1999).

1.3.7 Repetitive elements

Repetitive sequences or generic repeated signals in the DNA were previously termed as junk DNA or parasitic DNA (Orgel and Crick, 1980). Repetitive elements in humans make up 43% of the sequenced genome and also contribute to some part of heterochromatic part of the genome (International Human Genome Consortium, 2001). Repetitive elements can be of many types depending on their structural arrangement. VNTR (variable nucleotide tandem repeats), tandem array satellites, DNA transposons and LTR (long terminal repeat), SINE (short interspersed nucleotide element) and LINE (long interspersed nucleotide element) retrotransposons (Skelding et al., 2002). Tandem array satellites are large repeat elements, 100-200 bp, with thousands of copies of oligonucleotides that usually occur in the heterochromatic region of chromosomes such as telomeres and centromeres. They are known to have binding motifs for proteins involved in chromatin organisation (Henikoff et al., 2001). On the other hand, DNA transposons and retrotransposons are sequence specific mobile genetic elements, retroviral or non-viral in the latter case, that can incorporate within a genomic locus and can affect chromatin organisation, DNA replication and transcription. LINEs (or L1s) are the predominant class of retrotransposons in the genome and comprise 1.6% of human promoters. SINEs tend to be enriched in GC-rich DNA and comprises 5.3% of human promoters (Nigumann et al., 2002).

Transcription of genes is dependent on a lot of factors such as transcription initiation (promoter), stop sites (polyA addition signal) and the specific binding of transcription factors and other essential proteins to these motifs. Since there are canonical transcription start and end sites, cells can coordinate the expression of these RNAs and subsequently proteins. The occurrence of repetitive sequences in genome might be due to their importance in these DNA-protein interactions during DNA replication and transcription. The common repetitive signals found among these binding sites can, therefore, affect the chromatin states at multiple loci in a cell or tissue specific manner (Alberts and National, 2002). These common repetitive sequences found in the regulatory sites have motif specificity to DNA binding proteins. Differences or sequence alterations in one or two base pairs among similar repeat elements can, therefore, serve a regulatory role.

One of the best studied examples of LINEs are the LINE-1 in humans and the *gypsy* in *Drosophila melanogaster*. *gypsy* establishes chromatin domains and acts as an insulator or boundary element that separates the active and inactive parts of chromatin by binding the DNA to the nuclear matrix in *D. melanogaster* (Pelisson et al., 2002). The LINE-1 retrotransposon, on the other hand, is responsible for maintaining the heterochromatic regions in the inactive X chromosome in humans (Han et al., 2004). These mobile repetitive elements, when incorporated into any location in the genome, provides a cell with a tool to predetermine the physical organisation of chromatin (either closed or open).

1.4 Non-coding RNAs

The completion of the human genome project revealed that the portion of human genome encoding mRNAs is only 2%. However, it was found out later that the rest of the genome encodes thousands of non-protein-coding transcripts (Katayama et al., 2005). Recent advancement in deep sequencing technologies has revealed the presence of several non-coding transcripts which were previously considered to be transcriptional noises due to their low levels. The conservation of promoter region is comparable in both the protein-coding and non-coding genes indicating a conserved mode of regulation (Derrien et al., 2012, Derrien et al., 2011). Since ncRNAs interact with other RNAs and protein, conservation in ncRNAs could also be found at the level of how they function such as the formation of secondary hairpin structure. Non-coding RNAs can be categorised into housekeeping non-coding RNAs (expressed constitutively), such as rRNA (ribosomal RNA), tRNA (transfer RNA), snRNA (small nuclear RNA), which are highly conserved and secondly as regulatory RNAs (small ncRNAs and long ncRNAs). In this thesis, we will particularly be focussing on regulatory non-coding transcripts which have been classified into small ncRNAs (<200 nucleotide in length) such as miRNAs, siRNAs and piRNAs and long ncRNAs (>200 nucleotide in length) (Taft et al., 2010). Further details have been summarised in Table 1.4.1. Numerous studies have now shown these ncRNAs to be regulated during development, in a tissue or cell specific manner and are associated with several human diseases (Wang and Chang, 2011). The focus of this thesis is on long non-coding RNAs and understanding their importance in gene regulation during development (Fig 1.4.1).

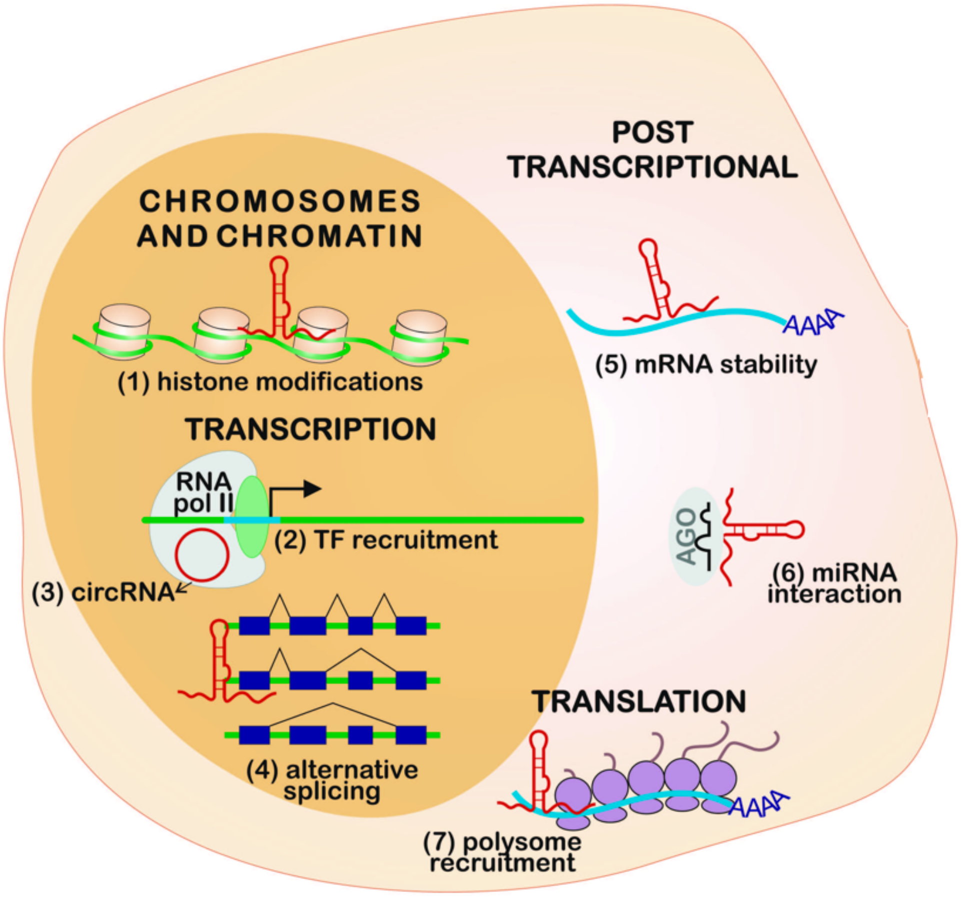


Figure 1.4.1 Non-coding RNAs in different subcellular compartments.

The schematics shows the function of ncRNAs in different mechanisms. In the nucleus, they are involved in histone modifications or changes in chromatin state (1), or regulation of transcription by TF recruitment (2) or binding to RNA Pol II (3) and in alternative splicing (4). In the cytosol, ncRNAs regulate post-transcriptional processes such as stabilising mRNAs (5), dsRNA degradation by miRNAs (6) recruitment of polysomes onto mRNAs (7). Figure taken from (Fernandes et al., 2019).

Table 1.4.1 List of types of non-coding RNAs

The table below outlines the different categories of non-coding RNAs based on the size and the function they carry out. Table adapted from (Taft et al., 2010).

Non-coding RNA class	Characteristic	Reference
Long non-coding RNAs (lncRNAs)	All non-coding RNAs >200nt	(Mercer et al., 2009, Wilusz et al., 2009)
Small interfering RNAs (siRNAs)	Small RNAs, less than 22nt, resulting from dsRNAs by Dicer	(Malone and Hannon, 2009b, Carthew and Sontheimer, 2009)
microRNAs (miRNAs)	Cleavage of RNA hairpins mediated by Dicer resulting in ~21-22 nt miRNAs	(Ghildiyal and Zamore, 2009, Winter et al., 2009)
PIWI-interacting RNAs (piRNAs)	Small RNAs (26-30 nt) produced by cleavage of lncRNAs	(Malone and Hannon, 2009a, Ghildiyal and Zamore, 2009)
Promoter-associated RNAs (PARs)	Long or short RNAs associated with promoters	(Belostotsky, 2009, Taft et al., 2009b)
Small nucleolar RNAs (snoRNAs)	Nucleolar localised RNAs that guide modification of rRNAs	(Taft et al., 2009a, Matera et al., 2007)

1.4.1 Small non-coding RNAs and their function

Small non-coding RNAs encompass RNAs ~21-22 nucleotide in length produced as a result of cleavage of double strand RNAs by Dicer, an RNaseIII endoribonuclease. The most commonly occurring small ncRNAs include siRNAs, miRNAs and piRNAs. The first evidence of siRNAs (small interfering RNA) gene silencing was found in the 1990s in *C.elegans* when a dsRNA was introduced and was found to be cleaved by Dicer into ~21 nt RNAs (Fire et al., 1998). Since then studies in plants and animals have identified many endogenous siRNAs which include miRNAs and PIWI-interacting RNAs (piRNAs) (shown in Fig 1.4.2). Other examples include small RNAs associated with the transcription start site (TSSa) RNAs such as the promoter-associated RNAs (PARs) and the transcription initiation RNAs (tiRNAs) (Malone and Hannon, 2009b, Kanhere et al., 2010). miRNAs, piRNAs and siRNAs so far have been the most well studied small ncRNAs. These small ncRNAs have been shown to be important in transposon silencing, chromatin remodelling and silencing of gene expression by Argonaute-mediated degradation of sequence specific RNAs (Hutvagner and Simard, 2008). piRNAs are slightly longer, ~26-30 nt, than the siRNAs and are produced by the cleavage of long ncRNAs either from host genome or transposons by the Argonaute proteins (Taft et al., 2010). miRNAs, on the other hand, are small RNAs, ~21 to 22-nt in length, produced by Dicer cleavage of RNA hairpins and integrated into RNA silencing complexes (RISC) to target sequence specific RNA for degradation (Huntzinger and Izaurralde, 2011).

Both piRNAs and miRNAs are shown to play important role during early development. For example, one of the well-known miRNAs involved in development is *let-7* which is responsible for the degradation of cell-cycle regulators such as *cdk-6, Ras* and several pluripotency factors such as LIN28 and *c-Myc* (Johnson et al., 2005). The *let-7* pre-miRNA is made by the drosha RNase in the nucleus from an RNA polymerase II transcript which is then cleaved into a ~22 nt miRNA by Dicer in the cytoplasm. This process is highly regulated at each step through a feedback loop. The pluripotency factor LIN28 binds to the *let-7* pre-miRNA and prevents its cleavage by the Drosha or Dicer endonuclease. In the process LIN28 also recruits ZCCHC1, a poly(A) polymerase TUTase4, which adds oligo-uridine tail to *let-7* pre-miRNA, thus, facilitating the degradation of *let-7* (Rybak et al., 2008, Abbott et al., 2005, Pauli et al., 2012). Another miRNA, miRNA430, is essential for zygotic genome activation and hence, development in zebrafish (Winter et al., 2009).

SnoRNAs or small nucleolar RNAs, are non-coding RNAs that can be categorised into two families, C/D RNAs and H/ACA RNAs, that are involved in the processing of ribosomal RNAs (rRNAs), mRNA splicing and guide ncRNA modification (Matera et al., 2007). The C/D RNAs methylate the hydroxyl group in the precursor and the H/ACA RNAs convert uridines into pseudouridines (pseudouridylation) in several rRNAs and snRNAs (Kiss, 2002). The two families of snoRNAs form small nucleolar ribonucleoprotein (snoRNPs) complexes with four core proteins where the proteins are needed for the catalytic activity and the snoRNA provides specificity to guide the complex to the target rRNA. The snoRNP complex, apart from being essential for rRNA function, is also responsible for the modification of snRNAs associated with the spliceosome (Jady and Kiss, 2001). These snoRNAs may express from a single transcription unit or are processed individually from multiple snoRNAs in a precursor (fungi and metazoans) and some can also be produced from the introns of mRNAs (as in humans). The biogenesis of the two ncRNAs, snoRNA and snRNA, are interdependent on each other with the snRNAs processing the snoRNA from the introns of mRNAs through splicing and the snoRNA, in turn, catalysing the modification of snRNAs (Roy and Chanfreau, 2012).

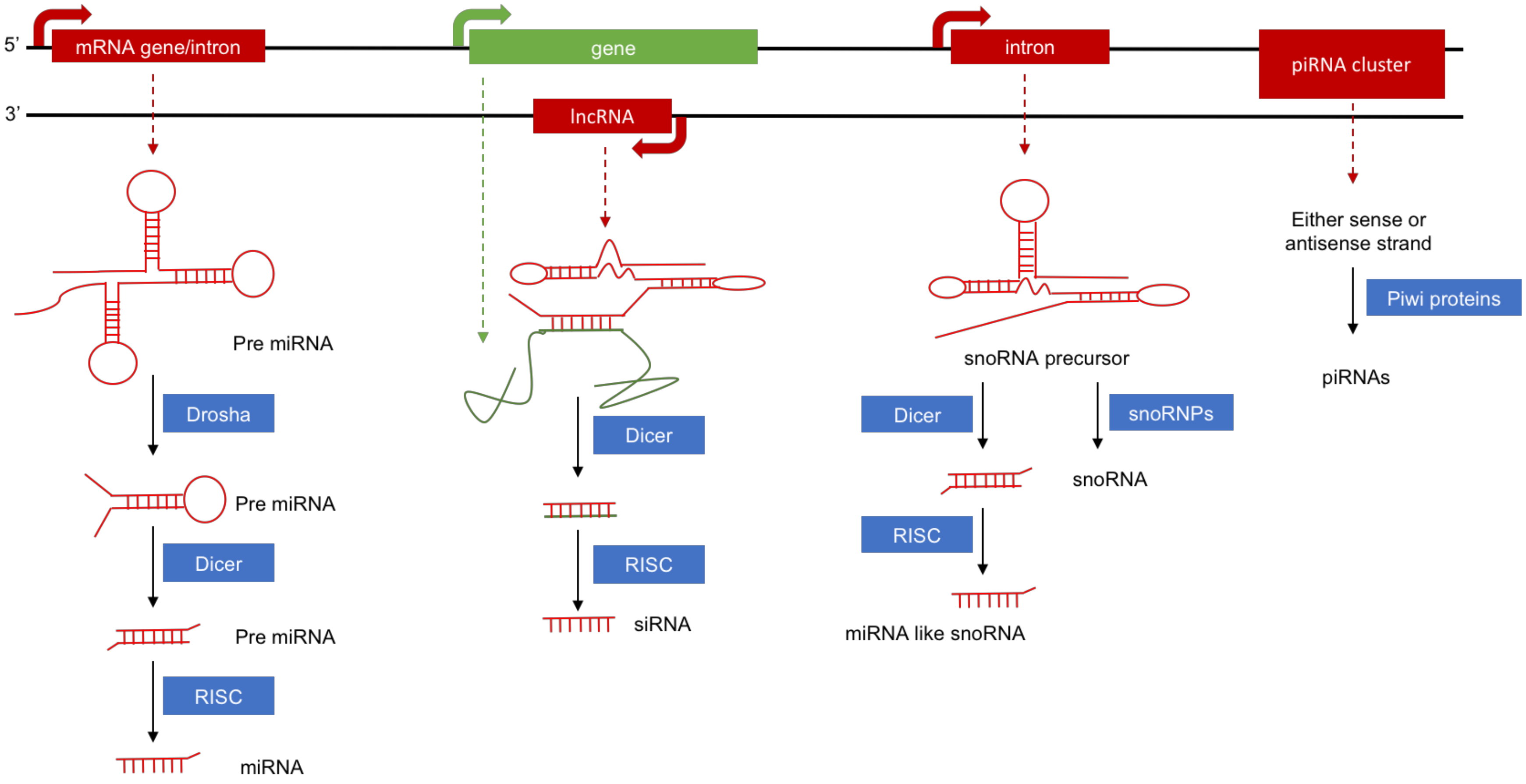


Figure 1.4.2 small ncRNAs biogenesis pathways.

The figure shows the synthesis of various types of small ncRNAs from either mRNA's (miRNA) or from a lncRNA overlapping an mRNA (green) on the opposite strand (siRNA). Some small ncRNAs are also synthesised from introns (snoRNA) or from piRNA cluster (piRNAs). Figure reproduced from (Sana et al., 2012).

1.4.2 Long non-coding RNAs (lncRNAs)

Long non-coding RNAs are a broad class that includes enhancer RNAs (eRNAs), long intergenic RNAs (lincRNAs) and non-coding transcripts overlapping a protein coding gene either in sense or antisense direction. Although lincRNAs do not encode for proteins (Guttman et al., 2013), they show many properties similar to protein-coding genes. Characterisation of chromatin-modification landscapes in mammalian genome revealed the association of H3K4 trimethylation, associated with promoters of protein coding genes, and H3K36 trimethylation, associated with entire transcribed region, with many long ncRNA genes (Guttman and Rinn, 2012). Similar to protein-coding genes, many lincRNAs are also spliced.

LncRNAs can regulate expression of its neighbouring genes *in cis* as well as genes at a distant locus in *trans*. The mechanisms by which lncRNAs regulate the function of protein coding gene expression is still poorly understood. A study in human cell lines (Andersson et al., 2014) identified ~3000 enhancer RNAs that positively regulate the expression of neighbouring protein-coding genes and display enhancer-like properties such as Pol II, p300 and CBP occupancy. LncRNAs can be grouped together into different types based on their functional mechanisms (Wang and Chang, 2011) as illustrated in Fig 1.4.3. LncRNAs can act as signals and express in a time and space dependent manner (Mohammad et al., 2009). LncRNAs may also act as negative regulators by titrating the effects of RNA binding proteins, such as chromatin modifiers or transcription factors (Martianov et al., 2007). On the other hand, they can help in recruiting chromatin-modifying complexes to nearby protein-coding genes either in *cis* or in *trans* (Leeb et al., 2009). They can also act as scaffolds thus bringing together different effector molecules involved in regulation (Gupta et al., 2010). In some cases just the act of lncRNA transcription could mediate regulatory function by recruitment of these complexes (Ebisuya et al., 2008).

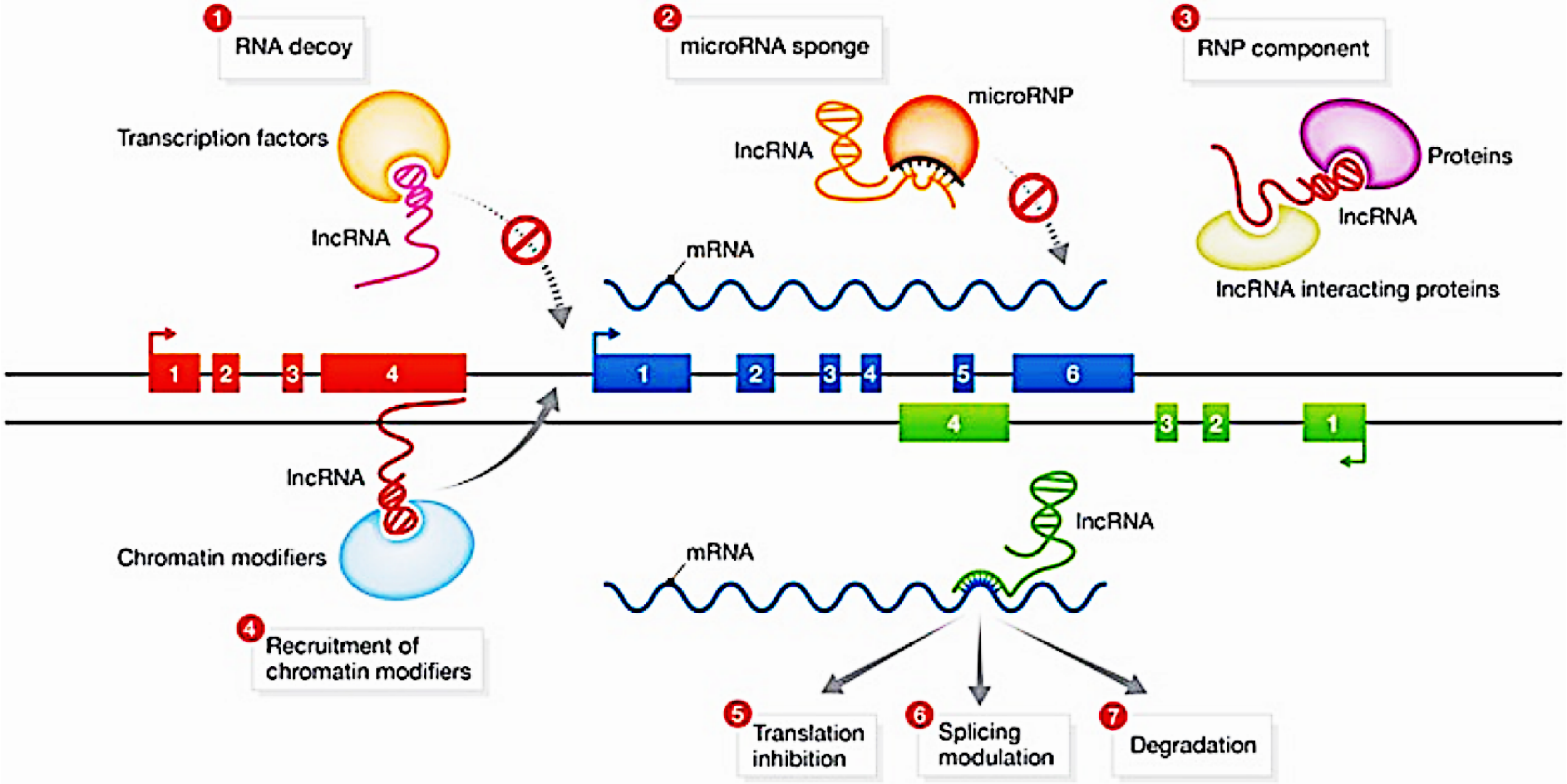


Figure 1.4.3 Representation of molecular mechanisms of LncRNAs.

The figure shows how lncRNAs regulate the expression of nearby genes either in cis or trans. lncRNAs act as signals whereby the expression of lncRNAs shows gene regulation in a space and time dependent manner, or they can act as decoys recruiting transcription factors either to or away from a target gene, or as guides where they engage chromatin modifying complexes or miRNAs to a target gene and form ribonucleoprotein (RNP) complexes or scaffolds bringing together factors necessary for gene expression. Figure taken from (Hu et al., 2012).

1.4.2.1 LncRNAs in development

There have been several studies in the past that have shown the importance of lncRNAs in regulation of dosage compensation and genomic imprinting, which are essential for normal development and ensures monoallelic expression of specific genes (reviewed in (Fatica and Bozzoni, 2014)). In mammals, both males and females have 22 pairs of chromosomes called autosomes and a pair of sex chromosomes, XX in females and XY in males. The X chromosome is large and gene rich as opposed to Y-chromosome, which is small and gene poor, thus, leading to chromosomal imbalance in the two sexes. To prevent this, a mechanism called dosage compensation exists in animals that maintains a balance in X-linked gene products between the two sexes (Brockdorff and Turner, 2015). There are three mechanisms of dosage compensation: complete inactivation of one of the X-chromosomes in females (mammals); secondly, the X chromosome in males is twofold upregulated (*Drosophila melanogaster*) and thirdly, partial inactivation of both the X chromosomes in females (*Caenorhabditis elegans*). A number of lncRNAs are shown to play a role in X-chromosomal inactivation. The best example of this is *Xist*, a lncRNA expressed from X-chromosome that regulates X chromosome inactivation in mammals by formation of repressive chromatin state and is a well characterized example of dosage compensation in mammals (Fig 1.4.4a-b) (Zhao et al., 2008). The X chromosome inactivation is carried out by the X-inactivation centre (Xic) which consists of seven regulatory ncRNAs, one of which is the lncRNA, *Xist*. *Xist* has an antisense, *Tsix*, which ensures silencing of *Xist* on the active X-chromosome (Xa) (Sun et al., 2006). While the *Jpx* lncRNA, on the other hand, upregulates the expression of *Xist* on the inactive X-chromosome (Xi) (Tian et al., 2010).

Kcnq1 overlapping transcript 1 (*Kcnq1ot1*) and *Airn* (antisense to Igf2r RNA) are paternally expressed lncRNAs that repress flanking protein coding genes, *kcnq1* and *Igf2r* allele, respectively in *cis*, which are essential in early development in mammals (Fatica and Bozzoni, 2014). *Kcnq1ot1* recruits the DNA methyltransferase DNMT1 on to *Kcnq1* gene thereby silencing the gene (Pandey et al., 2008) whereas the antisense *Airn* recruits G9a methyltransferase that induces deposition of H3K9me mark at nearby protein-coding gene, *Igf2r* (Nagano et al., 2008).

Another example of developmentally important lncRNAs is that of lncRNAs expressed from the HOX gene cluster. HOX genes encode transcription factors that contributes to cell specification and precise spatiotemporal coordination of expression during development. Apart from several protein coding genes, the *Hox* cluster also encodes several lncRNAs that regulate the expression of neighbouring protein coding genes. *HOTTIP*, *Hoxa* distal transcript, is a cis acting antisense lncRNA that regulates the expression of HOXA by recruiting the histone-modifying complex, MLL1 (Wang and Chang, 2011, Wang et al., 2011). *HOTTIP* regulates other genes by utilizing pre-existing chromosomal loops to move from its site of expression. The *Hoxc* gene cluster on chromosome 12 has an overlapping antisense called *HOTAIR* or *Hox* transcript antisense RNA that inhibits, in trans, the expression of *Hoxd* cluster present on chromosome 2 (Fig 1.4.4 c) (Maamar et al., 2013).

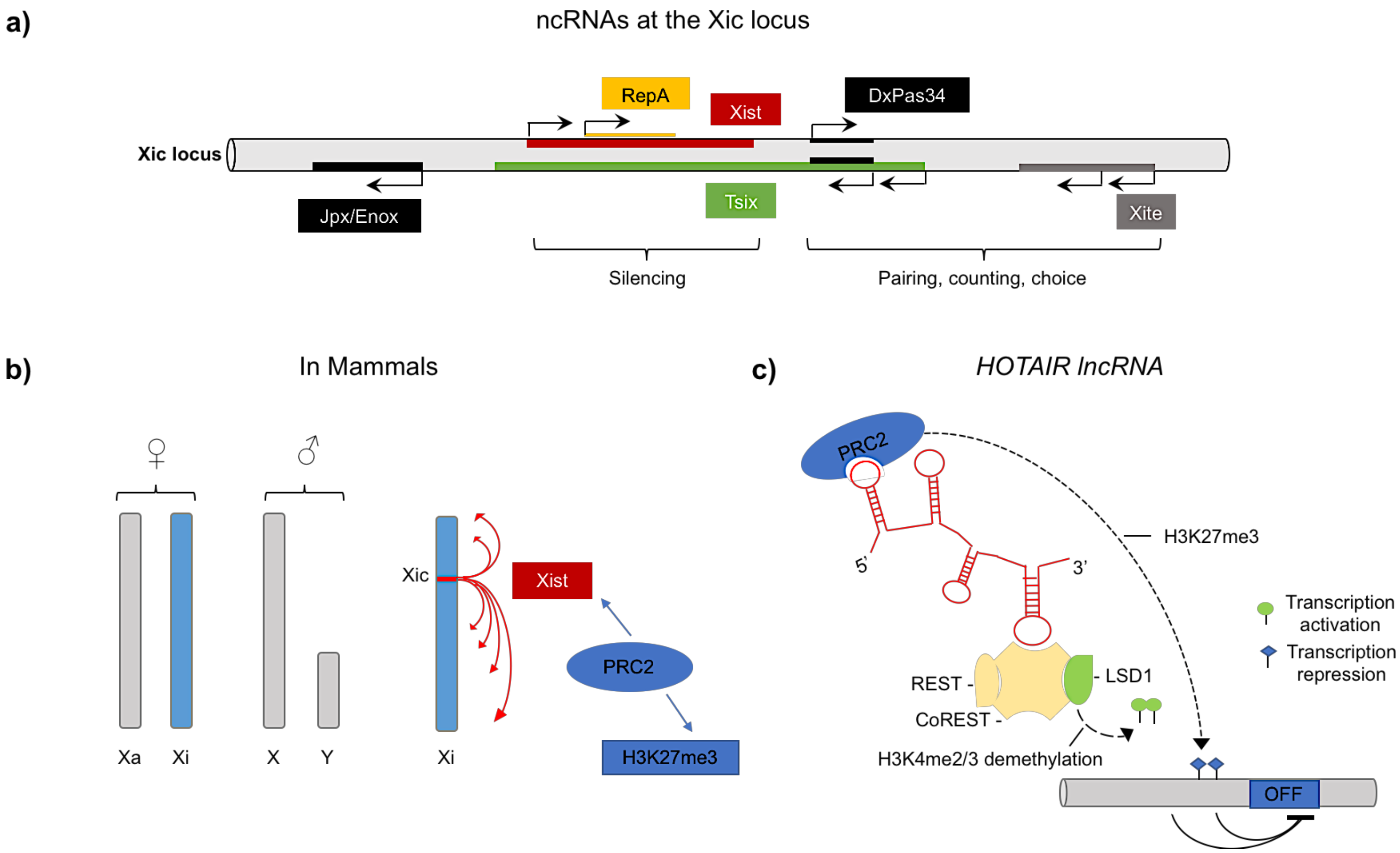


Figure 1.4.4 LNCRNAs in development
a) *the schematics show the arrangement of ncRNAs on the Xic locus (X-chromosome).* ***b)*** *in mammals X inactivation is carried out by Xist (red) which spreads from its site of transcription along the X chromosome. Xist associates with PRC2 (recruited by RepA) to establish H3K27me3 repressive marks on Xi (blue).* ***c)*** *The figure shows the mechanism of regulation by HOTAIR. It establishes repressive chromatin marks at the target gene by binding to PRC2 complex and the H3K4 demethylating lysine-specific demethylase1 (LSD1)-CoREST-REST complex. Figure reproduced from (Pauli et al., 2011).*

1.4.2.2 Antisense LncRNAs

Long non-coding RNAs are often transcribed as antisense transcripts that overlap protein-coding genes on the opposite strand (Duttke et al., 2015, Scruggs et al., 2015, Xu et al., 2009). Overlapping transcripts could involve a protein coding RNA overlapping another protein coding or non-coding RNA but natural antisense transcripts is a term restricted to a non-coding transcript. Due to their low level of expression and low evolutionary conservation, antisense RNAs (AS RNAs) were previously considered to be transcriptional noise. A study (Kiyosawa et al., 2005) in mice found ~15% of the genes to be transcribed as Sense-Antisense pair. The FANTOM consortium in 2005 reported that up to 72% of potential transcriptional units are transcribed in both directions in mice.

Antisense transcription is mostly observed near the 5’ end of protein-coding genes with the lncRNAs located mostly 1000bp downstream of the first exon intron junction. These antisense transcripts could be classified as *cis*-antisense, when it overlaps the sense transcript, and *trans*-antisense, when it is transcribed from another genomic location but still shows sequence overlap with the sense transcript. They can be further categorised, as shown in Fig 1.4.5, according to their genomic position with regards to the protein coding gene (Villegas and Zaphiropoulos, 2015). The antisense can occur nearby the 5’ end of protein-coding transcript; or nearby the 3’ end of protein-coding transcript. The categories include where the 3’ end of protein-coding transcript is localised close to the 3’ end of antisense (End-to-End); or 5’ end of protein-coding transcript intersects the 5’ end of antisense (start-to-start or divergent transcription) or when the protein-coding transcript completely overlaps the antisense and vice-verse (overlapping or convergent transcription).

The expression of antisense transcripts could be either positively or negatively correlated with the expression of the protein-coding transcript and thus, can be involved in the regulation of expression of the protein-coding transcript. There are several mechanisms by which antisense transcripts can regulate the expression of protein-coding transcripts. They have been previously shown to act before transcription through epigenetic modifications like DNA and histone methylation and acting as guides for several regulatory proteins, or post-transcriptionally through RNA-RNA interactions and transcriptionally as regulators (Villegas and Zaphiropoulos, 2015). A study (Mayer et al., 2015) on human cell lines, revealed convergent transcription is characterised by a lower expression of gene. They also showed the presence of repressive histone marks to be associated with these types of antisense. A more recent study (Lavender et al., 2016) on convergent transcription suggested their association with p300 and H3K27ac marks, which is a common feature of enhancers. They also looked at divergent transcription and found them to be associated with chromatin remodelling complexes, therefore, regulating nucleosome positioning at promoters.

Antisense lncRNA important in the regulation of the overlapping protein-coding gene have also been studied in zebrafish, for example, the antisense to the gene *tie-1*, which is a tyrosine kinase receptor for angiopoietin and is essential for vascular development during embryogenesis (Li et al., 2010). The *tie-1AS* is transcribed in the antisense direction from the 3'-UTR regions of *tie-1* gene. The *tie-1AS* down regulates the abundance of *tie-1* gene by RNA-RNA hybridization, thus, resulting in protein loss. In another example, *PU.1* AS regulates the expression of *PU.1* gene which is a transcription factor in myeloid and lymphoid cell development, thus, making it important in immune system development (Wei et al., 2014). *PU.1* AS also down regulates the expression of *PU.1* gene by forming RNA-RNA duplex and preventing its translation.

Most of the antisense lncRNAs discovered so far have been characterised from cultured cells or adult tissues and are highly cell type-specific transcripts. Therefore, it is important to study those non-coding RNAs which are expressed during specific developmental stages. Pauli et al. for the first time systematically identified lncRNAs expressed during zebrafish development. They identified 1133 embryonic transcripts out which 566 were antisense exonic overlapping lncRNAs revealing that most of these non-coding RNAs were actually antisense non-coding RNAs (Pauli et al., 2012).

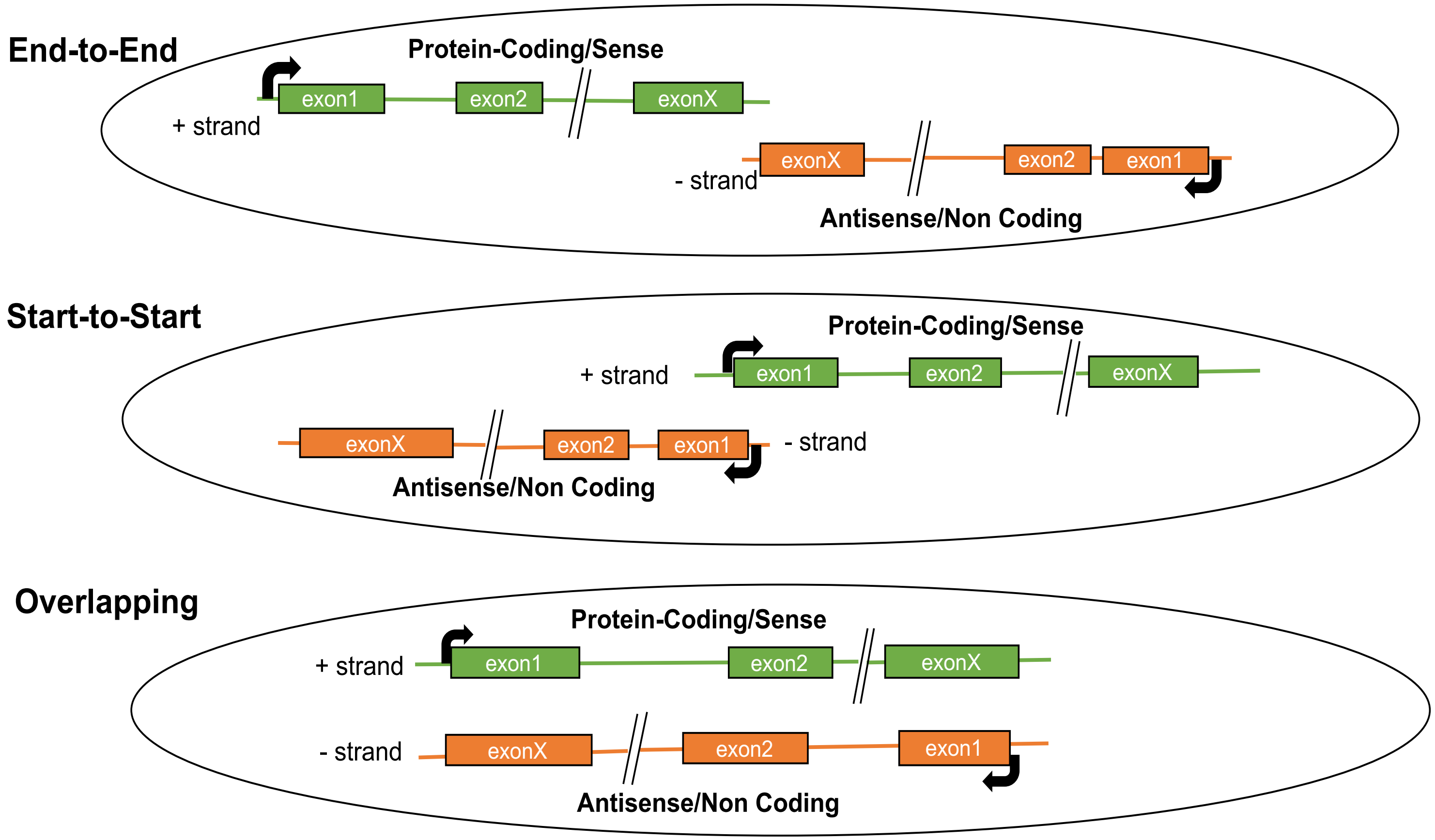


Figure 1.4.5 Types of antisense.
The figure shows the different types of antisense non-coding RNAs (AS) based on the localisation of the AS with regards to the protein-coding (Sense) gene (figure adapted from (Villegas and Zaphiropoulos, 2015)). End-to-End (convergent), where the tail of antisense overlaps with the tail of protein-coding, or Start-to-Start (divergent), where the antisense transcripts start overlaps with the start of the protein-coding transcript and lastly, the overlapping transcripts (another example of convergent), where either the antisense completely overlaps the protein-coding or vice-versa.

1.5 Aims & objectives

In zebrafish, previous RNA-seq have identified several long non-coding RNAs expressed or present during embryonic development (Ulitsky et al., 2011, Pauli et al., 2012). Even though we have a few pre-existing studies in zebrafish on non-coding RNAs, it is still largely unclear as to their role, if any, in early development. The Pauli et al. study identified around 1133 lncRNAs across eight stages of development, out of these 397 were intergenic lncRNAs, 184 intronic overlapping lncRNAs and the rest (566) were classified as exonic overlapping antisense lncRNAs (Pauli et al., 2012). The discovery of these abundant antisense lncRNAs opened questions about their role in early embryonic development which made the basis of our present study.

1. **Identification and characterisation of antisense ncRNAs and their overlapping protein-coding gene pairs during early embryonic development in zebrafish.** We aimed at identifying overlapping transcripts in zebrafish during development and then characterised them based on the correlation between the expression of antisense ncRNAs and overlapping mRNA transcripts based on the Pauli et al. RNA-seq (Pauli et al., 2012). To verify the observations from this study we further detected similarly overlapping transcripts in another vertebrate model organism (mouse embryonic stem cells).
2. **Non-coding RNA dynamics. Localisation of major RNAs with maternal-to-zygotic transition during zebrafish embryonic development.** The next important question that we wanted to address in this thesis was the subcellular localisation of the antisense ncRNAs identified with respect to their overlapping mRNA before and after zygotic genome activation. This

would help us in understanding how they might be regulating the expression of these mRNAs. We did a total RNA-seq on the nuclear and cytosolic fractions from zebrafish embryos in different stages of development.

3. **Identification of differential TSS usage and promoter organisation in the nuclear and cytosolic fraction during zebrafish embryogenesis.** To further understand the differences in the TSS or promoter usage for the nuclear and cytosolic RNAs identified during maternal-to-zygotic transition. We carried out Cap Analysis of Gene Expression (CAGE) sequencing to detect these changes in promoter utilisation during zebrafish embryonic development.

CHAPTER 2: MATERIALS AND METHODS

2.1 Collection and dechorionation of zebrafish embryos

Wild-type male and female zebrafish (AB-strain) were set up in breeding tanks overnight and the eggs were collected the next day as soon as (~10 min) they were laid (important to get synchronised embryos). About 100-500 (depending upon the stage) embryos were collected for each developmental stage (32 cells, 64 cells, 256 cells, 512 cells, high, shield, 24hpf, 50hpf). The embryos were treated with Pronase (protease from *Streptomyces griseus,* Sigma-Aldrich) to remove the chorion, 1mL of Pronase working solution (1mg/mL) was added to 2-3 ml of fish water containing embryos. The embryos were then washed twice to remove any traces of Pronase as that would interfere with other downstream processing.

2.2 Gene expression Analysis

2.2.1 RNA extraction from zebrafish embryos

Total RNA was extracted using the Nucleospin RNA-extraction kit (Macherey Nagel). Depending upon the stage ~ 50-1000 dechorionated embryos were transferred into a 1.5 mL eppendorf containing 350 µL of RA1 (with 3.5 µL β-mercaptoethanol, Sigma-Aldrich) buffer. The embryos were pipetted 10-15 times to dissolve in the buffer. After dissolving the embryos, one volume (350 µL) of 70% ethanol was added to the mix and pipetted at least 10 times. Approximately 700 µL of mix was transferred onto the Nucleospin column and centrifuged at 11,00 x g for 30 sec. The flow through was discarded and 350 µL of MDB buffer was added to the column. The column was then centrifuged at 11,00 x g for 30 sec and the flow through discarded. The column was then washed with 200 µL of RAW2 buffer, centrifuged as before, the flow through discarded and the column was placed in a new collection tube. The column was again washed with 600 µL of RA3 buffer, centrifuged as before and the flow through discarded. Finally, 250 µL of RA3 buffer was added to the column, centrifuged and flow through discarded. The column was given a blank spin to remove any traces of wash buffer. The RNA was eluted by addition of 30-40 µL of RNase free water to the column, centrifuged and the flow through collected in a 1.5 mL eppendorf.

2.2.2 Genomic DNase treatment

Removal of Genomic DNA was made sure by treating the RNA extracted using DNase treatment (Sigma-Aldrich, AMPD1-1KT). To an RNase free 1.5 mL eppendorf 8 µL of RNA extracted, 1 µL of 10X reaction buffer and 1 µL of DNaseI

(1 unit/µL) was added. The mix was incubated at RT (room temperature) for 15 mins. After the incubation 1 µL of STOP solution (50mM EDTA) was added to the mix and heated at 70°C for 10 mins to denature the DNaseI. After the incubation the eppendorf was chilled on ice. The quality of RNA extracted from zebrafish at different developmental stages were checked by running them on a gel.

2.2.3 cDNA synthesis from the extracted RNA

First strand cDNA synthesis was carried out using Bioline Tetro cDNA synthesis kit (BIO-65042) each for the different stages. In a 1.5 mL eppendorf upto 5 µg of DNase treated RNA, 1 µL of Random hexamer primers, 1 µL of 10mM dNTP mix, 1 µL of RNase inhibitor, 1 µL of Tetro Reverse transcriptase and 4 µL of 5 x RT buffer was added and made up to a final volume of 20 µL by addition of DEPC-treated water. The samples were mixed gently and incubated at RT for 10 mins. After the incubation the tubes were transferred to a heat block at 45°C for 30 mins before finally being incubated at 85°C for 5 mins and then chilled on ice. The quality of cDNA was tested using primers for genomic DNA and actin.

2.2.4 Expression analysis using PCR

PCR was carried using Bioline MyTaq kit (BIO-25043) as per the manufacturer's protocol. The reaction mixture was prepared using 50 ng of cDNA, 25 µL of MyTaq Red Mix, 10 µM of forward and reverse primers (Sigma-Aldrich) and RNase free water to make up the final volume to 50 µL. The mix was then PCR amplified using the conditions outlined in Table 2.2.1. After the run the PCR amplified samples were run on a 1% agarose gel. We shortlisted 3 protein-coding and antisense pairs, based on RNA-Sequencing analysis, from the two categories and checked the RNA levels using PCR. Primers were designed using PrimerBlast, in order, to specifically target only the protein-coding/protein-coding transcript (across exon-exon junction) or the non-coding antisense on the opposite strand (Fig 2.2.1). Table 2.2.2 gives a list of primers used and their sequence.

Table 2.2.1 PCR cycling conditions used for amplification.

Step	Temperature	Duration	Cycles
Initial Denaturation	95°C	1 min	1
Denaturation	95°C	15 sec	35
Annealing	60°C	20 sec	
Extension	72°C	10 sec	
Final Extension	72°C	1 min	1
Hold	4°C	∞	∞

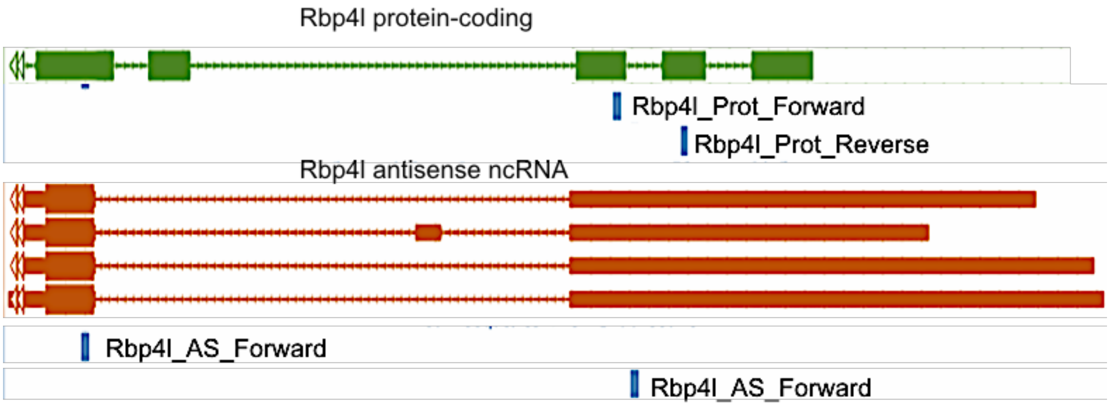


Figure 2.2.1 PCR Primer design
Example of a designed primers for PCR for Rbp4l protein-coding and antisense gene pair. Protein-coding transcript and non-coding antisense transcript are shown in green and red, respectively, while the designed primers are in blue.

Table 2.2.2 List of primer sequences used for expression analysis
Used to study the expression of the protein-coding transcript and the corresponding AS non-coding (Tm=melting temperature).

Primer Pair	Sequence	T_m	Size (bp)
Rbms3_AS_Forward	CGGAGCCAGAGGAATGATGG	63.8	533
Rbms3_AS_Reverse	GCAACAAAGCAAGCAGACGA	64.0	
Rbms3_Prot_Forward	TGTCAACAAAGGCAATCCTG	63.7	201
Rbms3_Prot_Reverse	ACCGGGAGGTTGGAGATGTA	63.5	
Rbp4l_AS_Forward	AGTTCTGTCTTTGACTTACCAGACT	59.6	575
Rbp4l_AS_Reverse	TTTCATCTCCGGCTGCTCAG	60.1	
Rbp4l_Prot_Forward	CGGAGGTGACAACTACTGGG	59.75	231
Rbp4l_Prot_Reverse	GTTTAGCAAGCACCAGACTGC	60.07	
Tfap2b_AS_Forward	GTTGGGAGTGGGGTTGTTCA	60.1	553
Tfap2b_AS_Reverse	CCATGTGGTTACGCAGCTAC	59.0	
Tfap2b_Prot_Forward	ATATTCCTCCACGGATCGCCA	61.1	555
Tfap2b_Prot_Reverse	ACGACGCTTTTGTGAGGAAC	59.06	
Gapdh_Forward	GTCTATAGCGAGAGGGACCCA	57.49	275
Gapdh_Reverse	TGACTCTCTTTGCACCACCC	59.06	

2.3 Nuclear & cytosolic fractionation of zebrafish embryos in different stages of development

Zebrafish embryos were collected and dechorionated as described above. Dechorionated embryos were then transferred to a 1.5mL microcentrifuge tube and then washed twice with 1 X PBS (Phosphate buffer solution, ThermoFisher) containing 0.25% bovine serum albumin without disturbing the embryos. 200 µL of RLN buffer (Table 2.3.1) was then added to the tube containing embryos and the yolk sac was disrupted by pipetting (200 µL pipette tip) releasing the cells buffer. The tube was then incubated on ice for 5 mins and centrifuged at 3700 r.p.m. for 2 mins. The resulting supernatant was collected and labelled as the cytosolic fraction and the pellet was further resuspended in 200 µL of RLN buffer for 30 mins. After the incubation the tube was centrifuged at 3700 r.p.m. for 2 mins and the supernatant discarded. The pellet was washed twice with 200 µL of RLN buffer and finally collected and labelled as the nuclear fraction. Table 2.3.1 gives a list of reagents used to prepare the buffers used for fractionation.

2.3.1 Extraction of RNA from Nuclear & Cytosolic fractions

To the cytosolic fraction, 700 µL of RLT buffer and 500 µL of 100% ethanol was added and the mix was pipetted 10 times to dissolve the fraction. The mix was applied to a column (700 µL at a time) and centrifuged at 8,000 x g for 15 sec. The flow through was discarded and the column washed once with 500 µL of RPE buffer and once with 80% ethanol. The flow through was discarded and the column was given a blank spin to remove any traces of ethanol. Finally, RNA was eluted in 35 µL of RNase free water. To the nuclear fraction, 350 µL of RLT buffer and 350 µL of

70% ethanol was added and the mix homogenised using a 1 mL syringe. The mix was applied to a column (700 µL at a time) and centrifuged at 8,000 x g for 15 sec. The flow through was discarded and the column washed once with 700 µL of RW1 buffer and then twice with 500 µL of RPE buffer. The column was then given a blank spin and the RNA eluted in 25-30 µL of RNase free water. The RNA extracted from the two fractions was treated for genomic DNase as explained before. A portion of the DNase treated RNA from both fractions was used to synthesize cDNA, as explained before, while the rest of it was kept for RNA and CAGE sequencing. The concentration of the RNA extracted was determined using the Nanodrop.

Table 2.3.1 List of reagents used for embryonic cell fractionation

Buffer	Stock	Working
Deyolking buffer (with Ca^{2+})	1M KCl, 5M NaCl, 1M $NaHCo_3$ & 1M $CaCl_2$	1.8mM KCl, 55mM NaCl, 1.25mM $NaHCo_3$ & 2.7mM $CaCl_2$
RLN buffer	1M Tris Cl (pH 8.0), 0.1M $MgCl_2$, 5M NaCl & 10% NP-40 (IGEPAL)	50mM Tris Cl (pH 8.0), 1.5mM $MgCl_2$, 140mM NaCl & 0.5% NP-40 (IGEPAL)

2.3.2 PCRs on RNAs from nuclear & cytosolic fractions

Based on the literature available so far on zebrafish embryogenesis we chose three markers to test our fractions. We used *hnrnpa1* RNA, *lsm1* (U6) snoRNA as nuclear marker and *gapdh* as a cytosolic marker. We also looked at the RNA levels of one of our protein-coding and antisense pairs, *Rbms3*, in these fractions. PCR was carried using Bioline MyTaq kit under the following conditions: 94°-1 min, 94°-15 sec, T_m15 sec, 72°-10 sec, 94°-1 min, 4°- $\propto$ for 35 cycles, and a total reaction volume of 50 μl with 50 ng of cDNA per reaction. Table 2.3.2 gives the list and sequence of primers used.

We also sequenced the PCR product we got for the Rbms3 AS to make sure that it is amplifying only the Rbms3 AS.

Table 2.3.2 List of primers used to test the fractions

Primer Pair	Sequence	T_m (melting temperature)	Product Size
hnrnapa1_Forward	CCAAAGAGCAACAGACCCCT	59.89	180bp
hnrnapa1_Reverse	TGACGAAGCCAAATCCCCTC	60.04	
lsm1(U6)_Forward	TCTTGCTTGGAGAAGTAGATCTGG	59.84	100bp
lsm1(U6)_Reverse	GCTTGCTGTTCTGTACGCTG	59.84	

2.4 RNA-sequencing

RNA was extracted from the nuclear and cytosolic fractions of the embryos in different stages as explained before using the RNeasy mini kit from QIAGEN and DNase treated using the Sigma kit (Fig 2.4.1). The quality of the RNAs extracted was detected on the RNA TapeStation using Agilent High Sensitivity RNA Screen Tape assay. Sequencing libraries were prepared using the TruSeq Stranded Total RNA library Prep kit with the Illumina Ribo-Zero rRNA removal kit (Human-Mouse-Rat). The whole process from quality check to library preparation and sequencing was carried out by the University of Birmingham Genomics facility.

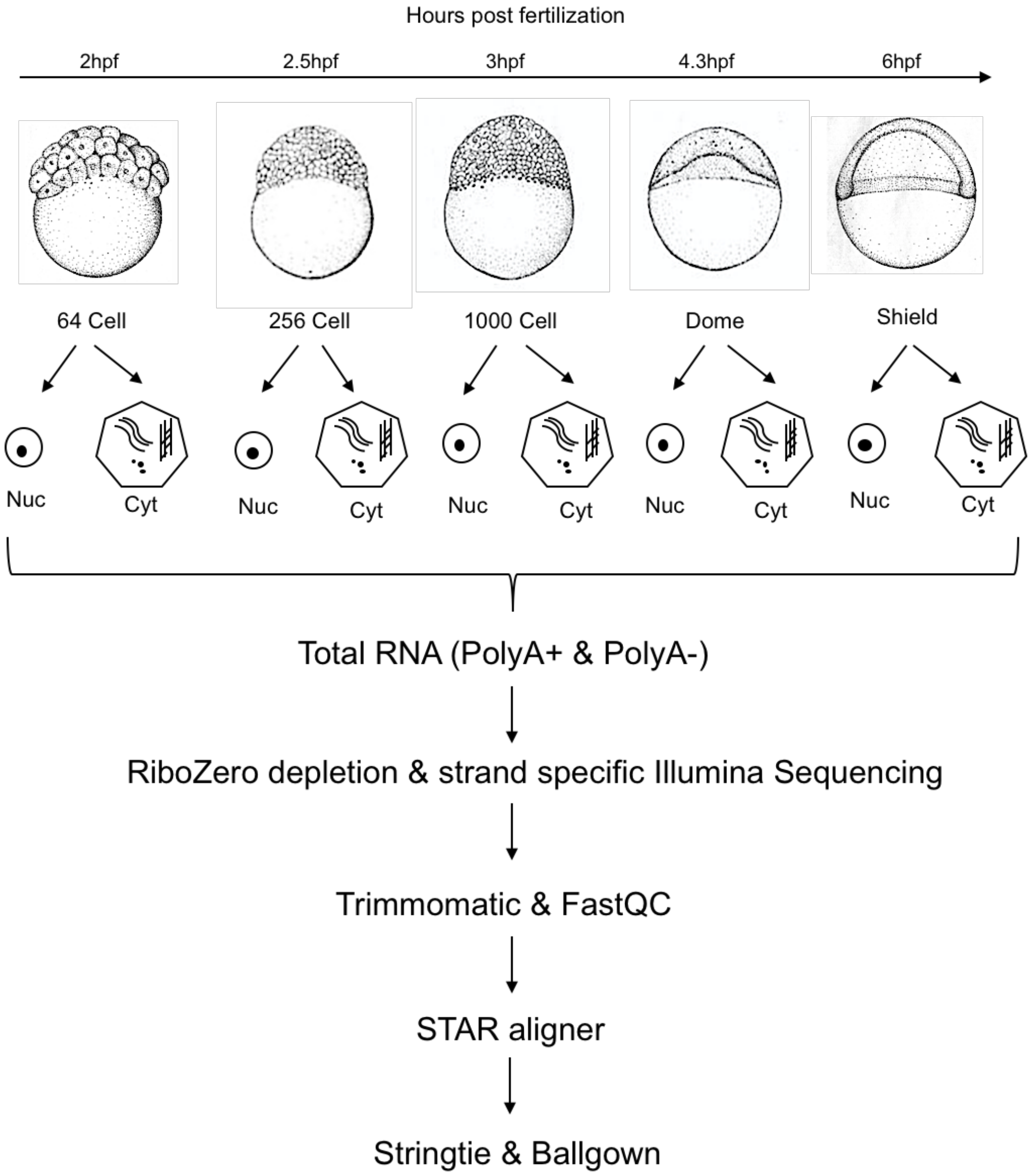


Figure 2.4.1 RNA library preparation.
The figure above shows a schematic flowchart explaining the fractionation step and library preparation from zebrafish embryonic cells in different stages of development (64-Cell, 256-Cell, 1000-Cell, Dome and Shield).

2.5 CAGE-sequencing (Cap Analysis of Gene Expression)

CAGE library preparation was done using a modified version of cap trapping from Carninci et al. (Carninci et al., 1996) (Fig 2.5.1). RNA was extracted using the RNeasy mini kit from QIAGEN and we started with 1µg of per sample. We divided each sample into four parts, 250ng of RNA to be pooled later after cDNA synthesis. The first step is the priming of cDNA at 65°C for 5min using a random primer mix from RNA and then synthesised the cDNA with an enzyme mix (5X FS buffer, 10mM dNTPs, 0.1X DTT, H_2O, SSIII (200U/uL, Thermo Fisher) at 25°C for 30sec, 50°C for 60mins and then cooled at 4°C to stop the reaction. This is followed by successive washing steps with RNACleanXP or magnetic beads. The next step is the oxidation of the 5' cap of RNA with 250mM NaIO4, followed by washing with beads and then successive biotinylation of the oxidised 5' cap with 100mM Biotin (long arm) Hydrazide (Sigma-Aldrich) at 40°C for 30mins. Before capturing the biotinylated double strand cDNA and RNA structure, any single strand RNAs which failed to extend to 5' ends are cleaved using RNase ONE (Promega™). This step leaves only complete cDNA-RNA ds structures intact which are captured in the next step by Cap-trapping using Streptavidin beads at 37°C for 15mins. This is followed by several steps of stringent washes with beads and then the release of 5' complete cDNAs from the capped RNAs and the addition of single strand 5' linkers (nAnTi 5'linker) at 30°C for 4hrs. Unlike the cap trapping we skipped the second strand cDNA synthesis and cleavage with *EcoP15I* but carried out the 3' end single strand linker (nAnTi 3'linker) ligation containing the Illumina primer sequence at 16°C for 16hrs. The concentration of the libraries was checked using KAPA SYBR® FAST qPCR kit (KR0389) and then on a Bioanalyzer (Agilent) with the final concentration of the libraries adjusted to 12pM. These were then sequenced on the Illumina HiSeq2500 and ran on High output mode (paired end, 50bp).

Library Preparation

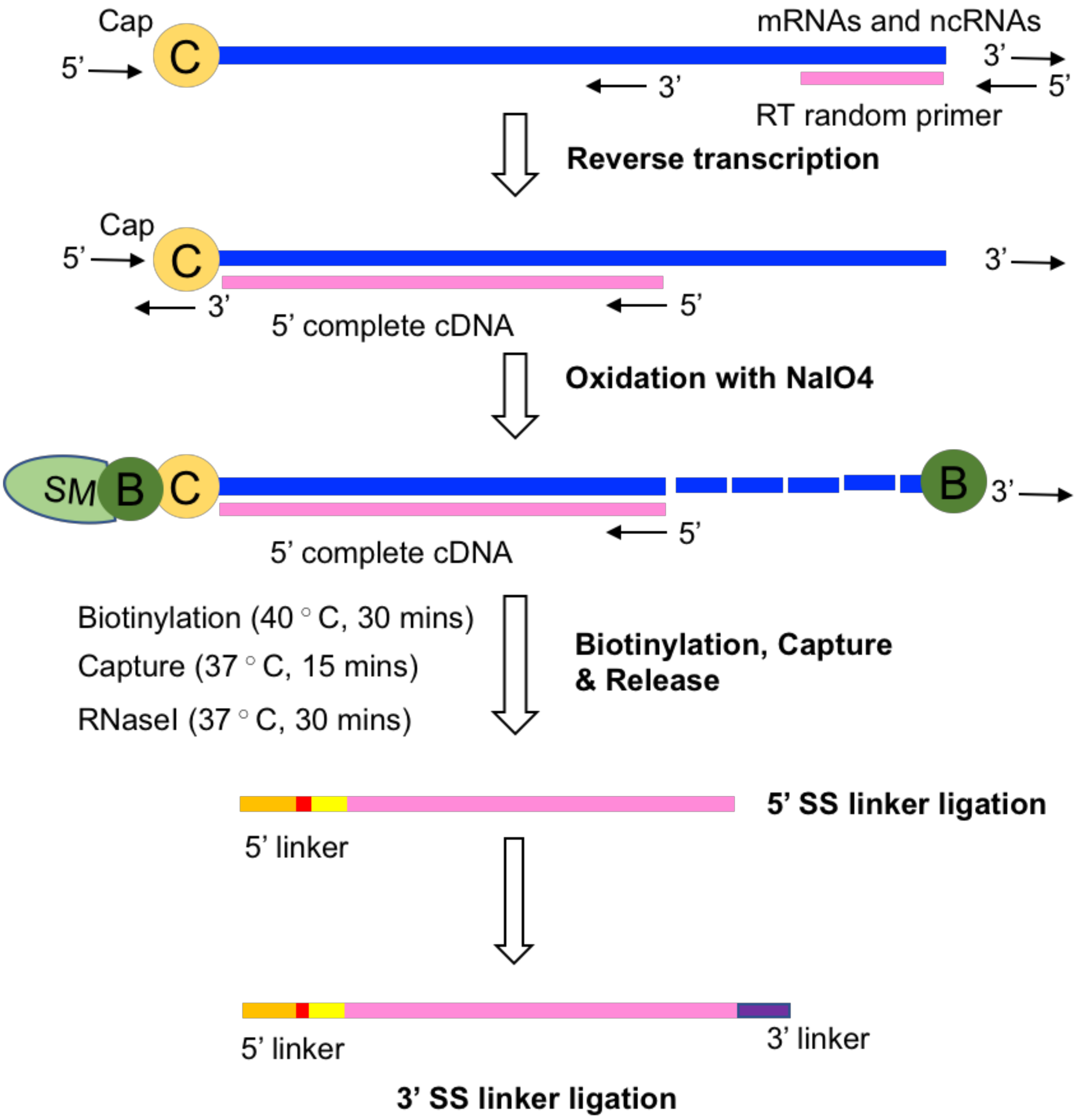


Figure 2.5.1 CAGE library preparation.
The figure above shows a schematic flowchart explaining the Low Quantity CAGE library preparation. The capped RNAs are shown in blue and the synthesised cDNA are in pink. The cap is represented as C, the B represents the biotin attached to the cap and the SM shows streptavidin beads.

2.6 Extraction of protein from Nuclear & Cytosolic fractions

To the nuclear fraction, in the form of pellet, we added 50-100 µL of lysis buffer and incubated on ice for ~1hr to lyse the nucleus. The cytosolic fraction, already in RLN buffer, obtained from the different stages was used as it is for western blotting. Both the fractions were quantitated using the Bradford reagent for estimation of protein.

2.6.1 Western Blotting

For western blotting, 2 µL of 2X SDS sample buffer per embryo was added to each nuclear and cytosolic fraction which were then incubated at 95°C for 10 mins. Samples were then centrifuged at max. speed to pellet any debris and the supernatant was used to run on the gel. We used 12.5% acrylamide gels and loaded equal amounts of nuclear and cytosolic fractions for semi-quantitative analysis purposes. Semi-dry transfer on a nitrocellulose membrane was used followed by blocking in 2.5% skimmed milk (1X PBS) at RT. Primary antibody incubation was carried out overnight at 4°C, washed three times for 5 mins with 1X TBST with shaking, incubated in secondary antibody at RT for 1hr and then finally, washed twice for 5 mins with 1X TBST and then once with 1X TBS. Signal was detected using the Odessey scanner from LI-COR Biosciences and analysed using the ImageStudio Biolite software. Antibodies used were RNAPol II (abcam), Ezh2 (CST), Gapdh (CST) and PCNA (CST) at concentrations suggested by the manufacturer.

Table 2.6.1 List of reagents used for Western Blotting

Reagent/Buffer	Stock	Working
12.5% SDS Polyacrylamide gel	1.5M Tris (pH8.8), 30% acrylamide, 10%SDS, 10%APS	For 4 gels, ddH_2O-9.91mL, 1.5M Tris (pH8.8)-6.25mL, 30% acrylamide-8.33mL, 10%SDS-250µL, 10%APS-250µl & TEMED-20µl
4% Stacking gel	0.5M Tris (pH6.8), 30% acrylamide, 10%SDS, 10%APS	For 4 gels, ddH_2O-5.55mL, 0.5M Tris (pH6.8)-2.5mL, 30% acrylamide-1.7mL, 10%SDS-100µL, 10%APS-100µl & TEMED-20µl
10X Running buffer	60g Tris & 288g glycine in 2L of ddH_2O	1X running buffer (100mL + 10mL 10% SDS, make upto 1L)
1X TBS	1M Tris (pH-8.0), 5M NaCl	1M Tris (pH-8.0)-40mL, 5M NaCl-60mL & make upto 2L with ddH_2O
1X TBST	1M Tris (pH-8.0), 5M NaCl	1M Tris (pH-8.0)-40mL, 5M NaCl-60mL & tween-2mL, make upto 2L with ddH_2O
Blocking buffer	-	5% skimmed milk in 1XTBS
Primary Antibody	As per provided	1:1000/1:2000 dilution in 3%BSA

2.7 Data Analysis

2.7.1 Identification & Categorisation of protein-coding and antisense pairs

To identify and characterize overlapping protein-coding mRNA and antisense non-coding RNA pairs we decided to consider all the protein-coding and antisense transcripts predicted in zebrafish by ENSEMBL (Zerbino et al., 2018). We took the annotations from the latter for these predicted protein-coding and antisense transcripts and ran a BedTools (with a -S option) to identify the pairs that overlap with each other.

2.7.2 Analysis of RNA-Sequencing data

2.7.2.1 Pauli et al. RNA-Seq Data analysis

We looked at the RNA-Sequencing data available during zebrafish embryogenesis (Pauli et al., 2012). Raw RNA-sequencing reads were taken from the study on 8 different developmental stages in zebrafish. The raw fastq files were downloaded and quality checked and trimmed using FastQC (developed by Andrew, S., Barbraham Bioinformatics) and Trimmomatic (v0.27) (Bolger et al., 2014), respectively. The reads, for 8 different developmental stages, were then mapped back to the Zv9 or DanRer7 genome assembly of zebrafish using TopHat (Bowtie2 v2.1.1) (Kim et al., 2013). The bam files (aligned reads) generated were then used for transcript assembly using Cufflinks v2.2.1. The resulting transcript expression data for each stage were used to extract the RNA levels of protein-coding and antisense transcripts in a pair (Trapnell et al., 2012). This was then used to calculate the correlation between the RNA levels of protein-coding and antisense transcripts in a pair. We then categorised them based on whether the protein-coding and antisense pairs correlated positively or negatively or did not show any correlation at all.

For visualisation purposes, we generated strand specific expression peaks using bedtools genomecov to convert the bam files into bedgraphs, and uploaded the tracks on the genome browser, specific to each strand.

2.7.2.2 Nuclear RNA-seq data analysis

For the RNA sequencing data generated by us we used a different approach for analysis. The raw sequencing reads obtained were mapped to the Zv9 genome assembly of zebrafish using STAR (v2.6.0a). As opposed to TopHat, STAR (Spliced transcripts alignment to a reference) is faster and more accurate in aligning the reads to the genome and detects complex RNA structures such as chimeric and circular RNAs (Dobin and Gingeras, 2015). The bam files generated were then assembled into transcripts using StringTie, which identifies more accurately and higher number of transcripts when compared to Cufflinks (Pertea et al., 2015). I used both and personally, found StringTie to be faster and identified more annotated transcripts than Cufflinks. In StringTie using the -B and -b options we produced files for differential expression analysis using Ballgown (Pertea et al., 2016). This gave us two csv format files, one for genes and the other for transcript expression (in transcript per million or TPM) in different stages of development. It also outputs exon (e.tab) and intron (i.tab) differential expression profiles. These files were used for all our downstream analysis.

2.7.3 ChIP sequencing analysis

The input fastq files for ChIP-Seq data on histone modifications were taken from the DANIOCODE repository, a database on existing transcriptional data in zebrafish. Again, for mapping the raw reads to the genome we used STAR aligner (v2.6.0a). We analysed the bam files, using HOMER to create tag directories and annotating peaks using the makeTagDirectory and findPeaks tools (-style factor and -o auto options), for the different marks H3K27me3, H3K4me3 and H3K27ac in the Dome (4hpf) stage of development (Vastenhouw et al., 2010, Zhang et al., 2014, Bogdanovic et al., 2012).

2.7.4 Association of AS non-coding and protein-coding genes with histone marks

We used ChIP-sequencing data to observe whether the protein-coding transcripts associated with antisense non-coding in the negatively and positively correlated categories showed any enrichments for histone modifications either at the TSS (transcription start sites). We looked at the histone modifications, H3K27me3 which is a histone modification associated with the repression of gene expression, H3K4me3 which is associated with transcription initiation and H3K27ac associated with promoter activation. We used a similar set of analysis to observe the distribution of CAGE tags across the distal peaks (± 2kb) for H3K27me3 mark and compared it to ChIP reads for H3K4me3 and H3K27ac.

2.7.5 CAGE Sequencing Analysis

The raw tags from CAGE sequencing were mapped using STAR aligner and the resulting BAM files were used in the bioconductor package CAGEr for further downstream analysis. CAGEr starts from mapped reads and does quality filtering, normalization, removal of the additional 5' end G nucleotide (added during the CAGE protocol) and the frequency of the usage of start sites (Haberle et al., 2015).

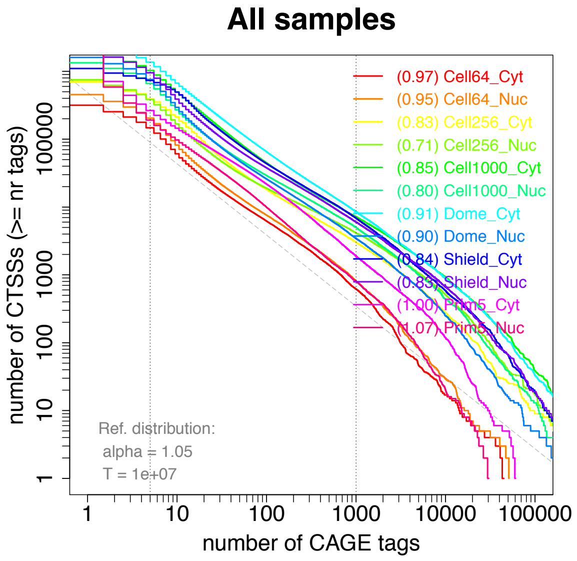
2.7.5.1 Normalization of CAGE tag counts using the power law distribution

The mapped raw tag counts from different samples, in this case stages, vary in sizes and therefore need to be normalized. Many studies on high-throughput sequencing, however, still uses number of tags per million as a measure of CAGE expression. In 2009, Balwierz et al. (Balwierz et al., 2009) proposed a reference power law distribution for normalization of CAGE tag counts in different samples. The *normalizeTagCount()* function plots a reverse cumulative distribution, between the number of CAGE TSSs that are equal or greater than the number of tags (Y-axis) and the number of CAGE tags (X-axis). This further assists us in choosing the appropriate parameters for normalization of dataset. Figure 15 below shows the output reverse cumulative plot for the 12 zebrafish samples we considered in this experiment. On plotting the raw tag counts (Fig 2.7.1a) we observed the slope alpha which is the median of slopes from individual samples and T being the total number of tags which is the power of 10 to the median sequencing depth of samples. Once we had the alpha (1.05) and T values (1e+07) we were able to plot another reverse cumulative plot with the normalized counts in which all the samples followed a power law distribution (Fig 2.7.1b).

2.7.5.2 Exporting the normalized CAGE signal to bedGraph for genome browser visualisation

The CAGEr protocol also supports the export of raw or normalized CTSS (CAGE TSS) files consisting of tag counts to a bedGraph format which can then be visualized on the genome browser. The *exportCTSStoBedGraph()* function exports the genomic location and tag counts of CTSSs. It produces two separate CTSSs files per sample for each strand.

a)



b)

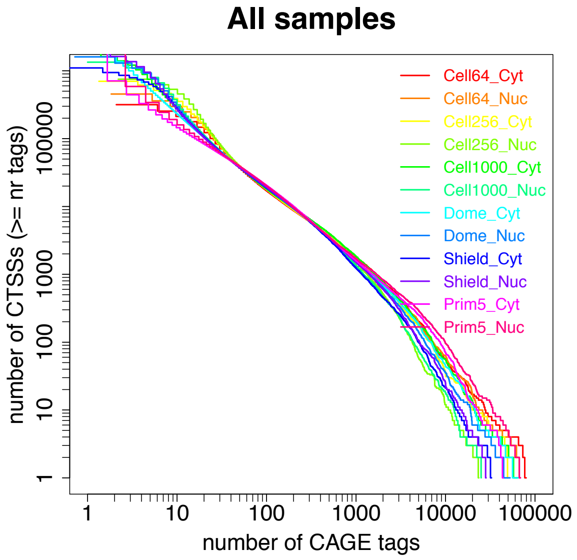


Figure 2.7.1 Power law distribution.
(a) The figure shows the distribution of CAGE tags per the number of CTSSs. The slope proposed by the power law is written next to each sample in colour while the grey dashed line represents the suggested power law distribution along with the parameters to be used for normalization. (b) distribution of CAGE tags after the normalization.

2.7.5.3 CTSS clustering and promoter width

Multiple CAGE tags that are within 20bp are clustered together into large transcriptional unit called tag clusters (TCs). Neighboring TSSs in close proximity to each other are mostly likely associated with the same promoter element and thus, are grouped into large clusters (TCs). The *clusterCTSS()* function outputs tag clusters for a given sample and estimates in each cluster the number of CTSSs, the position of the dominant CTSS. It also calculates the signal in tags per million (tpm) coming from each cluster as well as the dominant CTSS only. The *clusterCTSS()* function can be specifically told to perform clustering based on either distance or a parametric clustering based on the density of the signal. CTSSs with more than or equal to 1 TPM were considered for clustering.

Previous studies in human cell lines (Carninci et al., 2006) and during zebrafish development (Nepal et al., 2013) have identified two categories of promoters called “broad” promoters characterised by several CTSSs or “sharp” promoters with a distinct dominant CTSS. Thus, promoter width is an important feature to categorise different classes of promoters. Interquantile width is used to estimate the promoter width and is calculated by measuring the space between two quantiles of the total CAGE signal (also highlighted in Fig 2.7.2). These are calculated using the functions *cumulativeCTSSdistribution()* and *quantilePositions()* which can be exported as a BED file and viewed on the genome browser using the function *exportToBed()*. The *plotInterquantileWidth()* function plots histograms to compare promoter width across different samples.

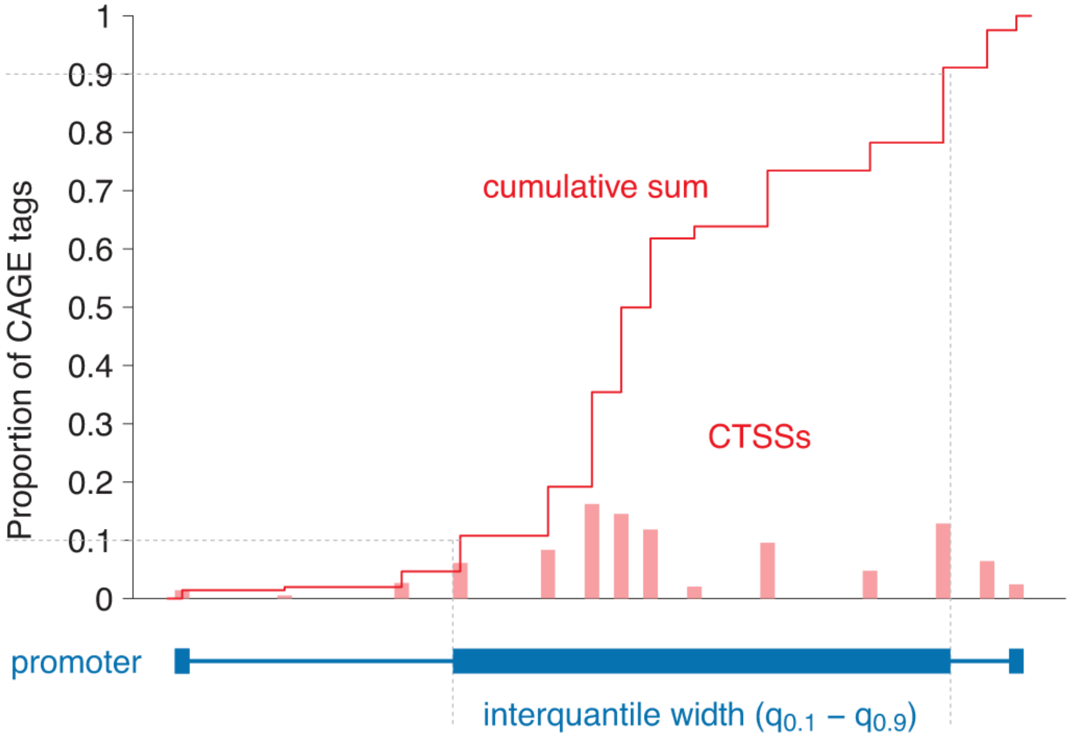


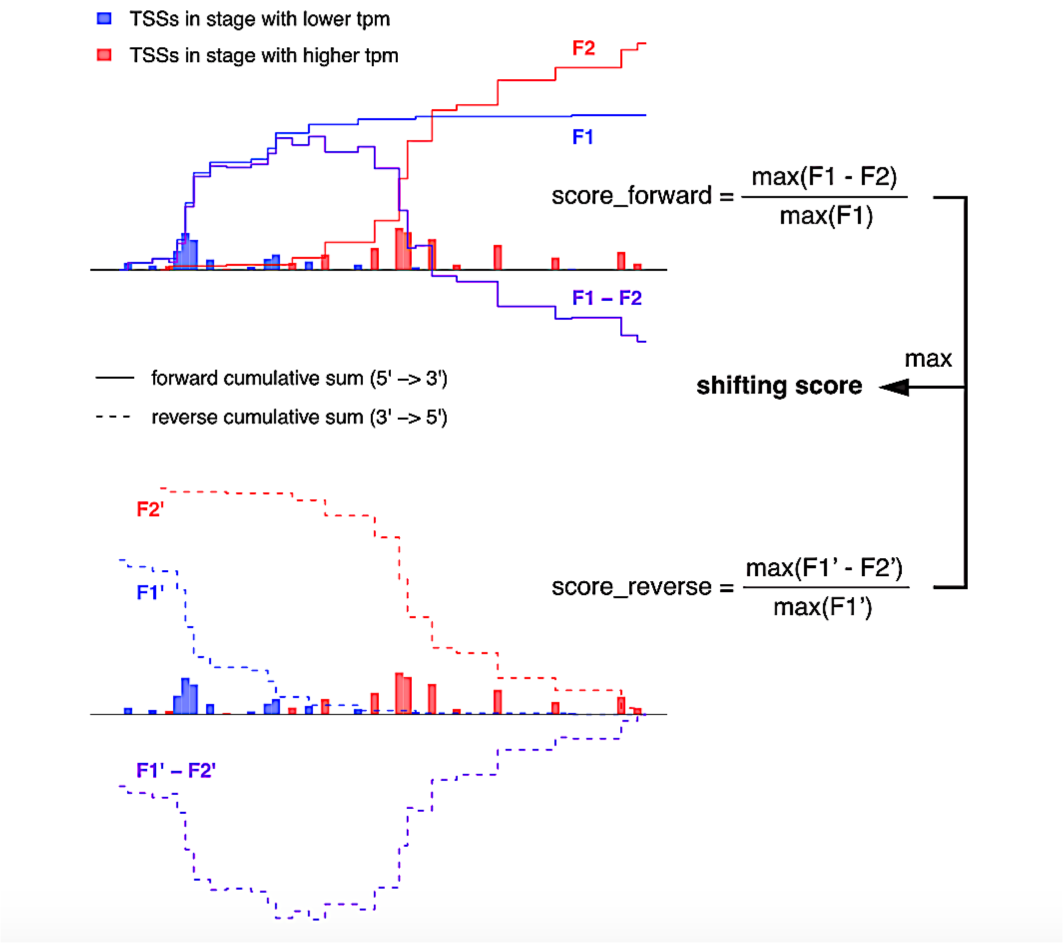
Figure 2.7.2 Interquantile width calculations.
The figure above explains how CAGEr calculates interquantile width. The X-axis represents promoter width which is the cumulative sum of all the tags that fall between the upper quantile position (q_{up}) and the lower quantile position (q_{down}). The Y-axis represents the proportion of CAGE tags. Figure taken from (Haberle et al., 2015).

2.7.5.4 Expression profiling

Since CAGE is also a measure of transcription it can be used for RNA expression analysis. Expression profiling can be done either at the level of individual CTSSs or on entire promoters. The former is done by considering CAGE signal at individual CTSSs across the different stages while in the latter case CAGE signal from an entire consensus cluster (promoter) is used. A consensus cluster is constructed using the *consensusClusters()* which aggregates all the overlapping TSSs and any neighboring TSSs within defined proximity of each other. CAGEr program uses two algorithms for clustering: *k*-means and the self-organizing maps (SOMS). These estimations are carried out by the *getExpressionProfiles()* function and then can be visualised using the *plotExpressionProfiles()* function. The *extractExpressionClass()* function helps in getting a specific expression class for further analysis. We used a CTSS threshold of 50TPM (as suggested in CAGEr protocol) for this analysis.

2.7.5.5 Shifting promoters

CAGEr also detects changes in the usage of TSS within the same promoter. It does so by comparing changes in cumulative distribution of CTSS within a consensus promoter from one sample to another. The function *scoreShift()* calculates a shifting score which is further explained in Fig 2.7.3. For example, a shifting score of 0.6 indicates that 60% (chosen randomly) of the CAGE signal detected in a stage is occurring outside the region that was used as TSS (within the same promoter) in another stage. Then using the *getShiftingPromoters()* function we could export the genomic coordinates, p-value and the shifting score for the promoters that show changes. It also gives us the position of the dominant TSS and expression levels. We used a shifting score of 0.6 to detect differential TSS usage pre-MZT between the 64-Cell stage and the 1000-Cell stage. We also observed differential TSS usage post-MZT between 1000-Cell and Prim-5 stage.



TSSs in stage with lower tpm
TSSs in stage with higher tpm
F2
F1
F1 – F2
score_forward = max(F1 - F2) / max(F1)
max
shifting score
forward cumulative sum (5' –> 3')
reverse cumulative sum (3' –> 5')
F2'
F1'
score_reverse = max(F1' - F2') / max(F1')
F1' – F2'

Figure 2.7.3 Promoter shift calculation.
The figure above shows the schematics of shifting score measurement. The blue and red lines indicate the cumulative distribution of CAGE signal within the same promoter between two samples. with F1 being the sample with the lower TPM and F2 with higher TPM. The purple line indicates the difference between F1 and F2. And the shifting score is the ratio of the maximum difference between F1 and F2 and the total CAGE signal at F1 (sample with lower TPM). Figure taken from (Haberle et al., 2015).

2.7.6 deepTools2.0

This software is used for quality control and normalisation of high throughput sequencing analysis such as RNA-Seq and ChIP-seq (Ramirez et al., 2016). Majority of the heatmaps in the thesis were plotted using deepTools computematrix and plotHeatmap tool. We also used deepTools for producing strand specific bigwig and bedgraphs (Fig 2.7.4) using the bamCoverage tool (--filterRNAstrand and –normalizeUsing BPM options).

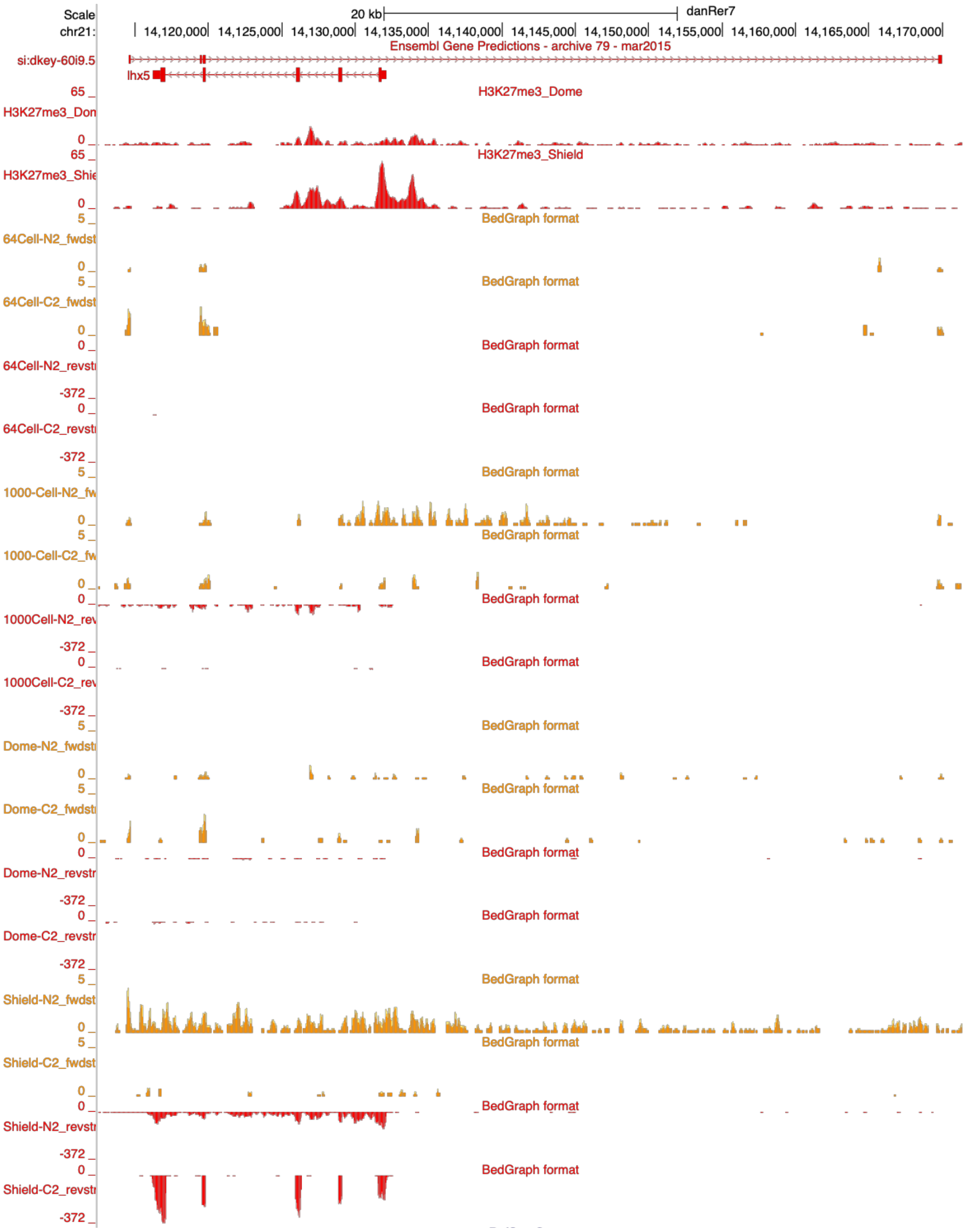


Figure 2.7.4 UCSC genome browser view of RNA-seq data
Strand-specific RNA expression data when viewed on the UCSC genome browser (Kent et al., 2002). H3K27me3 in dome and shield stages are ChIP-seq data from DANIOCODE. The RNA-seq data is from our fractionation sequencing with peaks in red and orange representing reverse and forward strand, respectively, in different stages of development.

2.7.7 BEDTools v2.28.0

BEDTools offers a range of tools to manipulate genomic features and use them for analysis (Quinlan and Hall, 2010). We used bedtools for calculating the coverage of repeats and genomic features such as introns, exons, 3’-UTR and 5’-UTR using the *coverage* tool with the -sorted -g “genomefile” -a “bedfiles” -b “BAMfiles” -s options. The annotations for the exons, introns and other gene features for Zv9 genome assembly were, in turn, downloaded using the R Bioconductor package “GenomicFeatures” (Lawrence et al., 2013).

Since BedTools (used to overlap the protein-coding and antisense transcripts) also gives us an idea about the overlap region using the *intersect* tool (-wo and -S options) between the protein-coding and antisense in a pair, we calculated the percentage overlap with respect to the length of the protein-coding transcript in the two categories.

2.7.8 Functional Analysis of overlapping protein-coding genes

We further looked at whether the protein-coding transcripts in the overlapping protein-coding and antisense pairs, belonging to the positively and negatively correlated group, are enriched in certain gene ontology based on their biological function. We used Database for Annotation, Visualisation and Integrated Discovery (DAVID 6.8) software for functional annotation of these transcripts (Huang da et al., 2009b, Huang da et al., 2009a). We especially focussed on the categories GOTERM (biological process), GOTERM (molecular function) and GOTERM (cellular component).

2.7.9 R-packages: ggplot2 and gplots

We used the ggplot2 bioconductor package for plotting all of our violin and box plots, using the geom_violin and geom_boxplot options (by H.Wickham 2016). We also used the geom_histogram() function to plot the histograms for showing the overlap region and the distance between the TSS of protein-coding genes and the antisense TES. The gplots R-package was used to create heatmaps showing the RNA levels of nuclear and cytosolic antisense and their overlapping protein-coding genes (used in sections 3.2 and 4.2).

CHAPTER 3: IDENTIFICATION AND CATEGORISATION OF ANTISENSE NCRNAS AND THEIR OVERLAPPING PROTEIN-CODING GENES DURING EARLY ZEBRAFISH DEVELOPMENT

3.1 Introduction

During vertebrate embryogenesis, one of the important events determining early development is the maternal-to-zygotic transition which is accompanied by zygotic genome activation. Previous studies on zebrafish embryogenesis have established the requirement for ncRNAs, such as miR-430 family, in the clearance of maternal transcripts. In the absence of miR-430 the embryos fail to develop beyond 30% epiboly (Giraldez et al., 2006). In addition, the role of ncRNAs in early developmental events such as the X-chromosome inactivation is well documented. A set of seven regulatory ncRNAs e.g. *Xist*, are required for X-chromosomal inactivation (Zhao et al., 2008). In addition, repression of developmental genes located in the *HoxD* cluster which are responsible for anterior-posterior patterning in all bilateral organisms is regulated by ncRNA *HOTAIR* or HOX antisense intergenic RNA (Rinn et al., 2007). Studies in humans have identified several cis-acting transcriptional start site-associated ncRNAs (TSSa RNAs) that repress developmental genes by recruiting histone modifying polycomb repressive complex, PRC2 (Kanhere et al., 2010). All these examples provide evidence towards the importance of lncRNAs in development.

On this background, identification of large number of lncRNA transcripts (~1133) during early zebrafish development highlights the need for the importance of these lncRNAs in development (Pauli et al., 2012). Interestingly, almost half of the lncRNA transcripts (566) identified during early development were expressed in an antisense manner showing exonic overlap with the protein-coding genes (Fig 3.1.1). These type of antisense lncRNAs, have been implicated in regulation of their overlapping protein-coding transcripts (Faghihi and Wahlestedt, 2009). Also, this class of ncRNA has not yet been explored in detail with respect to zebrafish embryonic development. One of the objectives of this thesis was to categorize these antisense ncRNAs into groups and understand their relationship with overlapping protein-coding transcripts based on the published deep RNA sequencing data for eight zebrafish developmental stages (Pauli et al., 2012).

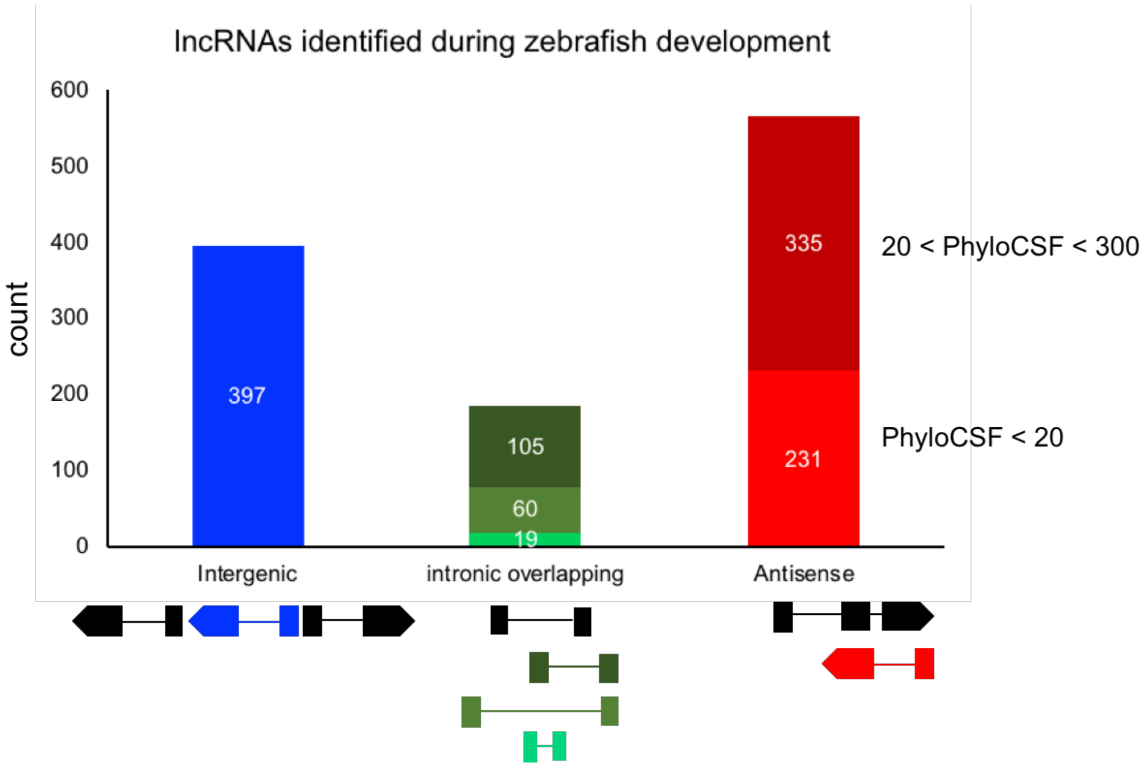


Figure 3.1.1 LNCRNAs present during zebrafish development
The figure shows the 1133 lncRNAs expressed zebrafish embryogenesis out which 397 were categorised as intergenic lncRNAs, 184 intronic overlapping lncRNAs and the rest 566 were antisense lncRNAs. The intergenic lncRNAs (labelled in blue) are localised in the genome between two protein-coding genes (in black). The intronic overlapping, in different shades of green, are lncRNAs found overlapping introns of protein-coding genes which could be further categorised into three categories (mRNA intron overlapping exon of ncRNA or all exons of mRNA overlapping intron of ncRNA or all exons of ncRNA overlapping intron of mRNA) e). The lncRNAs in the red are transcribed in antisense direction and overlap exons of protein-coding genes on the opposite strand (in black). (figure taken from (Pauli et al., 2012))

3.2 Results

3.2.1 Analysis of RNA-seq data for eight developmental stages from an existing RNA-seq study

The Pauli et al. study carried out deep sequencing of polyA tailed RNAs from 8 early stages of zebrafish development (Table 3.2.1). To measure levels of non-coding transcripts and protein-coding transcripts, the raw sequencing data from Pauli et al. was reanalysed (Pauli et al., 2012). The sequencing reads were quality filtered, trimmed to remove adapters and then mapped to the Zv9 genome of zebrafish. The alignment summary for each sample is tabulated in Table 3.2.1. On an average, out of 38-40 million reads for each stage 30-35 million reads mapped uniquely to the genome (i.e. ~ 88-90% reads). The aligned reads were further processed using Cufflinks program to calculate RNA levels of different transcripts (Pertea et al., 2015). The levels were measured as Fragments per kilobase per million (FPKM) values for all the transcripts (both annotated and novel) in the different stages of development.

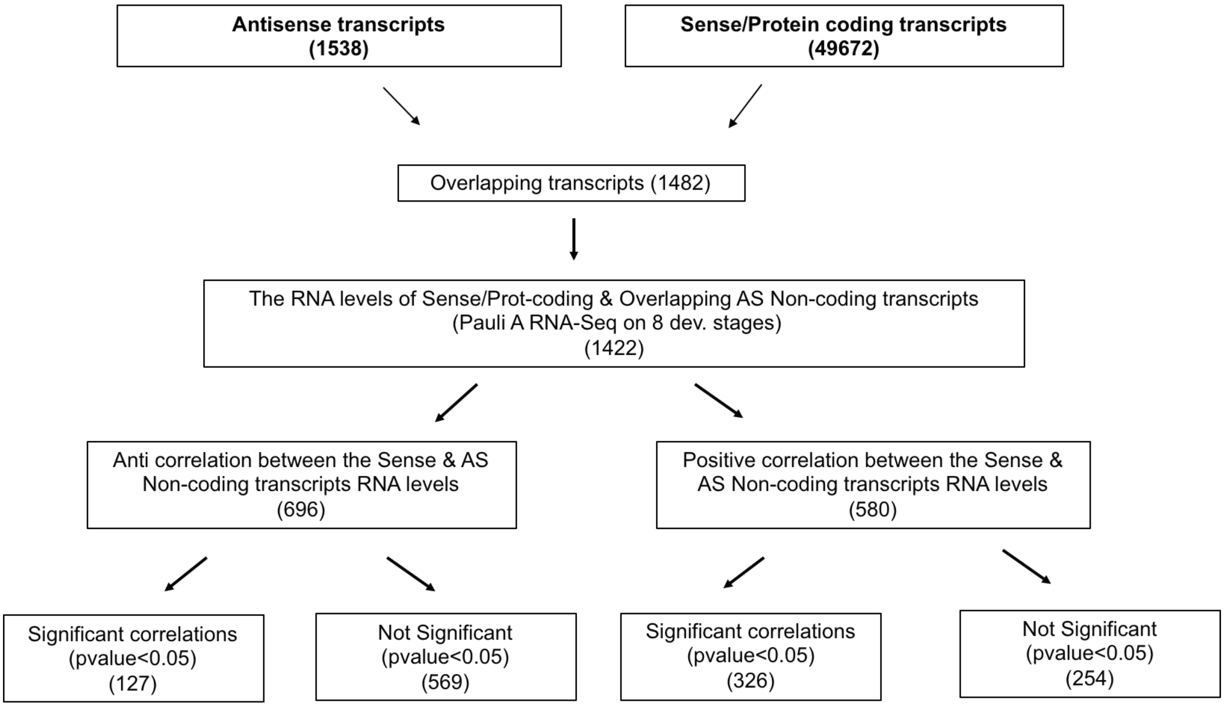
Table 3.2.1 Summary of RNA-seq alignment in 8 different stages of zebrafish development

Stage	Input Reads		Aligned Reads		Percentage Alignment
	Left reads	Right reads	Left reads	Right read	
2-4Cell (0.2hpf)	30,745,209	4,955,581	27,925,520	4,410,339	90.6%
1kCell (3hpf)	44,568,300	6,937,720	40,623,884	6,215,329	90.9%
Dome (4hpf)	42,579,162	6,761,161	38,687,577	6,049,445	90.7%
Shield (6hpf)	36,259,096	3,463,051	31,971,399	3,041,629	88.1%
Bud (10hpf)	43,055,780	7,549,530	38,679,487	6,798,714	89.9%
24hpf	41,952,775	7,123,408	37,531,430	6,409,422	89.5%
48hpf	43,256,341	7,577,505	38,786,945	6,836,188	89.7%

3.2.1.1 Characterisation of antisense ncRNAs and their overlapping protein-coding pairs during zebrafish embryogenesis

According to ENSEMBL, in Zebrafish genome (Zv9, zebrafish release 79), 1538 antisense ncRNAs and 49672 protein-coding (mRNA) transcripts are annotated (Zerbino et al., 2018). We intersected the ENSEMBL annotations for the protein-coding and antisense ncRNA genes and calculated the overlap as a minimum percentage (at least 10%) of length of antisense ncRNA gene. This revealed that there were 1482 overlapping protein-coding and antisense ncRNAs pairs in zebrafish (Fig 3.2.1).

A correlation value (as suggested by Richard Lowry, 2001-2009 for a sample size, N=8) was calculated between the levels of protein-coding RNA and antisense ncRNA in a pair using the RNA-seq (Pauli et al., 2012). Out of the 1482 protein-coding and antisense ncRNAs pairs, 60 pairs did not express at all in any of the developmental stages observed in the study and so were excluded. The 1422 protein-coding and antisense ncRNAs pairs, 696 protein-coding and antisense ncRNAs pairs were negatively correlated while 580 of them were positively correlated and the rest (146) showed no correlation. The pairs with significant correlations (p-value < 0.05 and r > 0.71) were retained for further analysis. As a result, 127 anti-correlated and 326 positively correlated protein-coding and antisense ncRNAs ncRNA pairs were shortlisted.



Antisense transcripts (1538)
Sense/Protein coding transcripts (49672)
Overlapping transcripts (1482)
The RNA levels of Sense/Prot-coding & Overlapping AS Non-coding transcripts (Pauli A RNA-Seq on 8 dev. stages) (1422)
Anti correlation between the Sense & AS Non-coding transcripts RNA levels (696)
Positive correlation between the Sense & AS Non-coding transcripts RNA levels (580)
Significant correlations (pvalue<0.05) (127)
Not Significant (pvalue<0.05) (569)
Significant correlations (pvalue<0.05) (326)
Not Significant (pvalue<0.05) (254)

***Figure 3.2.1 Pipeline used to identify and categorize overlapping protein-coding and antisense ncRNA pairs**.*
We took the annotations for all the antisense ncRNAs predicted by ENSEMBL and intersected with the annotated protein-coding genes. This gave us 1482 overlapping transcript information, out of which 1422 were present across the 8 developmental stages considered (Pauli et al., 2012) and showed either negative or positive correlation based on their expression.

3.2.1.2 Two different patterns of antisense ncRNAs and overlapping protein-coding transcript expression observed

The average as well as individual levels of the protein-coding RNAs and antisense ncRNAs in the positively correlated and negatively correlated groups were plotted (Fig 3.2.2). We found that the average abundance of all antisense ncRNAs in the negatively correlated group was higher in the pre-MZT stages while the protein-coding transcript levels increased post-MZT. Thus, suggesting that the antisense ncRNAs in the anti-correlated category were predominantly maternally deposited which eventually get degraded with zygotic genome activation (10hpf). In contrast, in the positively correlated category, both the antisense ncRNAs and the protein-coding transcripts were present throughout the 8 stages covered by RNA-seq data. Therefore, antisense ncRNAs and mRNAs belonging to the positively correlated group showed both maternal (0.75hpf to 10hpf) and zygotic contribution (24hpf onwards). These results were confirmed by plotting individual antisense – mRNA pair RNA levels in the negatively and positively correlated category in different stages of development (Fig 3.2.3a-b).

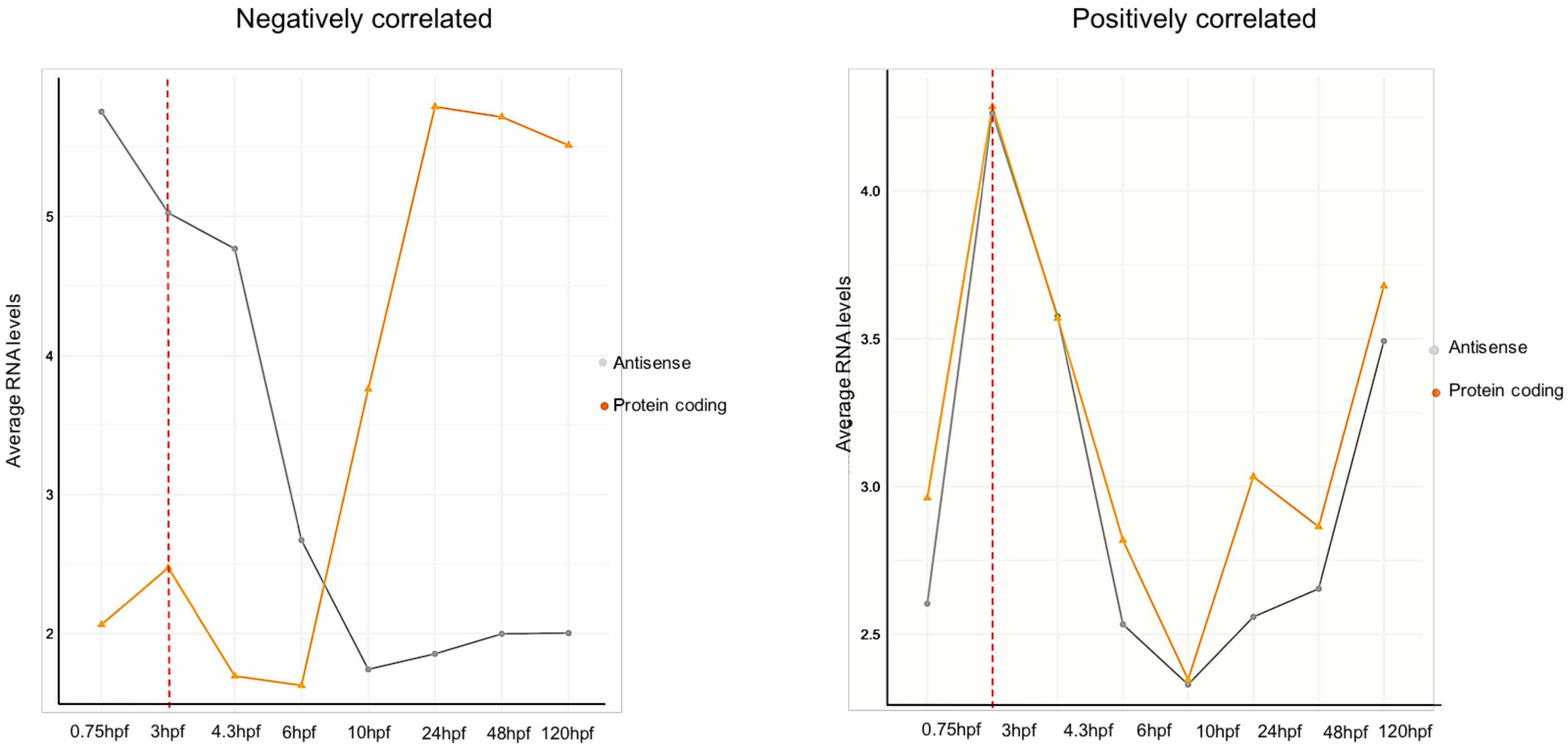
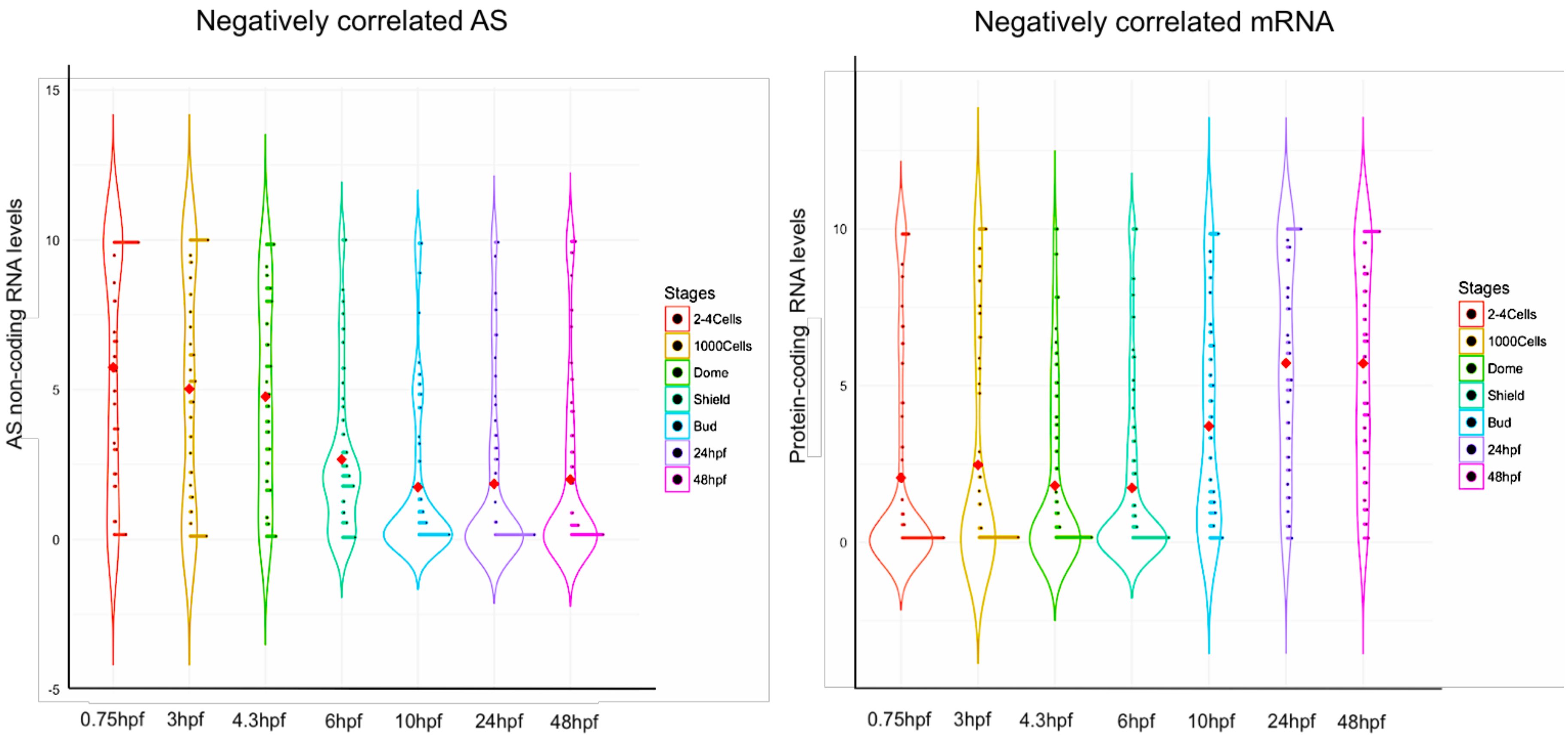


Figure 3.2.2 Two distinct patterns of antisense – protein coding expression.
The plot shows the average RNA levels of protein-coding and Antisense ncRNAs transcripts in 8 developmental stage of zebrafish. The X-axis shows hours post fertilisation while the Y-axis represents average RNA levels in FPKM which were colour coded as indicated by the key. And the red dotted line at 3hpf stage marks the activation of zygotic transcription. On the left is a plot of average RNA levels of antisense ncRNAs and corresponding protein-coding mRNAs in the anti-correlated group while on the right in the positively correlated category.

a)


Negatively correlated AS
AS non-coding RNA levels
15
10
5
0
-5
0.75hpf
3hpf
4.3hpf
6hpf
10hpf
24hpf
48hpf
Stages
2-4Cells
1000Cells
Dome
Shield
Bud
24hpf
48hpf
Negatively correlated mRNA
Protein-coding RNA levels
10
5
0
0.75hpf
3hpf
4.3hpf
6hpf
10hpf
24hpf
48hpf
Stages
2-4Cells
1000Cells
Dome
Shield
Bud
24hpf
48hpf

b)

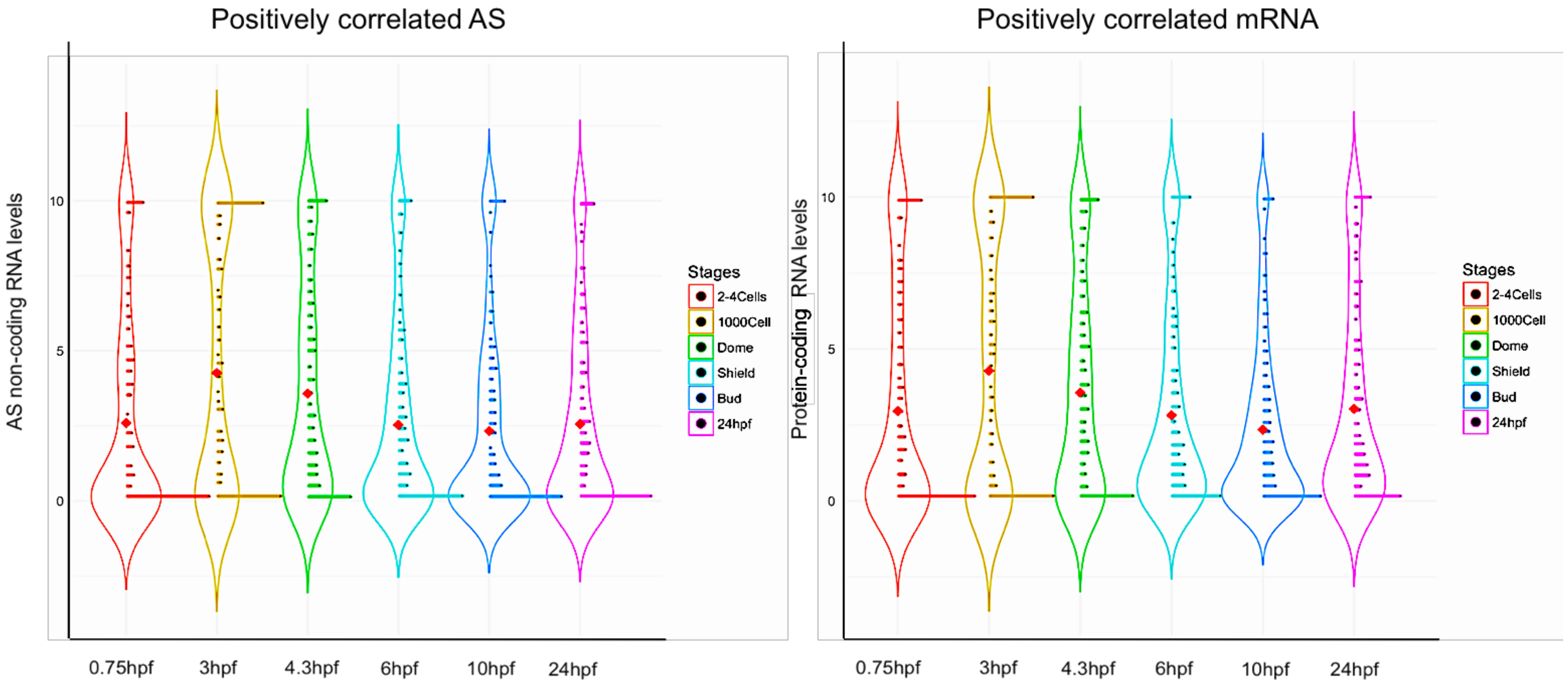


Figure 3.2.3 Individual RNA levels of antisense and mRNA in the two groups

Violin plot showing the RNA levels of individual antisense ncRNAs and protein-coding gene in both the negatively and positively correlated group. a) represents the antisense ncRNAs (left) and the protein-coding genes (right) in the negatively correlated category while b) shows the positively correlated group. The X-axis shows the hours post fertilisation which were colour coded as explained by the key on the right and the Y-axis represents the RNA levels (FPKM). The red dot marks the mean RNA levels.

3.2.1.3 Antisense examples confirmed the negative correlation observed for the protein-coding and antisense pairs in RNA-Seq data

Three examples of protein-coding and antisense ncRNAs pairs, *Rbms3*, *Rbp4l* and *Tfap2b* from the negatively correlated group were chosen for further downstream analysis based on their function during zebrafish development (Fig 3.2.4). Rbp4l, previously called purpurin, is a retinol binding protein that transports the retinol to developing retina. Morpholinos against *rbp4l* gene resulted in severe loss of cell differentiation of the retina in developing zebrafish embryos (Nagashima et al., 2009). The *rbp4l* antisense overlaps the entire length of the *rbp4l* protein-coding genes with the end of the antisense overlapping the start of the mRNA (Fig 3.2.4a). The antisense to *rbp4l* also overlaps two other protein-coding genes, *slc26a1* (full length overlap) and *PRC1* gene (start to start overlap). Rbms3 is an RNA-binding protein part of the *c-myc* single stranded interacting proteins called MSSPs (Penkov et al., 2000). Knockdown experiments of *rbms3* resulted in craniofacial defects in zebrafish suggesting Rbms3 by directly binding with *smad2* post-transcriptionally regulates the TGF-ßr (transforming growth factor ß receptor) pathway which is responsible for cartilage formation (Jayasena and Bronner, 2012). The *rbms3* gene has four isoforms with three shorter transcripts and a longer isoform that overlaps the whole length of the antisense (Fig 3.2.4b). TFAP2B is a DNA-binding protein with transcription factor activity and is essential for craniofacial development in zebrafish. It is co-expressed with another DNA-binding protein from its family of AP2 transcription factors, TFAP2A and promotes skeletal development in neural crest cells (Knight et al., 2005). The *tfap2b* gene shows five different isoforms with the start of all of them overlapping the end of the antisense (Fig 3.2.4c).

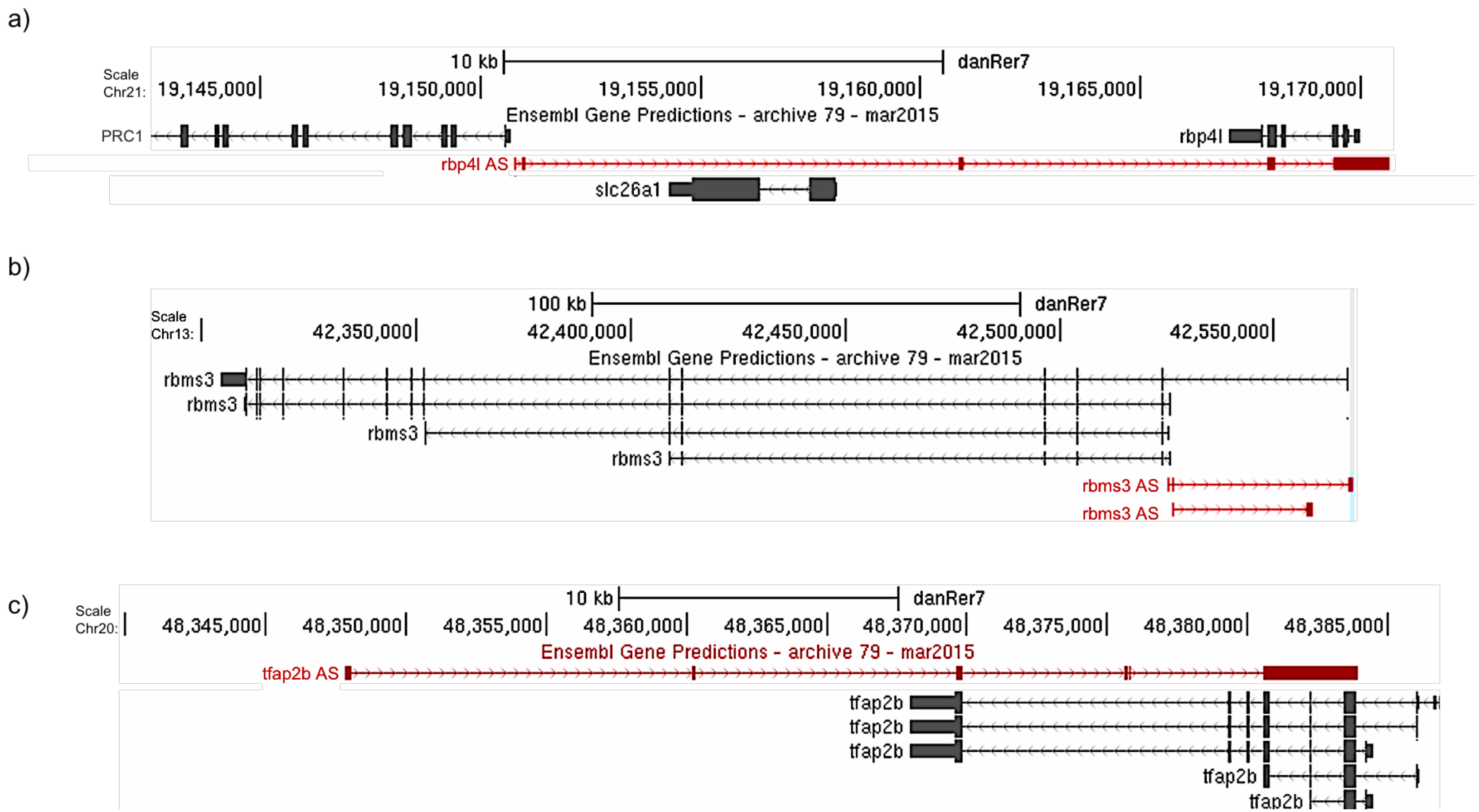


Figure 3.2.4 Localisation of antisense - protein-coding pairs on the UCSC genome browser (Kent et al., 2002).
a) Depicts the genomic position of rbp4l mRNA gene (grey) and its overlapping antisense ncRNA (red). It also shows two other genes slc26a1 and PRC1 that overlap the antisense ncRNA. b) The screenshot shows the arrangement of the Rbms3 mRNA gene (grey) and its overlapping antisense (red) in the genome. It also shows the presence of four different isoforms for the Rbms3 gene. c) Illustrates another antisense ncRNA (red) to the tfap2b gene (grey). The screenshot also shows the presence of five different isoforms for the tfap2b gene.

The expression pattern of the antisense RNAs and protein-coding mRNAs in these three pairs was profiled in different stages of development. We chose 3 stages of development, one at MZT (1000-Cell), post-MZT (Shield) and a later stage (24hpf) for our study (Fig 3.2.5). Rbp4l could be detected only in 24hpf (post-MZT) while the Rbp4l antisense ncRNA (Rbp4l-AS) expression goes down in 24hpf (Fig 3.2.6a). The example of Rbms3 was interesting as two protein-coding isoforms are annotated for this gene. The band at 533 bp is the desired band for the antisense. However, the higher band at ~800bp observed for Rbms3 AS in 24hpf (Fig 3.2.6b) is the full-length isoform of Rbms3 mRNA gene that starts expressing. The schematics explaining the position of the PCR primer used for Rbms3 AS and the PCR product sequenced (533 bp) is highlighted in Fig 3.2.5. The antisense to the Tfap2b mRNA gene also showed presence in the early stages but disappeared by 24hpf when the mRNA started expressing (Fig 3.2.6c). The PCR results for Rbp4l and Rbms3 protein-coding and antisense ncRNAs pairs suggest an increase in the expression of protein-coding RNA when the expression of the non-coding antisense ncRNAs decreases, which agree with the RNA-seq data.

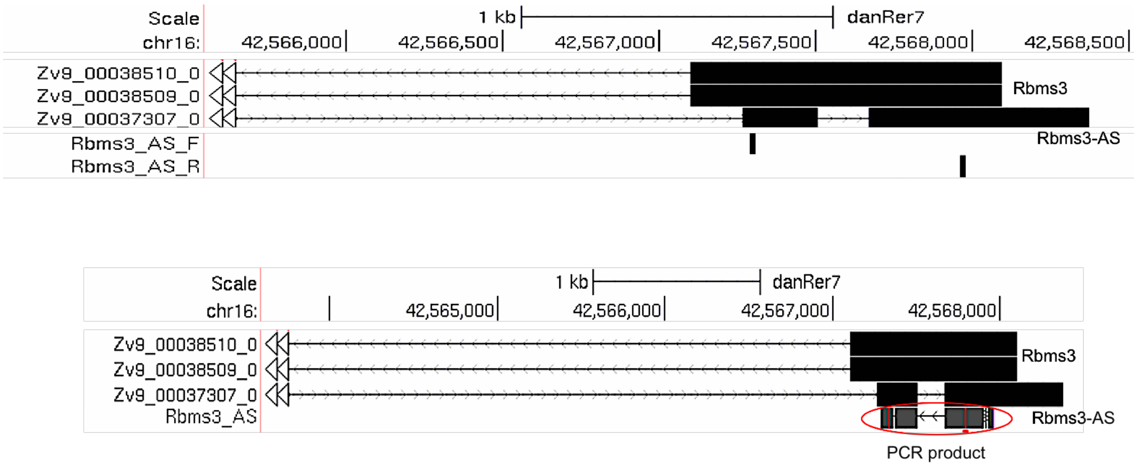


Figure 3.2.5 Primer design and PCR product for Rbms3 antisense.
The figure below gives a genome browser view of the primer design for Rbms3 antisense and the PCR product sequenced and mapped back to the zebrafish genome.

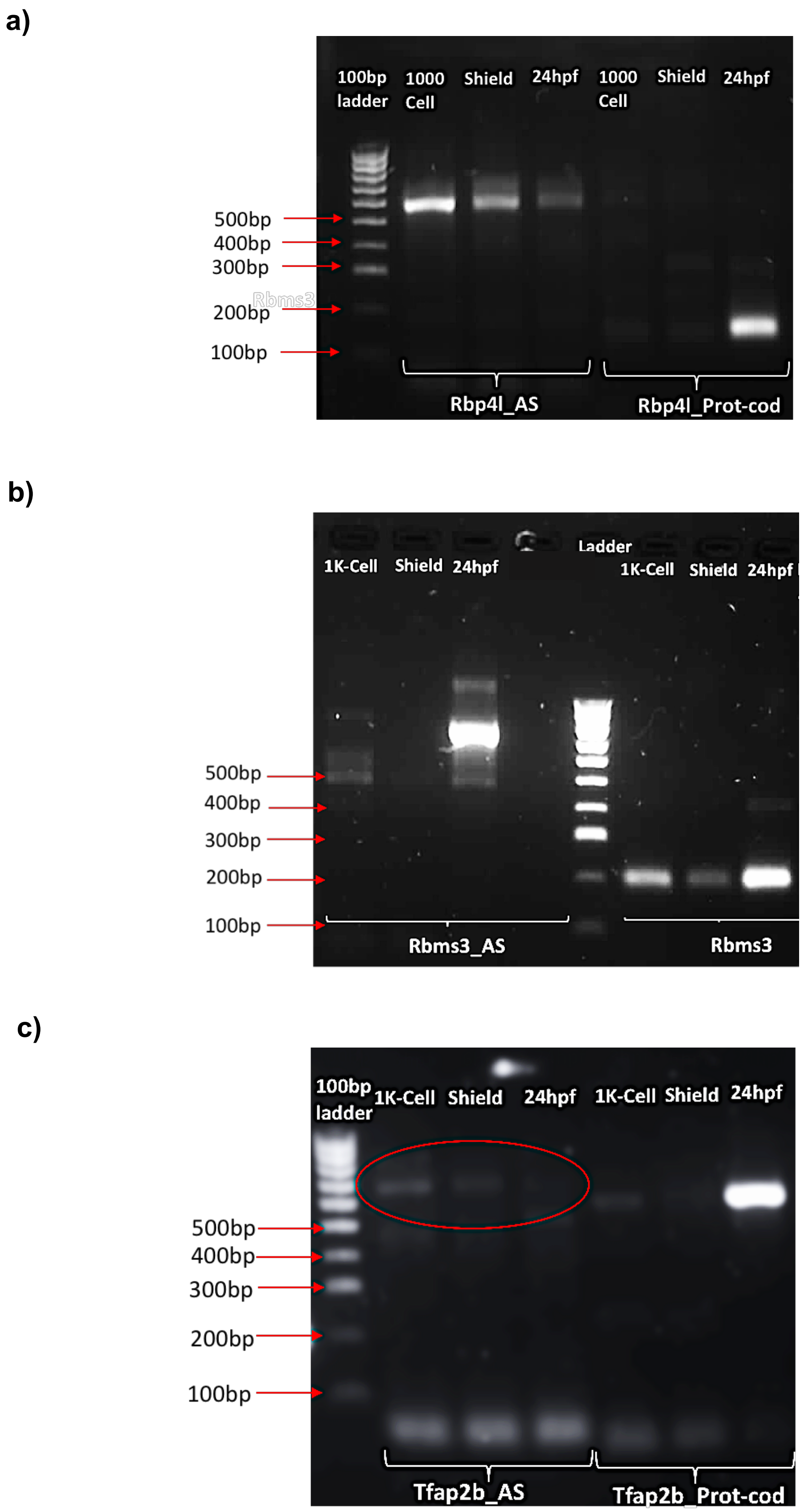

a)
100bp ladder
1000 Cell
Shield
24hpf
1000 Cell
Shield
24hpf
500bp
400bp
300bp
Rbms3
200bp
100bp
Rbp4l_AS
Rbp4l_Prot-cod
b)
1K-Cell
Shield
24hpf
Ladder
1K-Cell
Shield
24hpf
500bp
400bp
300bp
200bp
100bp
Rbms3_AS
Rbms3
c)
100bp ladder
1K-Cell
Shield
24hpf
1K-Cell
Shield
24hpf
500bp
400bp
300bp
200bp
100bp
Tfap2b_AS
Tfap2b_Prot-cod

Figure 3.2.6 Gel pictures showing the expression of antisense – mRNA pairs. *a) shows the expression of rbp4l antisense cDNA and rbp4l protein-coding cDNA in 3 stages, b) expression of rbms3 antisense cDNA against it overlapping protein-coding cDNA levels and c) comparison between the cDNA levels of tfap2b antisense and tfap2b protein-coding. We also additionally checked the quality of these RNAs using Gapdh as a control (APPENDIX, Fig 2).*

3.2.1.4 Protein-coding genes in the negatively correlated group were found enriched in developmental and homeobox proteins

Given the distinct expression pattern of genes in the positively and negatively correlated group, a pertinent question would be if the genes in the positively and negatively correlated group show differences in their biological and cellular functions. To answer this question, a Gene Ontology (GO) analysis was carried out using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang da et al., 2009b, Huang da et al., 2009a) on the protein-coding transcripts in the two groups. The two groups were enriched in distinct molecular functions (p-value < 0.05). The transcripts belonging to the negatively correlated group were enriched in homeobox containing (p-value of 7.65 x 10^{-12}), transcription regulation (p-value of 2.31 x 10^{-5}) and developmental genes with p-value equal to 2.06 x 10^{-7} (Fig 3.2.7). Thus, suggesting that these proteins are essential for transcription and development. The transcripts in the positively correlated group were enriched in housekeeping functions related to metabolism and signalling processes. The top category in this group were lipid metabolism (p-value of 0.0668), DNA binding (p-value is 0.082) and transport proteins with p-value equal to 0.052.

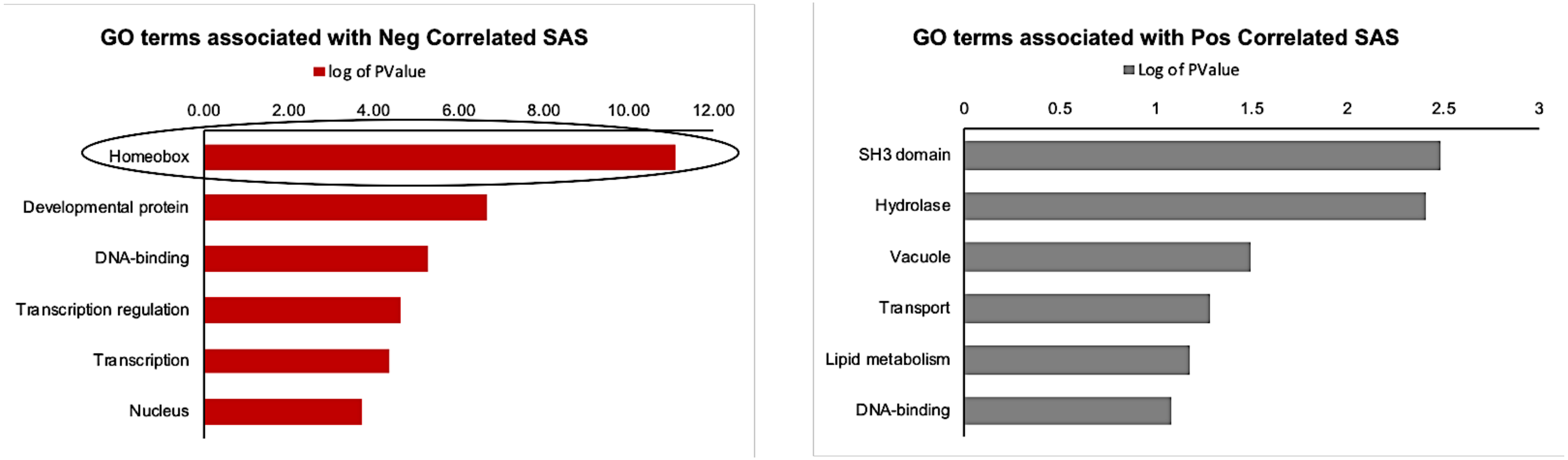


Figure 3.2.7 GO analysis on the protein-coding genes in the two groups.
The Y-axis shows the categories the genes were enriched in while X-axis shows the negative log of p-values. The circle represents the homeobox genes which were significantly enriched in the anti-correlated protein-coding genes. The chart on the left (red) shows the enrichment of mRNA genes in the negatively correlated group for functions related to DNA-binding and transcription regulation. The graph on the right (grey) shows gene ontology terms associated with the protein-coding genes in positively correlated category. The functional terms include hydrolase activity and transportation.

3.2.1.5 The overlap region between the antisense ncRNAs and the protein-coding gene was found to be longer in the negatively correlated category

It has been proposed that the antisense RNAs regulate expression of overlapping protein-coding genes (Wang and Chang, 2011). The analysis carried out here shows that the antisense RNAs can have either positive, negative or no correlation with the expression of overlapping protein-coding gene. This suggests that antisense RNAs might either repress or induce expression of overlapping protein-coding gene which is expressed in the opposite direction to the ncRNA. One manner in which the antisense ncRNAs keep the protein-coding genes repressed is by post-transcriptional RNA interference silencing pathway. The antisense RNA forms RNA-RNA hybrid with the mRNA, thus, preventing its translation (Villegas and Zaphiropoulos, 2015). Hence, it is important to further analyse the extent of overlap between the antisense ncRNA and its respective protein-coding gene.

The nature and extent of overlap between a protein-coding and its respective antisense ncRNAs (Fig 3.2.9) can be useful to gain insight into the mechanism by which antisense ncRNA might regulate the overlapping protein coding gene. Therefore, the extent of overlap between antisense and protein-coding pairs belonging to the two groups was calculated. Interestingly, average base pair overlap in case of negatively correlated group (13003bp) is almost three-times as compared to the protein-coding and antisense ncRNAs pairs in the positively correlated category (4673bp). The antisense in the no correlation category, however, showed higher overlap region (6722bp) compared to the positively correlated group with the corresponding protein-coding genes. The percentage overlap with respect to the length of the protein-coding transcript was also found to be much higher for the negatively correlated group (54.1%) in comparison to the positively correlated category (31.92%). The antisense in the no correlation group (33.2%) showed percentage overlap similar to that of antisense in the positively correlated group. Thus, possibly suggesting a role of higher overlap in how these antisense might be regulating the expression of protein-coding genes.

3.2.1.6 The distance between the TSS of the protein-coding gene and the TES of overlapping antisense

To further investigate the nature of overlap between protein-coding gene and its antisense RNA, we analysed the overlap region and the distance between the TSS of mRNA and the end or TES of antisense RNA (Fig 3.2.8). The reason behind this analysis was the fact that majority of the antisense observed on the browser in the negatively correlated group showed overlap with the protein-coding gene start (Fig 3.2.4). The histograms for the overlap region and TSS-TES distances again reflected that the negatively and positively correlated groups are quite distinct (Fig 3.2.10). In general, the antisense RNA – protein-coding genes in the negatively correlated group intersected more compared to those in positively correlated group. It was also found that the distance between the TES of antisense and the TSS of mRNA gene was higher for the positively correlated category in comparison to the negatively correlated group indicating that the transcription of antisense RNAs terminated in close proximity of TSS of overlapping gene.

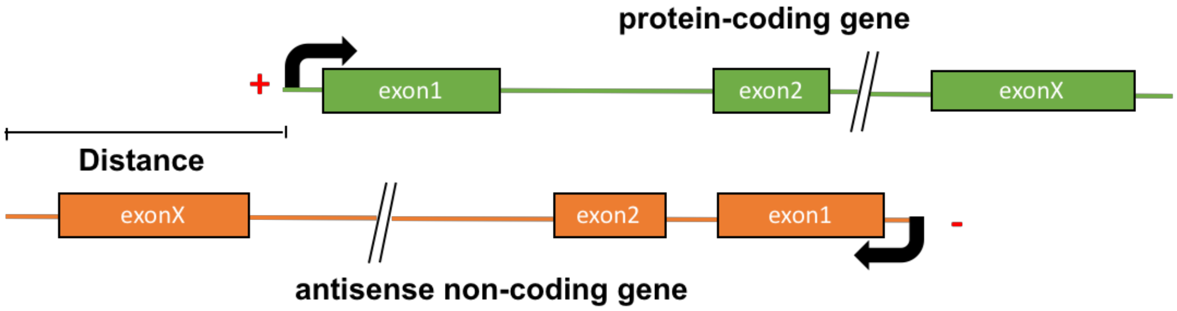


Figure 3.2.8 Schematics of distance between antisense TES and mRNA TSS
The figure above is a schematic representation of how the distance in bp was calculated between the TES of antisense (shown in orange) and the TSS of the overlapping protein-coding gene (in green). The arrows indicate the direction of transcription and the symbols in red the strand information (±).

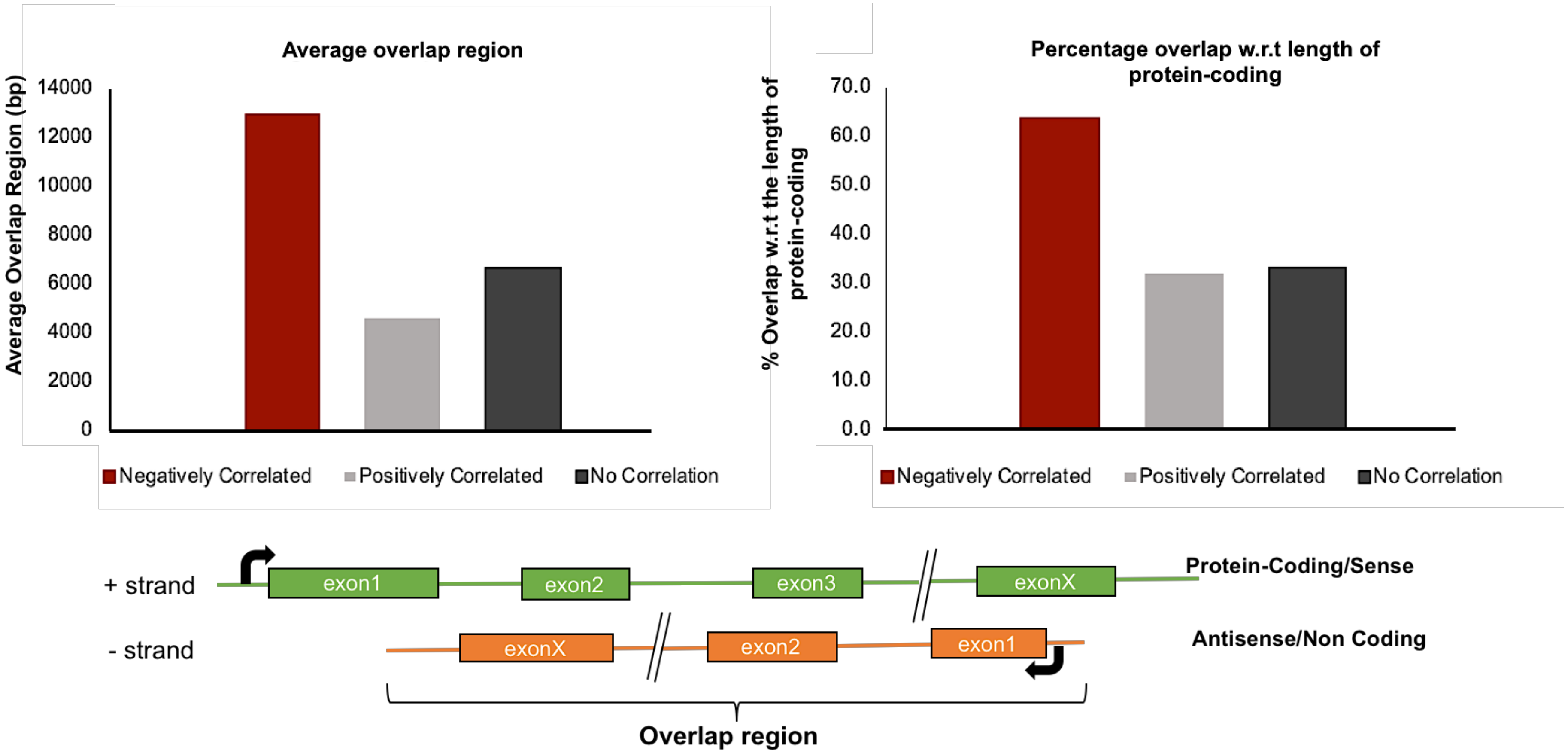


Figure 3.2.9 Average overlap region between antisense and mRNA

Average Overlap region between the antisense ncRNAs and protein-coding transcripts in negative, positive and no correlation category. The bar plot on the left shows the average overlap region in three categories which are colour coded as indicated by the labelling below each plot. The plot on the right shows the percentage overlap region with respect to the length of the protein-coding gene. The schematics below the plot explains how we calculated the overlap region.

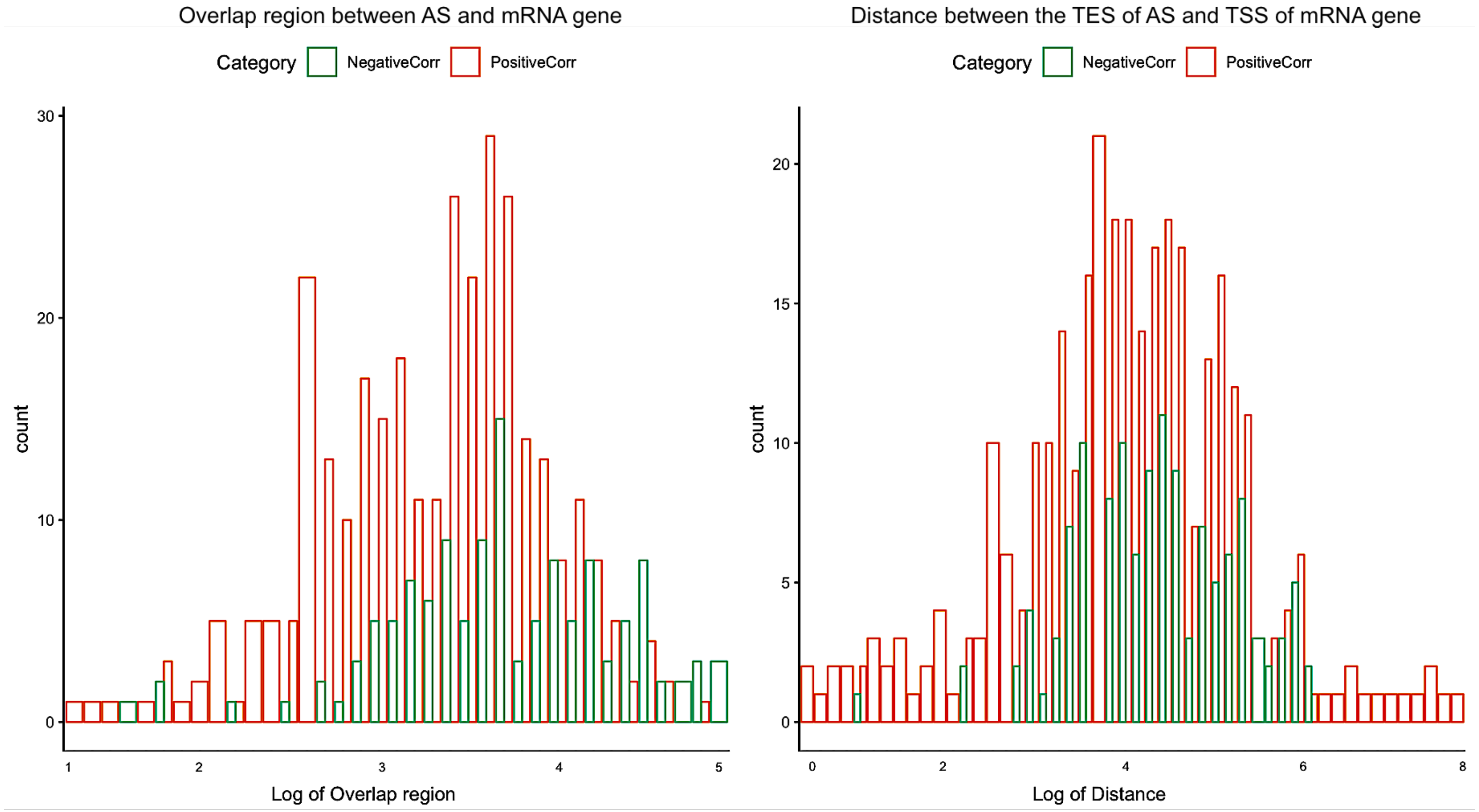


Figure 3.2.10 Histograms representing overlap region and distance

Histogram on the left shows the distribution of individual overlap region between the antisense and protein-coding gene in the negatively (green) and positively correlated group (orange). The histogram on the right shows the distance between the TES of antisense and the TSS of mRNA genes in the two categories. The X-axis represents the count or the number of times an event occurs while the Y-axis shows the log of overlap region (left) and distance (right) in bp.

3.2.1.7 Association of antisense and mRNA expression with chromatin modification revealed a plausible mechanism of regulation by antisense during development

It is well-documented that developmental genes are associated with polycomb group of proteins which are responsible for deposition of H3K27 trimethylation (H3K27me3) histone modification marks at the promoter of these genes (Aloia et al., 2013). As the protein-coding genes in the negatively correlated group were found to be involved in developmental processes, it will be interesting to study the dynamics of H3K27me3 mark at these genes. Past studies in human cell lines and mouse have also suggested the association of numerous lncRNAs associated with chromatin or repressive chromatin marks (Guttman et al., 2011, Bogu et al., 2015). So, we mined the zebrafish database for ChIP-seq (Chromatin immunoprecipitation) data profiling genome-wide enrichment of H3K27me3 mark in the dome stage of development. A heatmap showing enrichment of H3K27me3 across the TSS of antisense and overlapping mRNA in the two categories (Fig 3.2.11) showed distinct patterns of H3K27me3. Along with the H3K27me3 data we also plotted H3K4me3, associated with transcriptional initiation and H3K27ac marks that is usually enriched at transcriptionally active genes.

As expected, the protein-coding genes in negatively correlated group showed high enrichment for H3K27 trimethylation which is typically seen at developmental genes. Interestingly, antisense RNAs in the negatively correlated group showed little or no repressive marks around the TSS. On observing the H3K27ac and H3K4me3 marks, which represent activation and initiation marks at the promoter regions, we could not find any significant enrichment for the marks on both the antisense RNAs and protein-coding genes. Considering the antisense RNAs are maternally deposited (Fig 3.2.2) and the overlapping mRNAs are not expressed in early stages (Fig 3.2.2), H3K27ac and H3K4me3 marks show expected pattern. The histone modification marks, at antisense-protein-coding pairs in the positively correlated category, are distinct. The antisense as well as the protein-coding genes in the positively correlated category did not show any enrichment for the H3K27me3 repressive mark but they did show significant enrichment of activation marks H3K27ac and H3K4me3 across their TSS. This is expected given the genes in positively correlated group is enriched in house-keeping genes. This provides additional evidence that the antisense and protein-coding in the positively correlated category show both maternal and zygotic contribution.

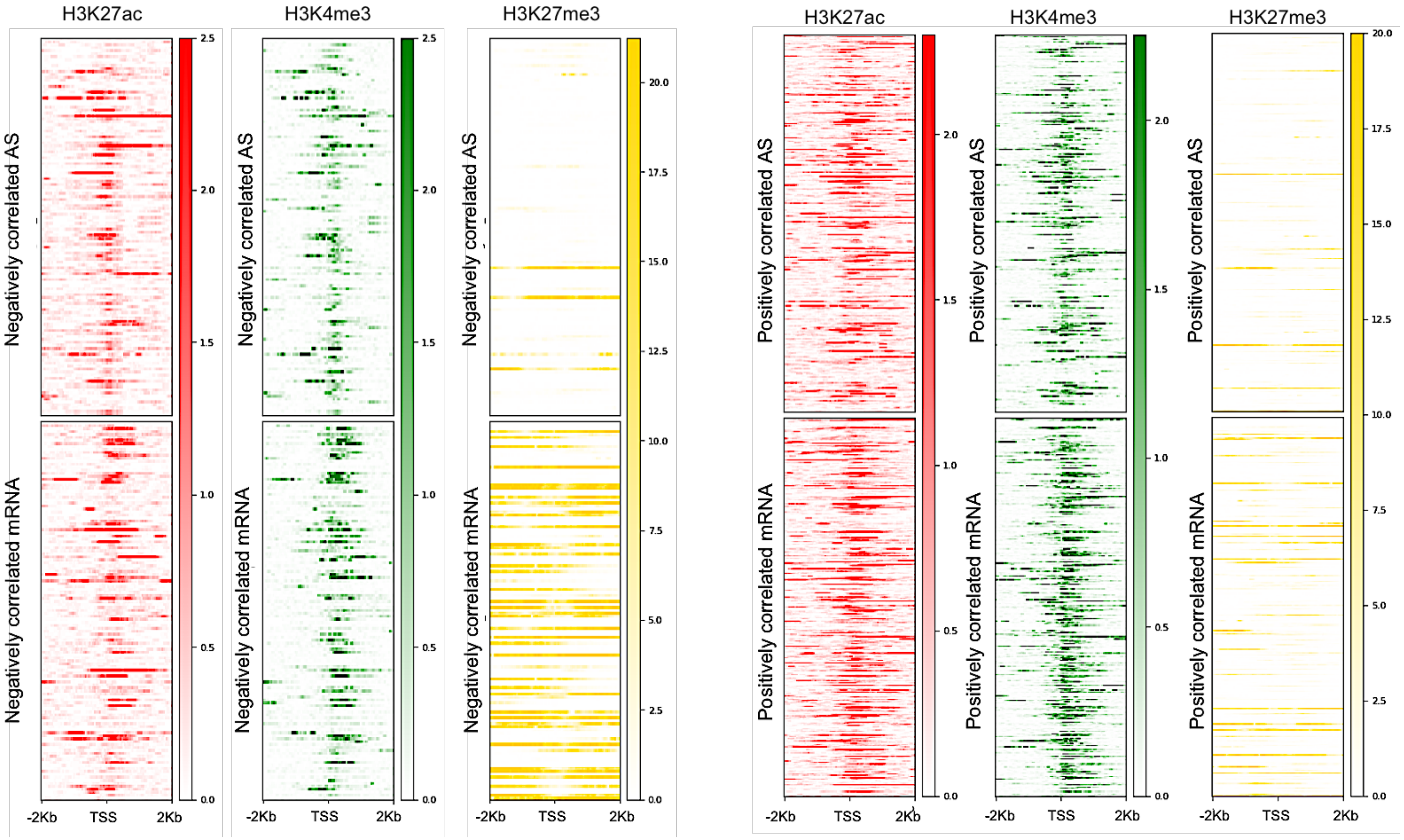


Figure 3.2.11 Chromatin modification at TSS of antisense and overlapping mRNA
Heatmap showing the distribution of ChIP marks (H3K27ac, H3K4me3 and H3K27me3) across the TSS of antisense and protein-coding gene in the negatively and positively correlated group. the H3K27ac marks are in red, H3K4me3 marks are in green and H3K27me3 marks in yellow. The X-axis represents the TSS and 2kb upstream and downstream of TSS of antisense (upper window) and protein-coding gene (lower window).

3.2.2 Identification and categorisation of antisense ncRNAs and overlapping protein-coding pairs in mouse embryonic stem cells (mESCs)

To verify whether the correlations observed between antisense ncRNAs and protein coding genes during early zebrafish development are also present in early development of other vertebrates, data from mouse which is another commonly used vertebrate model was mined. Mouse embryonic stem cells (mESCs) were chosen as transcriptomics data for stages equivalent to that used in our zebrafish (maternal-to-zygotic transition) analysis was available for this model organism (Gloss et al., 2017). As in case of Zebrafish data, we divided mESC transcripts into two groups based on the correlation between the RNA levels of antisense ncRNAs and protein-coding transcripts.

According to ENSEMBL annotations (Zerbino et al., 2018) there are 5946 antisense ncRNAs non-coding RNAs and 1,242,256 protein-coding transcripts in mouse. An intersect between the antisense and protein-coding transcripts resulted in 2772 overlapping pairs. Ten time points for every 6hrs during differentiating mESCs was chosen from an existing RNA-seq study (Gloss et al., 2017). As in case of Zebrafish data, we divided mouse transcripts into two groups based on the correlation between the RNA levels of antisense ncRNAs and protein-coding transcripts.

Based on the expression data across the ten timepoints considered, we got 206 overlapping pairs that showed either positive (67) or negative correlation (139) (Gloss et al., 2017). Among the correlated pairs 48 were significantly anti-correlated and 26 significantly positively correlated with p-value < 0.05 (Fig 3.2.12). The very low number of antisense to protein-coding pair in mESCs could be due to the fact that majority of the antisense do not express in early stages considered in our study.

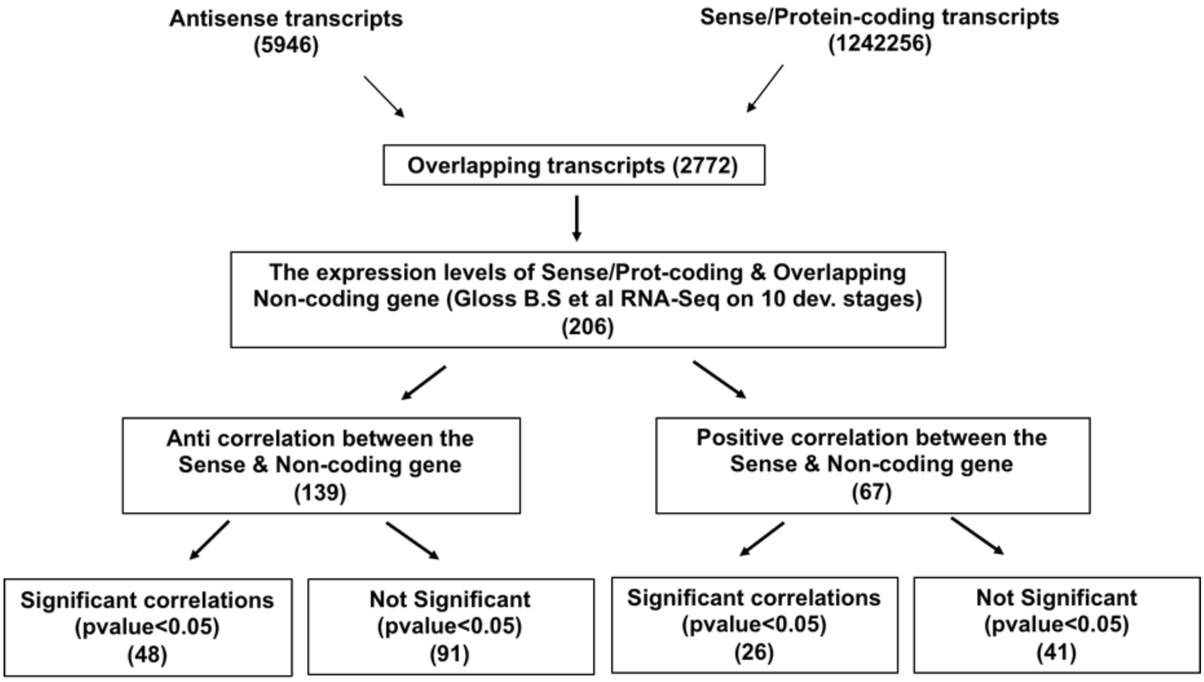

Antisense transcripts
(5946)
Sense/Protein-coding transcripts
(1242256)
Overlapping transcripts (2772)
The expression levels of Sense/Prot-coding & Overlapping Non-coding gene (Gloss B.S et al RNA-Seq on 10 dev. stages)
(206)
Anti correlation between the Sense & Non-coding gene
(139)
Positive correlation between the Sense & Non-coding gene
(67)
Significant correlations
(pvalue<0.05)
(48)
Not Significant
(pvalue<0.05)
(91)
Significant correlations
(pvalue<0.05)
(26)
Not Significant
(pvalue<0.05)
(41)

Figure 3.2.12 Pipeline used to identify and categorise antisense – mRNA pairs
Schematic representation of workflow for identifying antisense ncRNAs and overlapping protein-coding transcript pairs in mouse embryonic stem cells. we found a total of 2772 overlapping transcripts out of which only 206 were expressing in the ten stages of differentiation.

3.2.2.1 No distinct pattern of expression was observed among the antisense ncRNAs in the negatively and positively correlated group in mESCs

On plotting the average expression levels (Fig 3.2.13) of the antisense ncRNAs against the expression of their overlapping protein-coding transcripts in the two categories, no significant differences in their levels during differentiation was observed. Both the antisense ncRNAs and the protein-coding transcript in the positively and negatively correlated group seemed to be expressing throughout differentiation. The plot for the antisense and mRNAs in the negatively correlated group did not reveal any pattern compared to what we observed during zebrafish development. However, the average RNA levels of antisense, in the positively correlated category, did seem to go down by 24hrs and then again increased by 120 hours of differentiation proposing both early and late developmental expression. The average RNA levels of overlapping antisense ncRNAs and protein-coding transcripts (Fig 3.2.13) in both the negatively and positively correlated category did not show specific characteristic of being either maternal or zygotic transcripts. The absence of any specific correlation between the antisense and the overlapping protein-coding genes in any of the two categories (negatively and positively) could be due to the very low number of pairs observed.

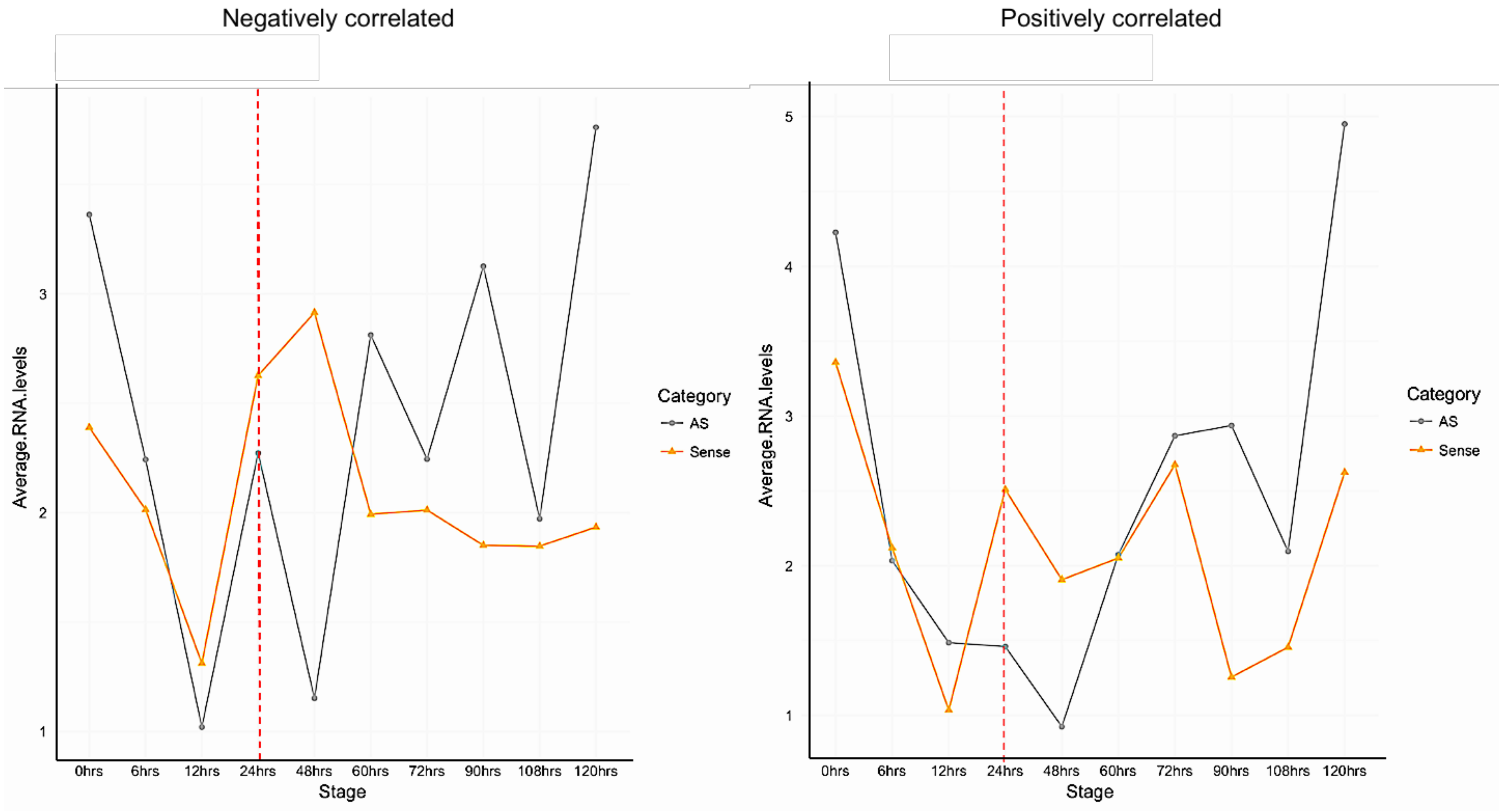
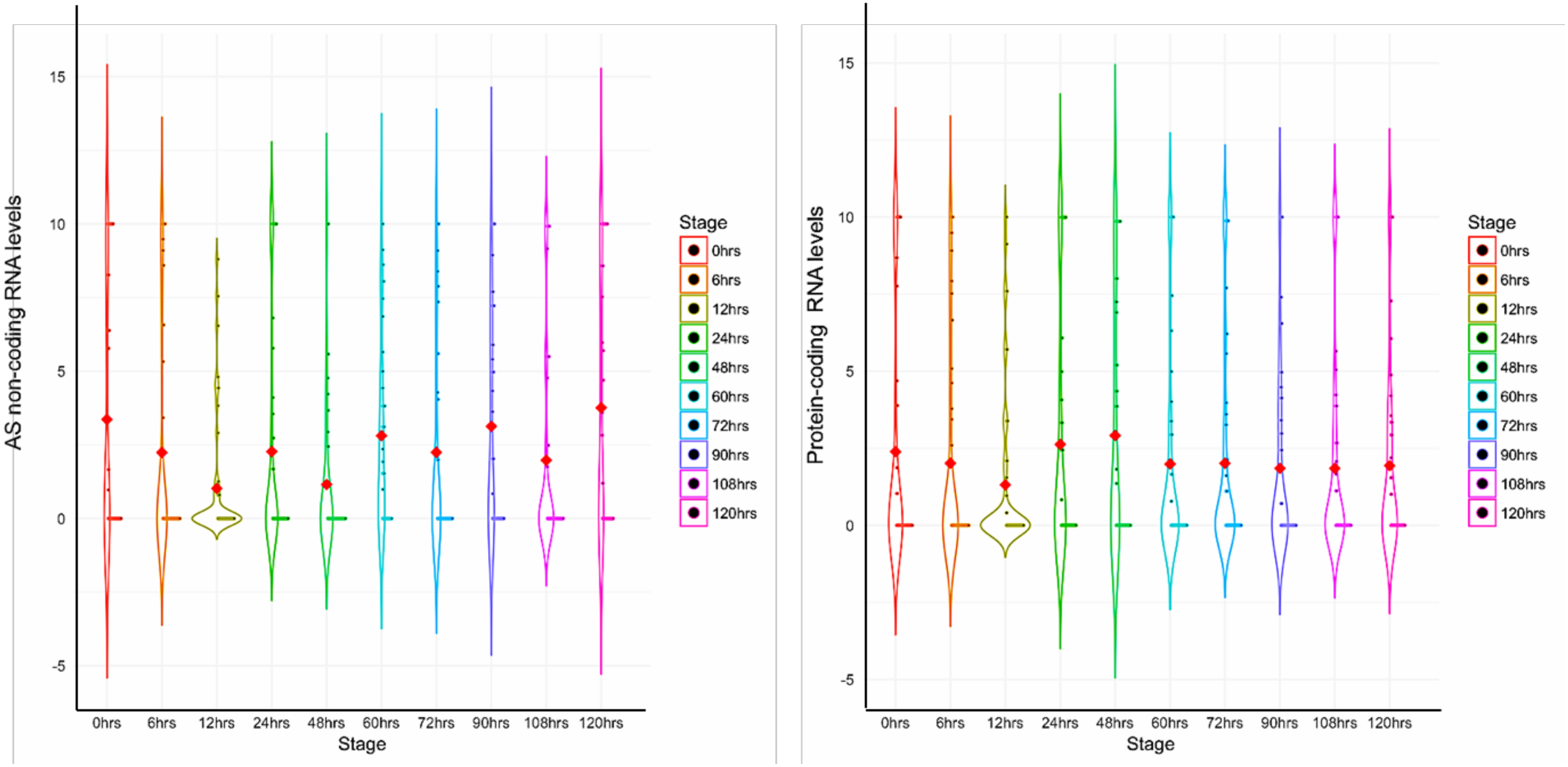


Figure 3.2.13 Average RNA levels of antisense and mRNA in the two categories
The plot shows the average RNA levels of overlapping protein-coding and antisense ncRNAs transcripts during differentiation in mESCs. The X-axis shows the timepoints in development while the Y-axis represents average RNA levels. The red dotted line at 24hrs of differentiation marks the activation of zygotic genome.

a) Negatively correlated category



b) Positively correlated group

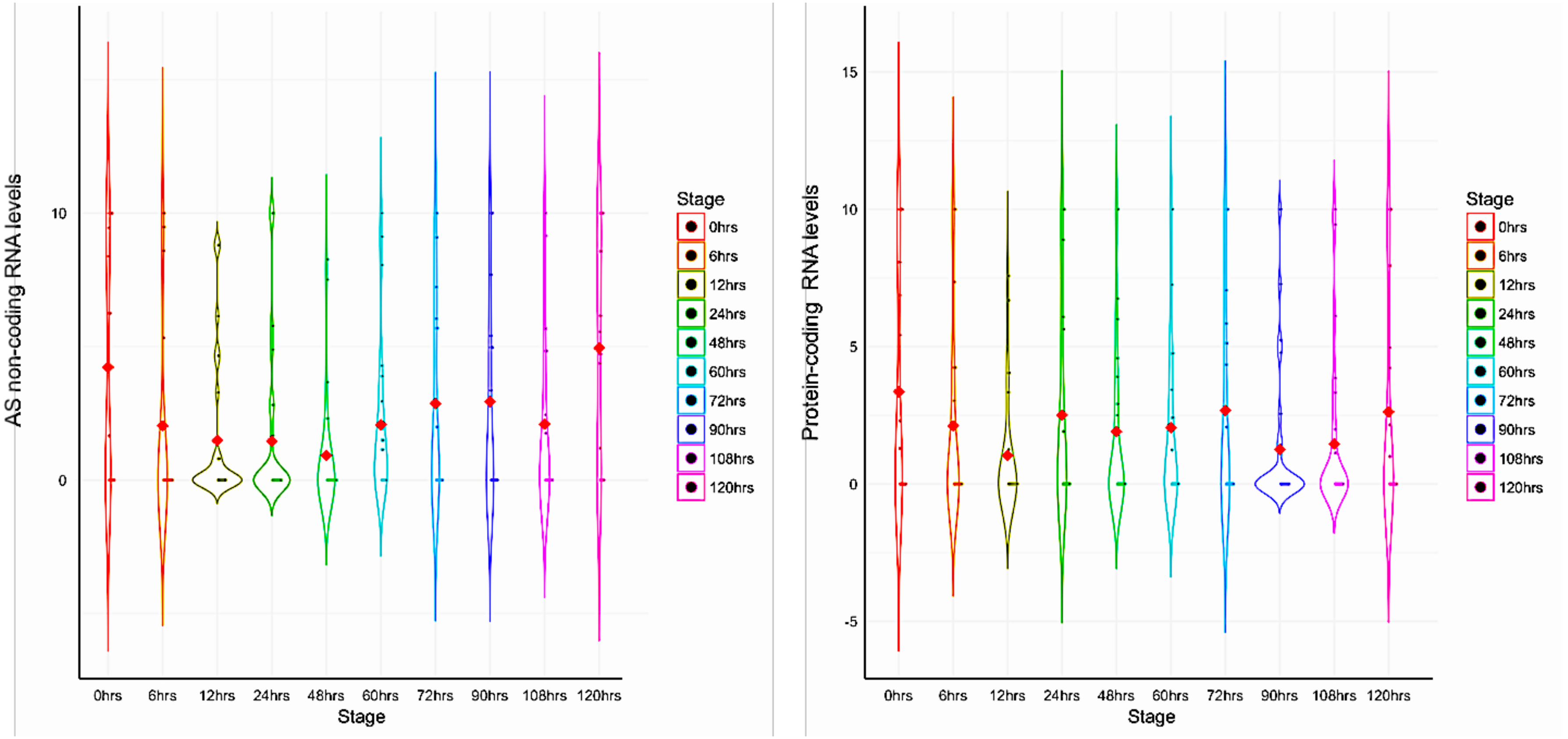


Figure 3.2.14 Individual antisense – mRNA expression plot in the two categories

Violin plot showing the RNA levels of antisense ncRNAs and overlapping protein-coding transcripts in both the a) negatively and b) positively correlated group. Y -axis represents the RNA levels of antisense ncRNAs (left) and mRNAs (right) in each category. The X-axis shows the ten timepoints considered in the study which are colour coded (colour key).

3.2.3 Conservation of antisense – protein-coding pairs from negatively correlated group in humans and mouse genome

To observe whether the characterized zebrafish antisense – protein-coding pairs from the negatively correlated category (Fig 3.2.1) are conserved, the genomic location of the overlapping developmental genes were viewed in human and mouse genomes. We looked at the developmental gene examples (Fig 3.2.7), *hoxb* gene cluster (*hoxb2*, *hoxb3*, *hoxb4*, *hoxb5* and *hoxb6*), *six1b* and the *lhx5* genes. RNA-seq dataset from Pauli et al. (Pauli et al., 2012) was used for observing expression of genes. For mouse and human the available resource on UCSC genome browser for RNA-seq (CSHL long RNA-seq from ENCODE/Cold Spring Harbor, (Affymetrix and Cold Spring Harbor Laboratory, 2009) and CAGE-seq (RNA subcellular CAGE localization from ENCODE/RIKEN, (Carninci et al., 1996) was used (Kent et al., 2002).

In zebrafish all genes in the *hoxb* (*hoxb2a, hoxb3a, hoxb5a* and *hoxb6a*) cluster shares the same antisense, *si:ch73-4e5.1*, and all the protein-coding gene expression showed anticorrelation with the antisense (Fig 3.2.14). The antisense is seen present in the early stages of zebrafish development while all the overlapping *hoxb* genes started expressing in the later stages. In mouse (Fig 3.2.14b) the expression of *hoxb2* and *hoxb3* mRNA gene also decreased with an increase in the expression of antisense (*hoxb3os*) on comparing adrenal A8 and kidney AB RNA-seq datasets for these genes. The *hoxb5* and *hoxb6* mRNA gene in mouse (Fig 3.2.15b) also showed a decrease in their expression when the antisense, *0610040B09Rik*, expression increased. In the human cell lines, however, we observed a positive correlation between the expressions of *hoxb2* and *hoxb3* and their antisense, *hoxb-AS1*, as well as in the expression of *hoxb5* and *hoxb6* genes and their antisense, *hoxb-AS3*. This was possibly due to the differences in the localisation of antisense with respect to its overlapping protein-coding. The antisense and the overlapping mRNA gene in both zebrafish and mouse genomes showed a very similar arrangement (Fig 3.2.14-3.2.15 a and b), however, in humans its different with multiple transcripts for both the antisense and mRNA. The *six1b* gene in zebrafish is completely overlapped by its antisense and their expression shows negative correlation (Fig 3.2.16a). The antisense to *six1* gene is annotated in both humans and mouse (Fig 3.2.16b-c), however, CAGE-seq and RNA-seq showed presence of read on the opposite strand suggesting the presence of an antisense. In mouse, an anticorrelation between the *six1* gene and antisense was prominent in the tissue specific cell lines but in humans it shows a positive correlation. The localisation of the antisense and *six1* gene in both mouse and zebrafish showed complete overlap unlike in humans in which they are head-to-head, divergent lncRNAs (Fig 3.2.16). The antisense to *lhx5* gene in zebrafish completely overlaps it and showed a negative correlation in their expression pattern during development (Fig 3.2.17). The mouse genome also showed presence of an antisense overlapping the protein-coding gene, *lhx5* (Fig 3.2.17b). However, in human genome, there is no annotated antisense but RNA-seq confirmed the presence of antisense on the opposite strand (Fig 3.2.17c). The antisense and the *lhx5* gene in mouse and human cell lines showed a positive correlation in their expression. This again could be due to the localisation of antisense with respect to its overlapping protein-coding gene which in human and mouse are arranged as divergent RNAs (Fig 3.2.17b and c). But in zebrafish again the antisense completely overlapped the *lhx5* gene (Fig 3.2.17a).

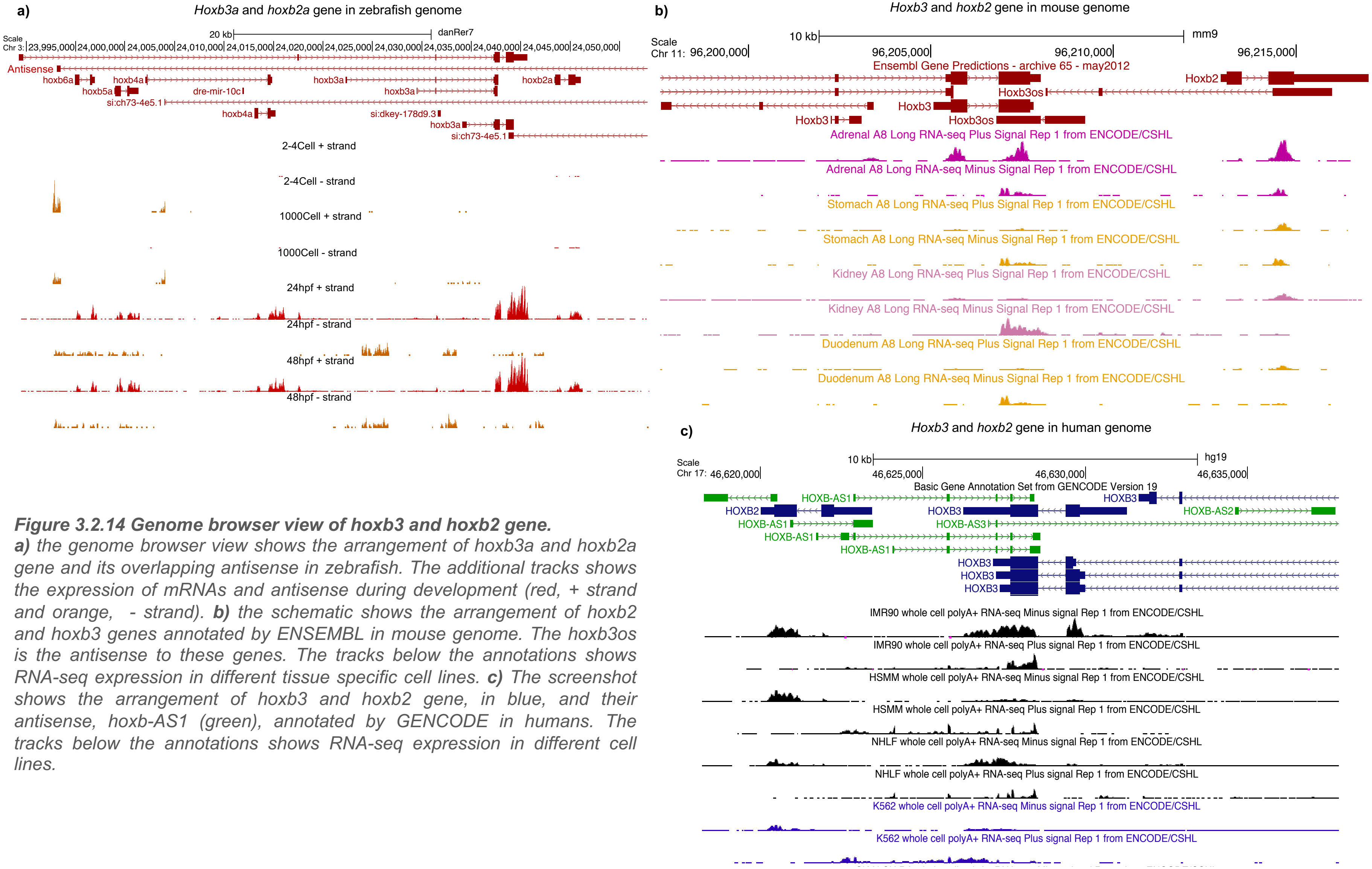
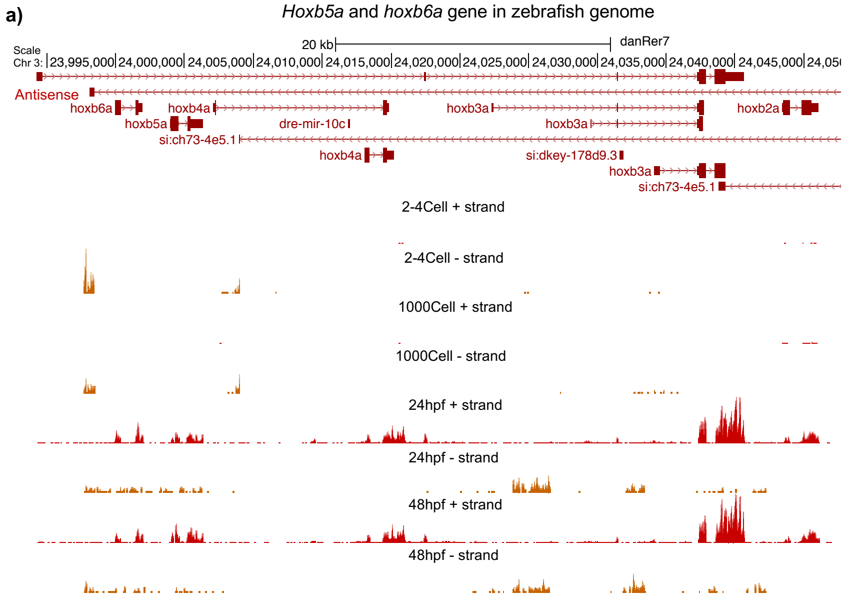

a)
Hoxb3a and *hoxb2a* gene in zebrafish genome
Scale
Chr 3:
20 kb
danRer7
23,995,000
24,000,000
24,005,000
24,010,000
24,015,000
24,020,000
24,025,000
24,030,000
24,035,000
24,040,000
24,045,000
24,050,000
Antisense
hoxb6a
hoxb5a
hoxb4a
hoxb4a
si:ch73-4e5.1
dre-mir-10c
hoxb3a
hoxb3a
si:dkey-178d9.3
hoxb3a
hoxb2a
si:ch73-4e5.1
2-4Cell + strand
2-4Cell - strand
1000Cell + strand
1000Cell - strand
24hpf + strand
24hpf - strand
48hpf + strand
48hpf - strand
b)
Hoxb3 and *hoxb2* gene in mouse genome
Scale
Chr 11:
10 kb
mm9
96,200,000
96,205,000
96,210,000
96,215,000
Ensembl Gene Predictions - archive 65 - may2012
Hoxb2
Hoxb3os
Hoxb3
Hoxb3
Hoxb3os
Adrenal A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Adrenal A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
Stomach A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Stomach A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
Kidney A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Kidney A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
Duodenum A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Duodenum A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
c)
Hoxb3 and *hoxb2* gene in human genome
Scale
Chr 17:
10 kb
hg19
46,620,000
46,625,000
46,630,000
46,635,000
Basic Gene Annotation Set from GENCODE Version 19
HOXB-AS1
HOXB3
HOXB2
HOXB3
HOXB-AS2
HOXB-AS1
HOXB-AS3
HOXB-AS1
HOXB-AS1
HOXB3
HOXB3
HOXB3
IMR90 whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
IMR90 whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
HSMM whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
HSMM whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
NHLF whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
NHLF whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
K562 whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
K562 whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL

Figure 3.2.14 Genome browser view of hoxb3 and hoxb2 gene.
a) *the genome browser view shows the arrangement of hoxb3a and hoxb2a gene and its overlapping antisense in zebrafish. The additional tracks shows the expression of mRNAs and antisense during development (red, + strand and orange, - strand).* ***b)*** *the schematic shows the arrangement of hoxb2 and hoxb3 genes annotated by ENSEMBL in mouse genome. The hoxb3os is the antisense to these genes. The tracks below the annotations shows RNA-seq expression in different tissue specific cell lines.* ***c)*** *The screenshot shows the arrangement of hoxb3 and hoxb2 gene, in blue, and their antisense, hoxb-AS1 (green), annotated by GENCODE in humans. The tracks below the annotations shows RNA-seq expression in different cell lines.*



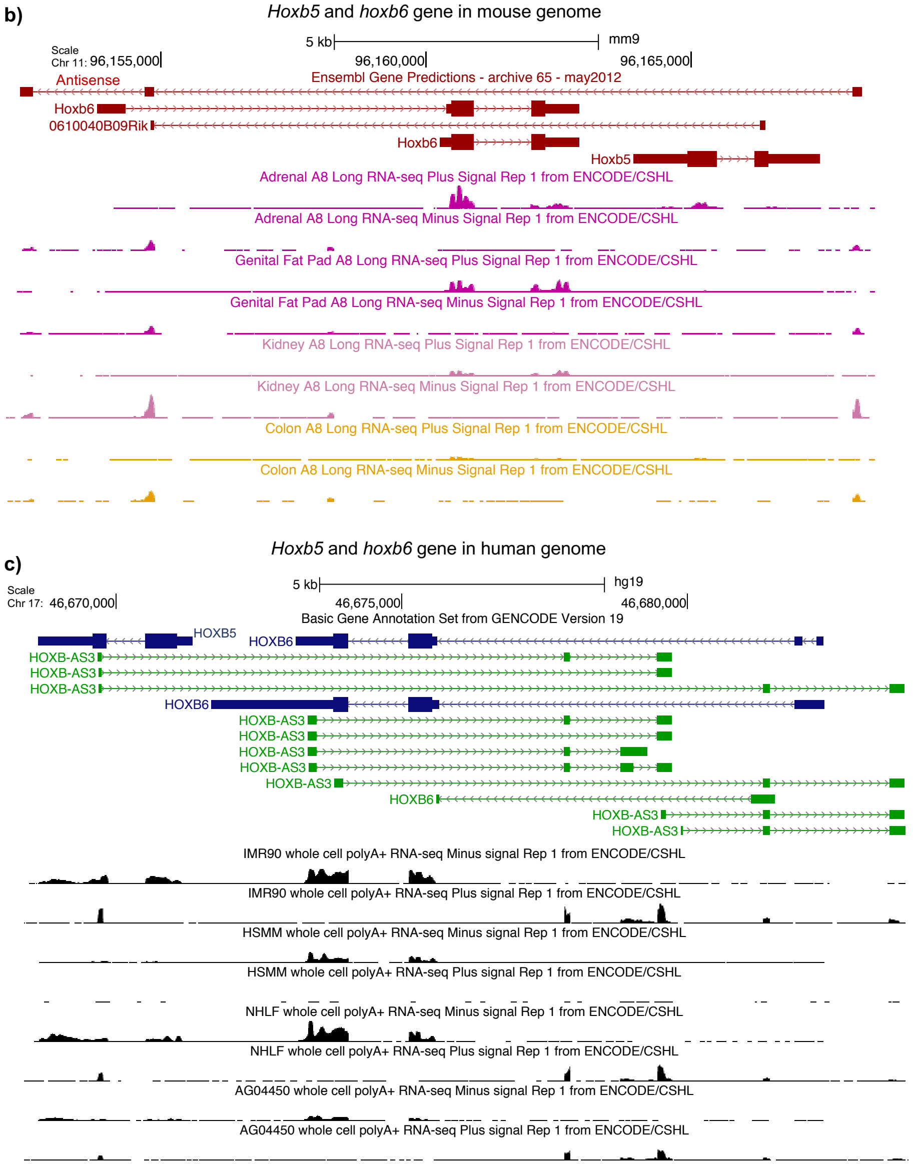

b)
Hoxb5 and hoxb6 gene in mouse genome
Scale
Chr 11:
5 kb
mm9
96,155,000
96,160,000
96,165,000
Antisense
Ensembl Gene Predictions - archive 65 - may2012
Hoxb6
0610040B09Rik
Hoxb6
Hoxb5
Adrenal A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Adrenal A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
Genital Fat Pad A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Genital Fat Pad A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
Kidney A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Kidney A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
Colon A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Colon A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
c)
Hoxb5 and hoxb6 gene in human genome
Scale
Chr 17:
5 kb
hg19
46,670,000
46,675,000
46,680,000
Basic Gene Annotation Set from GENCODE Version 19
HOXB5
HOXB6
HOXB-AS3
HOXB-AS3
HOXB-AS3
HOXB6
HOXB-AS3
HOXB-AS3
HOXB-AS3
HOXB-AS3
HOXB-AS3
HOXB6
HOXB-AS3
HOXB-AS3
IMR90 whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
IMR90 whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
HSMM whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
HSMM whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
NHLF whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
NHLF whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
AG04450 whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
AG04450 whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL

Figure 3.2.15 Genome browser view of hoxb5 and hoxb6 gene.
__a)__ the genome browser view shows the arrangement of hoxb5a and hoxb6a gene and overlapping antisense in zebrafish. The additional tracks shows the expression of mRNAs and antisense during development (red, + strand and orange, - strand). __b)__ the schematic shows the arrangement of hoxb5 and hoxb6 genes and their antisense annotated by ENSEMBL in mouse genome. The tracks below the annotations shows RNA-seq expression in different tissue specific cell lines. __c)__ the screenshot shows the arrangement of hoxb5 and hoxb6 gene in blue, and their antisense, hoxb-AS3 (in green), annotated by GENCODE in humans. The tracks below the annotations shows RNA-seq expression in different cell lines.

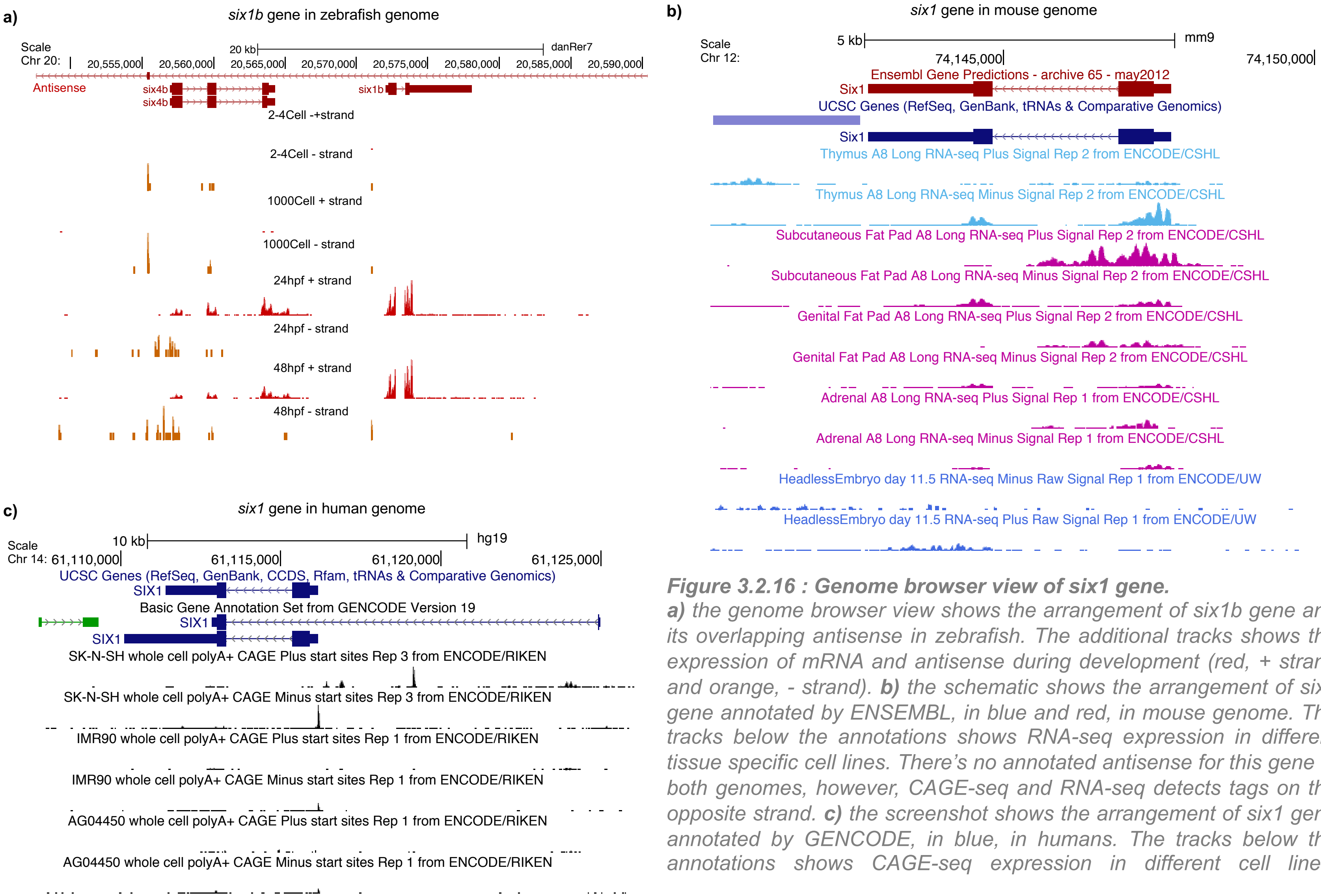

a)
six1b gene in zebrafish genome
Scale
Chr 20:
20 kb
danRer7
20,555,000
20,560,000
20,565,000
20,570,000
20,575,000
20,580,000
20,585,000
20,590,000
Antisense
six4b
six4b
six1b
2-4Cell -+strand
2-4Cell - strand
1000Cell + strand
1000Cell - strand
24hpf + strand
24hpf - strand
48hpf + strand
48hpf - strand
b)
six1 gene in mouse genome
Scale
Chr 12:
5 kb
mm9
74,145,000
74,150,000
Ensembl Gene Predictions - archive 65 - may2012
Six1
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Six1
Thymus A8 Long RNA-seq Plus Signal Rep 2 from ENCODE/CSHL
Thymus A8 Long RNA-seq Minus Signal Rep 2 from ENCODE/CSHL
Subcutaneous Fat Pad A8 Long RNA-seq Plus Signal Rep 2 from ENCODE/CSHL
Subcutaneous Fat Pad A8 Long RNA-seq Minus Signal Rep 2 from ENCODE/CSHL
Genital Fat Pad A8 Long RNA-seq Plus Signal Rep 2 from ENCODE/CSHL
Genital Fat Pad A8 Long RNA-seq Minus Signal Rep 2 from ENCODE/CSHL
Adrenal A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Adrenal A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
HeadlessEmbryo day 11.5 RNA-seq Minus Raw Signal Rep 1 from ENCODE/UW
HeadlessEmbryo day 11.5 RNA-seq Plus Raw Signal Rep 1 from ENCODE/UW
c)
six1 gene in human genome
Scale
Chr 14:
10 kb
hg19
61,110,000
61,115,000
61,120,000
61,125,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
SIX1
Basic Gene Annotation Set from GENCODE Version 19
SIX1
SIX1
SK-N-SH whole cell polyA+ CAGE Plus start sites Rep 3 from ENCODE/RIKEN
SK-N-SH whole cell polyA+ CAGE Minus start sites Rep 3 from ENCODE/RIKEN
IMR90 whole cell polyA+ CAGE Plus start sites Rep 1 from ENCODE/RIKEN
IMR90 whole cell polyA+ CAGE Minus start sites Rep 1 from ENCODE/RIKEN
AG04450 whole cell polyA+ CAGE Plus start sites Rep 1 from ENCODE/RIKEN
AG04450 whole cell polyA+ CAGE Minus start sites Rep 1 from ENCODE/RIKEN

Figure 3.2.16 : Genome browser view of six1 gene.
__a)__ the genome browser view shows the arrangement of six1b gene and its overlapping antisense in zebrafish. The additional tracks shows the expression of mRNA and antisense during development (red, + strand and orange, - strand). __b)__ the schematic shows the arrangement of six1 gene annotated by ENSEMBL, in blue and red, in mouse genome. The tracks below the annotations shows RNA-seq expression in different tissue specific cell lines. There's no annotated antisense for this gene in both genomes, however, CAGE-seq and RNA-seq detects tags on the opposite strand. __c)__ the screenshot shows the arrangement of six1 gene annotated by GENCODE, in blue, in humans. The tracks below the annotations shows CAGE-seq expression in different cell lines.

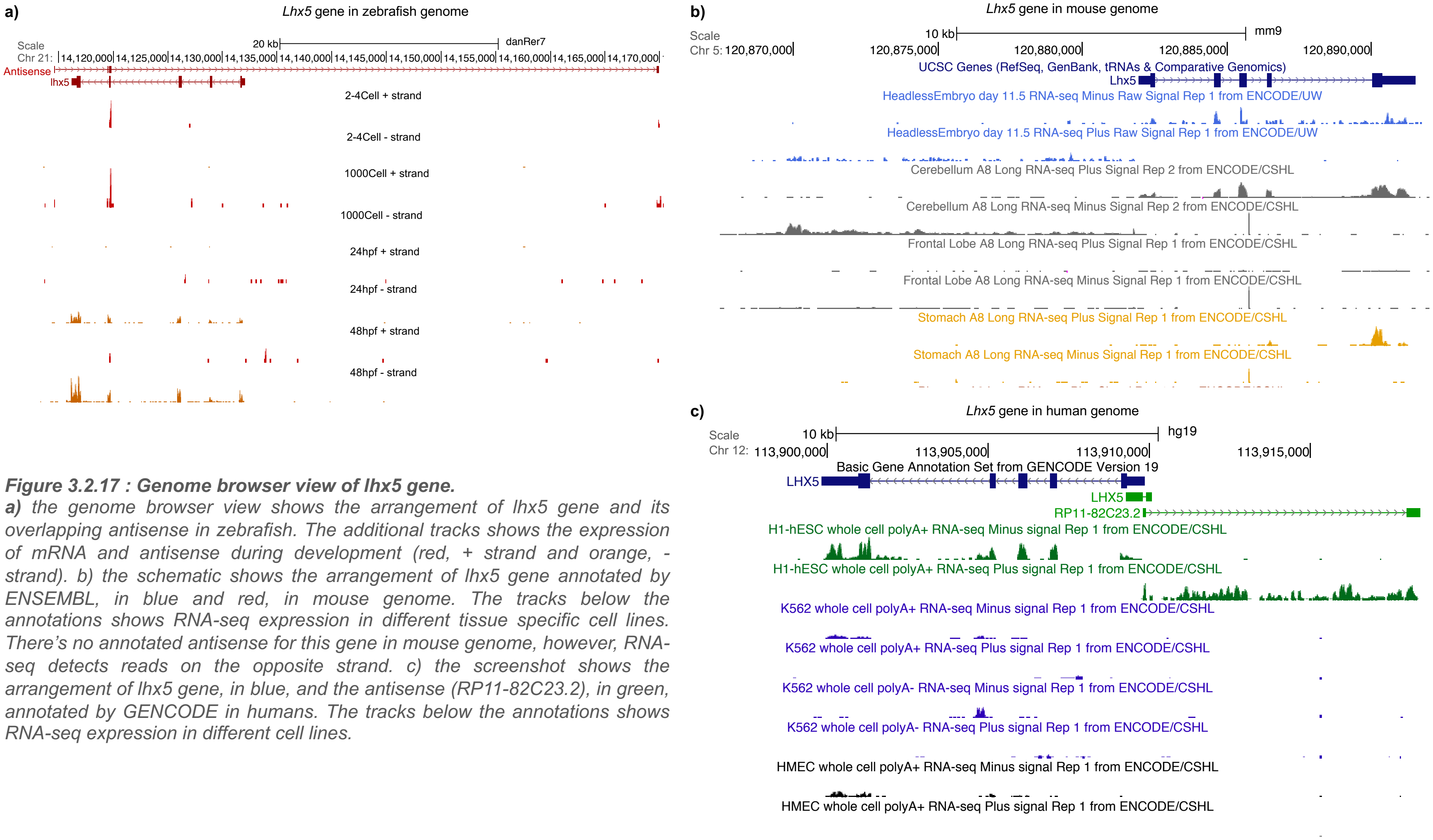

a)
Lhx5 gene in zebrafish genome
Scale
20 kb
danRer7
Chr 21: 14,120,000 14,125,000 14,130,000 14,135,000 14,140,000 14,145,000 14,150,000 14,155,000 14,160,000 14,165,000 14,170,000
Antisense
lhx5
2-4Cell + strand
2-4Cell - strand
1000Cell + strand
1000Cell - strand
24hpf + strand
24hpf - strand
48hpf + strand
48hpf - strand
b)
Lhx5 gene in mouse genome
Scale
10 kb
mm9
Chr 5: 120,870,000 120,875,000 120,880,000 120,885,000 120,890,000
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Lhx5
HeadlessEmbryo day 11.5 RNA-seq Minus Raw Signal Rep 1 from ENCODE/UW
HeadlessEmbryo day 11.5 RNA-seq Plus Raw Signal Rep 1 from ENCODE/UW
Cerebellum A8 Long RNA-seq Plus Signal Rep 2 from ENCODE/CSHL
Cerebellum A8 Long RNA-seq Minus Signal Rep 2 from ENCODE/CSHL
Frontal Lobe A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Frontal Lobe A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
Stomach A8 Long RNA-seq Plus Signal Rep 1 from ENCODE/CSHL
Stomach A8 Long RNA-seq Minus Signal Rep 1 from ENCODE/CSHL
c)
Lhx5 gene in human genome
Scale
10 kb
hg19
Chr 12: 113,900,000 113,905,000 113,910,000 113,915,000
Basic Gene Annotation Set from GENCODE Version 19
LHX5
LHX5
RP11-82C23.2
H1-hESC whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
H1-hESC whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
K562 whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
K562 whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL
K562 whole cell polyA- RNA-seq Minus signal Rep 1 from ENCODE/CSHL
K562 whole cell polyA- RNA-seq Plus signal Rep 1 from ENCODE/CSHL
HMEC whole cell polyA+ RNA-seq Minus signal Rep 1 from ENCODE/CSHL
HMEC whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL

Figure 3.2.17 : Genome browser view of lhx5 gene.
a) *the genome browser view shows the arrangement of lhx5 gene and its overlapping antisense in zebrafish. The additional tracks shows the expression of mRNA and antisense during development (red, + strand and orange, - strand). b) the schematic shows the arrangement of lhx5 gene annotated by ENSEMBL, in blue and red, in mouse genome. The tracks below the annotations shows RNA-seq expression in different tissue specific cell lines. There's no annotated antisense for this gene in mouse genome, however, RNA-seq detects reads on the opposite strand. c) the screenshot shows the arrangement of lhx5 gene, in blue, and the antisense (RP11-82C23.2), in green, annotated by GENCODE in humans. The tracks below the annotations shows RNA-seq expression in different cell lines.*

3.3 Discussion

In zebrafish so far, there are a few antisense which have been identified during embryogenesis. The *tie-1* gene which is essential in vascular development in zebrafish embryos has been shown to be downregulated by its antisense *tie-1* AS (Li et al., 2010). In this chapter, the aim was to characterise antisense ncRNAs expressed during early stages of zebrafish embryogenesis and their relation to the overlapping protein-coding partner which is transcribed from opposite strand. It has been shown in human and mouse ESCs that lncRNAs have a significantly higher expression correlation with their closest gene (less than 5kb) compared to protein-coding genes (Luo et al., 2016). We characterised antisense ncRNAs into different groups based on whether they correlated positively or negatively with the expression of their overlapping protein-coding partners. A very clear pattern of anti-correlation and positive correlation was observed for all stages (Fig 3.2.1). In doing so, we determined that the antisense ncRNAs in the negatively correlated group showed characteristics of maternal transcripts with high RNA levels during the early stages (0.75hpf) and degradation around 10hpf stage with little or no zygotic contribution. This was in accordance with what has been predicted for the lncRNAs identified in the early stages of zebrafish development (Pauli et al., 2012). The study proposed either a maternal (majority of RNAs) or a paternal contribution (sperm provided RNAs) for these lncRNAs (Lalancette et al., 2008). The antisense ncRNAs belonging to the positively correlated category, on the other hand, showed characteristics of housekeeping genes or genes involved in organogenesis with both maternal and zygotic contribution. Thus, for the first-time we have provided evidence for the presence of two different class of antisense during zebrafish embryogenesis. We also established the importance of the protein-coding genes that these antisense ncRNAs overlapped in the two groups with the negatively correlated mRNA genes being enriched for developmental genes and the positively correlated mRNA genes in metabolism.

In some cases, antisense downregulate their overlapping protein-coding gene function by forming an RNA-RNA hybridisation with the mRNA thus preventing its translation (Faghihi and Wahlestedt, 2009). A similar mechanism can be explained by the higher overlap region observed between the antisense genes and their protein-coding partner (Fig 3.2.9). The proximity of the antisense TES from the overlapping mRNA TSS or convergent transcription, in negatively correlated group (Fig 3.2.10), might be a possible mechanism of regulation of overlapping mRNA genes by antisense (Shearwin et al., 2005).

We also established that the developmental genes observed in the negatively correlated group were associated with H3K27 trimethylation marks or PcG complexes in agreement with another study (Aloia et al., 2013). The antisense in this category did not show enrichment for neither H3K4me3 nor H3K27ac hinting towards their maternal inheritance. We also proved that the mRNA genes in the positively correlated category are housekeeping genes as they showed enrichment for both initiation (H3K4me3) and activation marks (H3K27ac) but not for H3K27me3. A past study has shown that majority of these chromatin marks are established during or after zygotic genome activation (Vastenhouw et al., 2010). Therefore, we speculate that after the zygotic genome activation which establishes the H3K27me3 marks onto the protein-coding genes, their expression is regulated by the PRC2 complex. Before zygotic genome activation, however, the maternally deposited antisense ncRNAs probably through their complementarity with the overlapping mRNA prevents its translation. This is reflected by the higher overlap region between the mRNAs and their antisense (Fig 3.2.9). We propose a post-transcriptional mechanism of regulation by the antisense in the negatively correlated group which is a backup process to prevent the developmental genes from expressing before their absolute requirement. The antisense in the positively correlated category, in contrast, may regulate the overlapping mRNA genes by association with activating chromatin modifying complexes (Fig 3.2.11) (Orom et al., 2010b, Orom et al., 2010a). To further prove this mechanism of transcription regulation we checked the localisation of these antisense ncRNAs transcripts with respect to their corresponding protein-coding transcripts, which will be the objective of our next set of experiments (CHAPTER 4).

Our study also included further analysis whether antisense ncRNAs and protein-coding pairs show similar correlations during early development in other vertebrates. The average RNA levels of overlapping antisense ncRNAs and protein-coding transcripts (Fig 3.2.13) in both the negatively and positively correlated category did not show specific characteristics of maternal and zygotic transcripts. This might be due to the difference in the timing of maternal-to-zygotic transition which in zebrafish occurs at 3hpf and is accompanied by rapid cell divisions. While in mouse, it occurs at 24hpf, with comparatively slower cell divisions and longer developmental time (21 dpf) compared to zebrafish (3 dpf) (Tadros and Lipshitz, 2009). Therefore, it is possible that majority of the overlapping transcripts (2772) that we miss in our analysis are expressed later in development. Majority of the antisense and overlapping developmental genes characterised in our negatively correlated category showed conservation in both mouse and human genome. This has been previously observed in *Fugu rubripes* (Woolfe et al., 2005) and mouse (Ponjavic et al., 2009) that identified lncRNAs to be conserved at syntenic level as well as colocalised with transcription factor genes during vertebrate development. The *hoxb* genes, for example, showed comparable arrangement in both zebrafish and mouse with a similar expression correlation (negative) between the antisense and overlapping protein-coding gene. However, in human genome all of the antisense observed showed a positive correlation with the expression of the mRNA gene. Previous correlation analysis between the expression of lncRNAs antisense to the coding genes showed positive correlation in tissue specific human cell lines (Derrien et al., 2012). The arrangement or genomic localisation of the antisense in humans with the protein-coding gene was different from zebrafish and mouse for majority of the genes observed. The results suggested that the positioning of antisense and the protein-coding gene plays an important role in how the antisense might be regulating the expression of overlapping mRNA gene. A study, in mouse and human ESCs, revealed divergent lncRNAs to be associated with important developmental genes (Luo et al., 2016). Studies on yeast (Qi and Arkin, 2014) and bacteria (Xu et al., 2011) has also revealed similar positioning of antisense with specific genes that needs to be activated owing to environmental changes such as stress response genes.

In summary zebrafish development showed presence of several maternal antisense ncRNAs that might be involved in the regulation of overlapping protein-coding gene expression. These maternal antisense also showed conservation in both human and mouse genomes. The analysis also revealed the importance of antisense positioning with respect to the protein-coding gene.

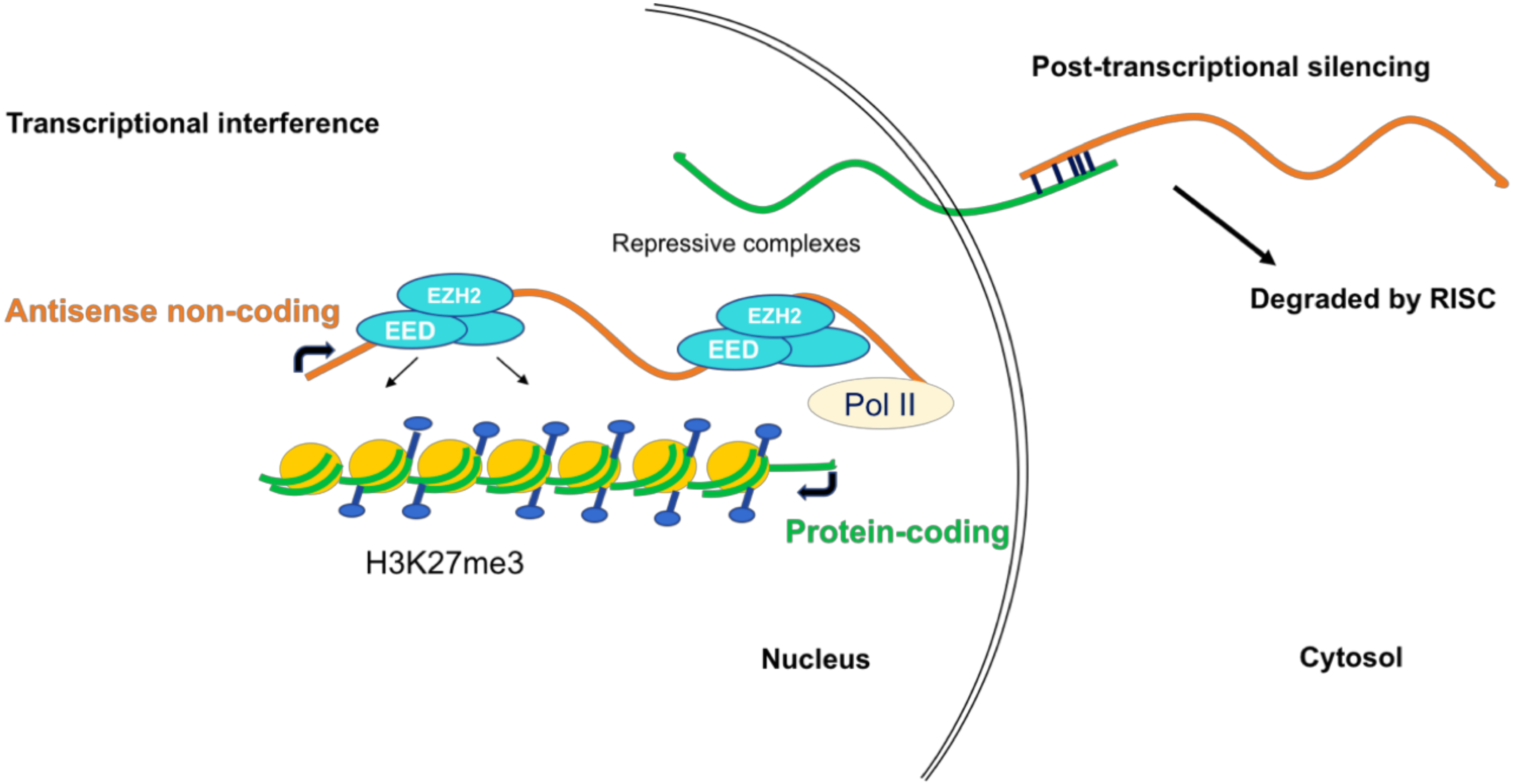


Figure 3.3.1 Mechanism of regulation of protein-coding genes by antisense ncRNAs.
The figure shows a hypothesis for the possible mechanism of mRNA transcription regulation by the overlapping antisense ncRNA. Two mechanisms speculated is either transcriptional interference, where the antisense modulates transcription of mRNA by recruiting PRC2 or repressive complex, or post-transcriptional silencing of mRNA gene expression by antisense by forming double strand RNA-RNA hybridisation.

CHAPTER 4: SUBCELLULAR LOCALISATION OF RNAS WITH MATERNAL-TO- ZYGOTIC TRANSITION DURING ZEBRAFISH DEVELOPMENT

4.1 Introduction

During early stages of vertebrate development, zygotic genome is inactive and initial functioning of embryo takes place with the help of maternally deposited RNAs as well as proteins. Zygotic genome activation, which results in maternal-to-zygotic transition (MZT) of transcriptome, is a crucial step responsible for embryonic development. There is much that we still do not understand about MZT and one such question is the cellular localisation of both types of RNAs, maternally provided and zygotically expressed, prior to MZT and post-MZT respectively. The subcellular localisation will be important in understanding genomic function of these RNAs and will contribute in gaining insight into early developmental events. Similar studies such as (Djebali et al., 2012, Derrien et al., 2012, Diermeier and Langst, 2014, Mitchell et al., 2012, Dhaliwal and Mitchell, 2016, Bai and Laiho, 2016) for example have revealed interesting dynamics of RNAs. The GENCODE study (Derrien et al., 2012) in human cell lines revealed that when compared to the coding RNAs, non-coding RNAs are localised in the nucleus relative to cytosol. The levels of polyA+ and polyA- RNAs varied in their subcellular localisation. It has been observed that small non-coding RNAs show highly specific localisation, with miRNAs and tRNAs being abundant in the cytosolic fraction and snoRNAs being predominantly present in the nuclear fraction (Djebali et al., 2012). However, snRNAs were distributed equally in both the compartments and on further examination were highly abundant in chromatin-associated fraction possibly due to their role in splicing (co-transcriptional mechanism).

RNA dynamics and localisation during early vertebrate development has not yet been studied. In this chapter, we sought to identify the subcellular localisation of RNAs. Total RNA-Seq (both polyA+ and polyA-) on these fractions can not only provide information on the localisation of developmental mRNAs, but also several processed and unprocessed non-coding RNAs. The fractionation of RNAs will also allow detection of lowly expressed RNAs which cannot be detected in whole-cell RNA-seq due to low transcript levels and being restricted to the nucleus.

In this chapter we analyse the localisation and expression of annotated ENSEMBL (Zerbino et al., 2018) mRNA and non-coding RNAs in either the nuclear or cytosolic compartment during zebrafish development. Our RNA-seq data not just provides a resource of RNA landscape but also unravels important regulatory mechanism of ncRNAs during vertebrate development.

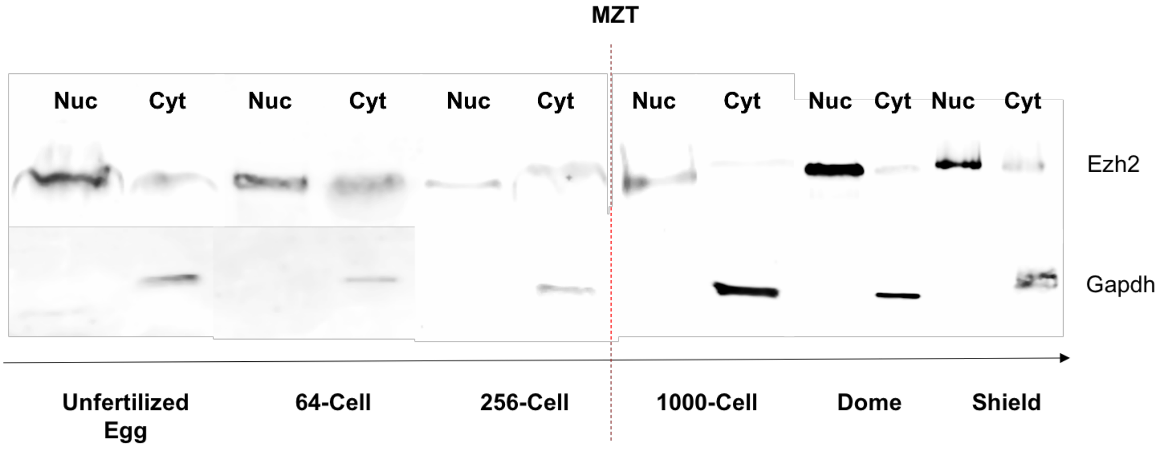
4.2 Results

4.2.1 Nuclear and Cytosolic fractionation of zebrafish embryos

In order to understand compartmentalization of RNAs during early development, we collected cytosolic and nuclear fractions from 8 different stages of zebrafish embryonic development, which included the 64-Cell (2hpf), 128-Cell (2.25hpf), 256-Cell (2.5hpf), 512-Cell (2.75hpf), 1000-Cell (3hpf), High (3.3hpf), Dome (4.3hpf) and Shield (6hpf) stages. The fractions were validated by checking for the presence of known nuclear and cytosolic protein markers in humans. For cytosolic fraction *Gapdh* and for nuclear fraction *Ezh2* was used as a marker (San et al., 2016, Thisse et al., 2004). In human cell lines, these proteins are highly localised in corresponding compartments. In the collected embryonic fractions, as expected GAPDH was only present in the cytosolic fraction indicating that cytosolic fraction is highly enriched and its absence from nuclear fraction indicated that nuclear fraction is free from any cytosolic contamination (Fig 4.2.1a). The nuclear marker EZH2, a chromatin modifying protein (subunit of the polycomb repressive complex II associated with chromatin) is much more enriched in the nuclear fraction (Fig 4.2.1a) indicating that nuclear fractions are enriched for nuclear proteins and RNAs.

Although zygotic transcription begins only after MZT, we do not know much about RNA Pol II expression levels and localisation in these early stages. After confirming the quality of fractions, we tested the expression levels and localisation of RNA Pol II (Fig 4.2.1b). We found that RNA Pol II was absent or present at low levels before MZT (2hpf and 2.5hpf). However, following zygotic genome activation (4.3hpf and 6hpf) its levels increased and as expected there was higher enrichment of RNA Pol II in the nuclear fractions when compared to the cytosolic fractions (Fig 4.2.1b).

a)



b)

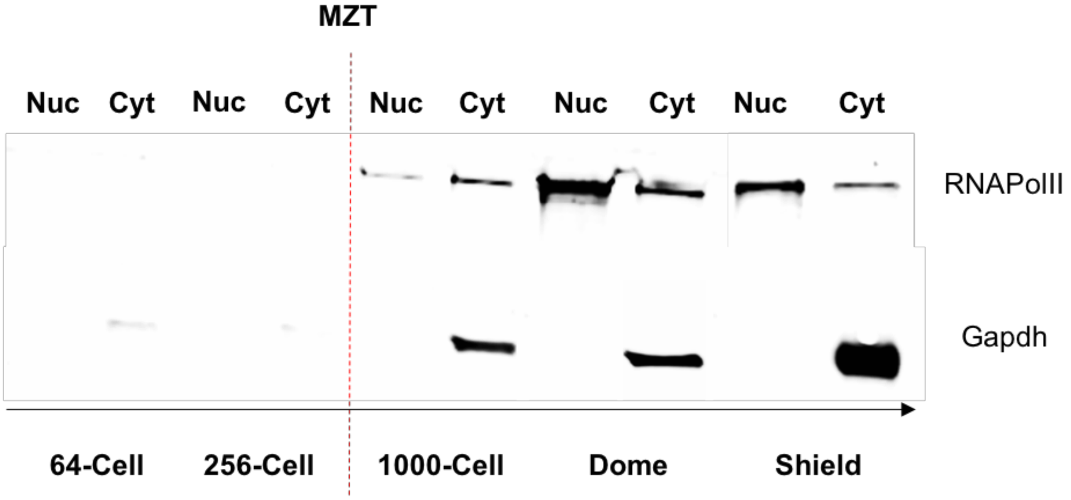


Figure 4.2.1 Expression analysis of Ezh2, gapdh and RNA Pol II proteins in the fractions.

Western blots showing expression of markers in Nuclear and Cytosolic fractions of zebrafish embryonic cells. a) Expression of Ezh2 and Gapdh protein in the fractions from five (64-Cell, 256-Cell, 1000-Cell, Dome and Shield) stages of development and unfertilised egg. b) Expression of RNA Pol II protein in the fractions from five stages of development along with the loading control (Gapdh).

4.2.2 Further validation of nuclear and cytosolic fractions using RNA markers

Both cytosolic and nuclear fractions at different stages of zebrafish development were further validated by checking localisation of nuclear and cytosolic RNA markers by PCR. Since there is no documented nuclear or cytosolic RNA markers for early stages of development, we relied on detection of other RNA features. For example, because splicing takes place in nucleus, presence of unspliced RNA would serve as a good indicator of nuclear fraction. We used *hnrnpa1* mRNA to test this hypothesis. *hnrnpa1* is an important gene for alternative splicing of pre-mRNAs (Despic et al., 2017) and is, therefore, expected to be present ubiquitously throughout development. Hence it will be anticipated that it is amongst one of the first to be synthesised during zygotic genome activation. The PCRs for *hnrnpa1* in the nuclear fractions from different stages suggested the presence of an additional higher band (~280bp) along with the predicted PCR product (180bp) in the stages following ZGA (Fig 4.2.2a). We further sequenced the higher PCR band that we see on the gel as it occurs only in the nuclear fraction. On aligning the sequence of the additional PCR product onto the genome browser it mapped back to the regions in the unprocessed transcript of *hnrnpa1* RNA explaining why it is present only in the nuclear fraction after transcription activation in the 1000-Cell stage. The control, GAPDH mRNA, on the other hand, did not show much difference in the levels between the cytosolic and nuclear fractions.

Additionally, we also investigated the RNA levels of the RBMS3-antisense ncRNA, one of the antisense RNAs we had previously identified in CHAPTER 1, with respect to its overlapping protein-coding gene (*Rbms3*) in the cytosolic and nuclear fractions with development. We found that the levels of *Rbms3* antisense ncRNA (533bp product) was predominantly nuclear before the MZT. The lower band (~400bp) for Rbms3 antisense, visible in the cytosolic fraction, is the spliced RNA while the higher band (533bp) observed in the nucleus is the unspliced antisense RNA (Fig 4.2.2). In a similar fashion, the levels of the overlapping protein-coding transcript (201bp product) was more enriched in the nucleus in the pre-MZT stages, and then switches to be more cytosolic in the post-MZT stages (Fig 4.2.3b and c). We believe that the switch in the levels of the protein-coding mRNA is due to the presence of two different transcripts, one starting at the end of antisense and the other transcript beginning at the TSS of antisense.

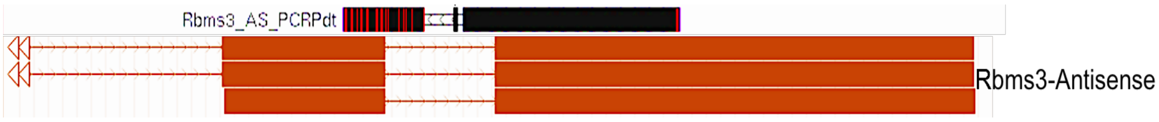


Figure 4.2.2 BLAT search result for Rbms3 AS PCR product on the browser
The schematics above is a genome browser view of Rbms3 AS PCR product (533bp, sequenced) observed in the nuclear fraction. We can see the sequence covers the intron region in the antisense.

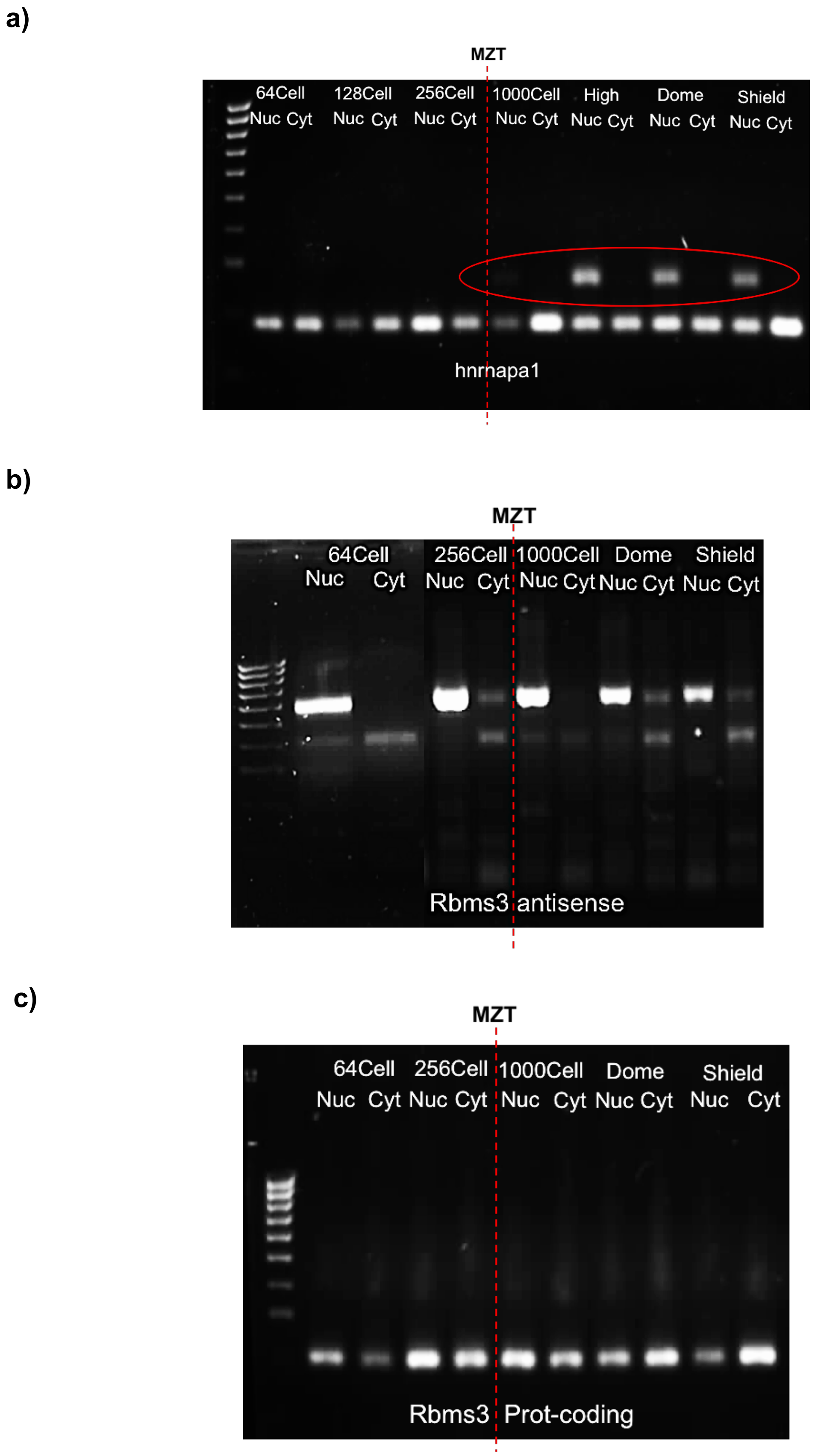

a)
MZT
64Cell Nuc Cyt
128Cell Nuc Cyt
256Cell Nuc Cyt
1000Cell Nuc Cyt
High Nuc Cyt
Dome Nuc Cyt
Shield Nuc Cyt
hnrnapa1
b)
MZT
64Cell Nuc Cyt
256Cell Nuc Cyt
1000Cell Nuc Cyt
Dome Nuc Cyt
Shield Nuc Cyt
Rbms3 antisense
c)
MZT
64Cell Nuc Cyt
256Cell Nuc Cyt
1000Cell Nuc Cyt
Dome Nuc Cyt
Shield Nuc Cyt
Rbms3 Prot-coding

Figure 4.2.3 RNA expression analysis of antisense – mRNA pairs.

PCRs showing cDNA levels of markers in the nuclear and cytosolic fractions of zebrafish embryonic cells. a) cDNA levels of hrnapa1 (mRNA), the upper band encircled in red is the unspliced mRNA while the lower band is the desired PCR product. b) cDNA levels of Rbms3 antisense (non-coding RNA) in the fractions, the upper bright band is strongly detected in the nucleus. c) cDNA levels of Rbms3 protein-coding mRNA in the fractions from different stages of development

4.2.3 Deep sequencing of nuclear and cytosolic fraction of zebrafish development

Once the purity of cellular fractions from the five stages of development was established i.e. two pre-MZT stages, 64 Cell and 256 Cell; MZT stage i.e 1000 Cell stage; and two post-MZT stages, dome and shield, were deep sequenced using Illumina's next generation sequencing technology. A total of ~420 million reads, each 200 nucleotides in length, were generated for each replicate (2x). For each replicate, ~35-40 million reads were uniquely mapped to the Zv9 zebrafish genome (Table 4.2.1).

Table 4.2.1 RNA-seq alignment summary
Showing the raw reads after quality control (trimming) and the uniquely mapped reads for the above 5 stages in the nuclear and cytosolic fractions

Fraction	Stages	Raw reads	Uniquely mapped reads	% aligned reads
Nuclear	64-Cell	40,109,064	36,430,406	90.8283624
	256-Cell	44,723,452	38,076,187	85.136959
	1000-Cell	38,770,159	35,447,369	91.4295167
	Dome	58,894,287	48,407,288	82.1935207
	Shield	49,410,255	40,372,982	81.709722
Cytosolic	64-Cell	37,710,788	33,947,606	90.0209404
	256-Cell	34,046,933	24,364,110	71.5603664
	1000-Cell	36,611,958	33,464,055	91.4019813
	Dome	44,205,046	37,017,755	83.7410168
	Shield	32,459,547	22,944,418	70.6861929

4.2.3.1 The antisense in the negatively correlated group were found to be more specifically localised through MZT than the positively correlated category

One of our major objectives to carry out nuclear and cytosolic fractionation was to determine the localization of the antisense characterized by our study on the Pauli et al. RNA-seq data as this would help us in understanding how these antisense ncRNAs regulate their overlapping protein-coding genes (Pauli et al., 2012). Studies on yeast (Long et al., 1997) and flies (Johnstone and Lasko, 2001) protein-coding genes showed very specific subcellular localisation of mRNAs to be important in development. Therefore, the specific subcellular localisation of antisense with zebrafish embryogenesis might also be essential for its regulatory function. For this analysis the antisense ncRNAs, in the negatively and positively correlated categories, were divided based on whether they were nuclear or cytosolic antisense across the five stages (Fig 4.2.4-4.2.5). To define an RNA as nuclear or cytosolic, we used the log2 ratio of nuclear and cytosolic RNA levels (Nuclear/Cytosolic). An RNA was categorised as nuclear, if the ratio was greater than 0.65 (1.5-fold enrichment), and cytosolic, if it was less than -0.65 (1.5-fold enrichment). And anything in between the ratios (0.65 and -0.65) was considered to be more or less equally present in both cellular fractions. A similar analysis was done to observe the distribution of overlapping protein-coding genes with zygotic genome activation (Fig 4.2.6).

Majority of the antisense belonging to the negatively correlated group were localised stage specifically with a subset being more enriched in cytosol (purple) or nucleus (red) post-MZT. The cytosolic antisense quite possibly suggests the importance of these antisense in post-transcriptional regulation of overlapping mRNAs while the nuclear localised antisense observed after MZT implies their requirement during transcription of the overlapping protein-coding genes. There were a few antisense (purple) that showed cytosolic localisation in the 64-Cell stage and a few others that become nuclear (red) in the 256-Cell stage. On the other hand, in the positively correlated category the fraction of antisense showing specific localisation with development was less. With majority of antisense being present in the nucleus (orange and red) in the dome and the shield stages. This was an observation made on plotting the average RNA levels for the antisense in the positively correlated group in CHAPTER 3 (Fig 3.2.2). The antisense levels increased with zygotic genome activation suggesting its transcription and therefore being nuclear localised.

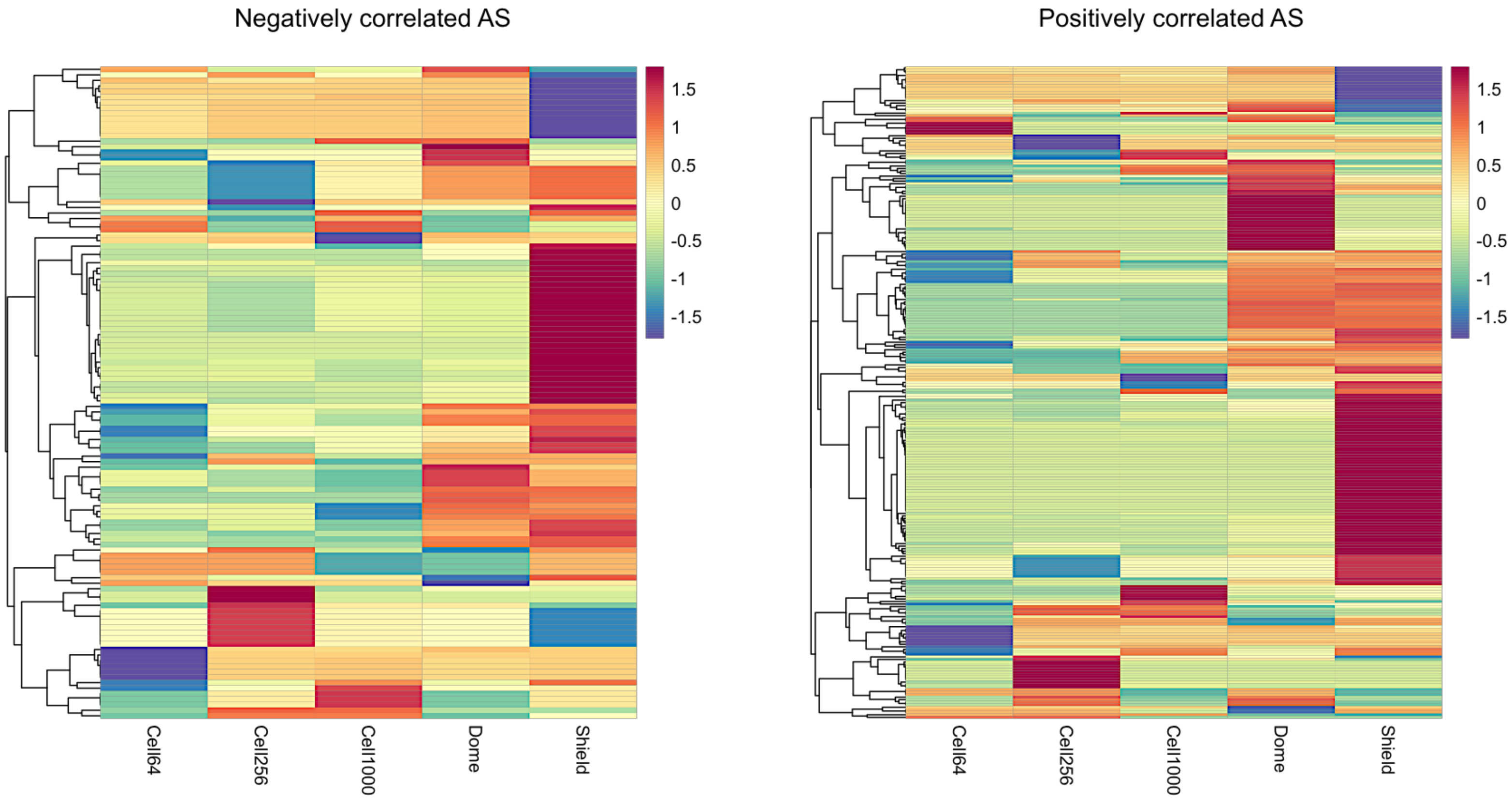


Figure 4.2.4 Localisation of antisense in the two categories.
The heatmap on the left shows the distribution of antisense in the negatively correlated group in the nuclear and cytosolic fraction through development based on the log2 ratio. The colour key suggests the colour used for different log2 ratio values (blue for cytosolic and red for nuclear). The heatmap on the right represents the distribution of antisense in positively correlated category in the nuclear and cytosolic fraction.

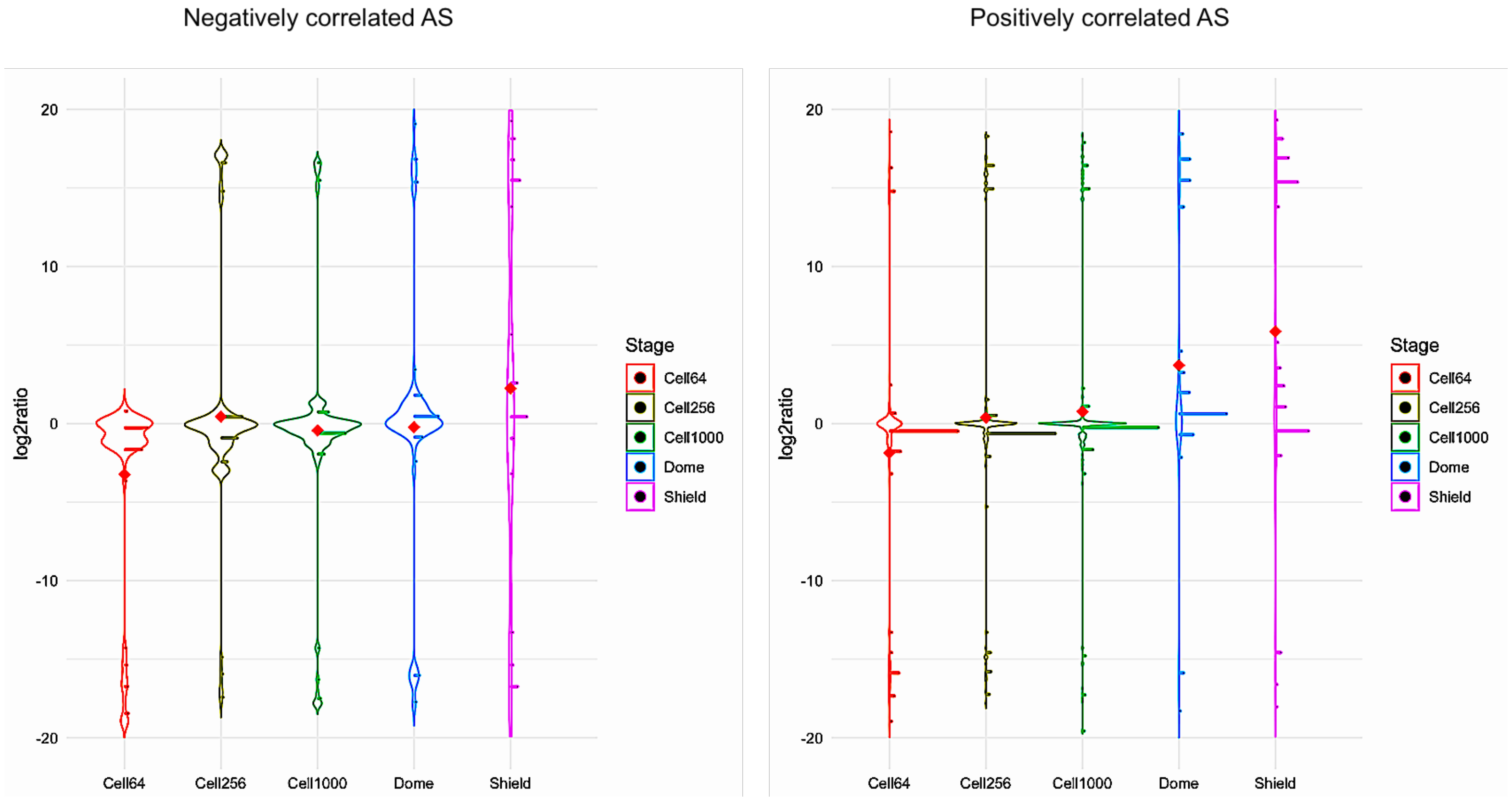


Figure 4.2.5 Localisation of antisense in the two categories.
The violin plot on the left shows the distribution of antisense in the negatively correlated group in the nuclear and cytosolic fraction through development based on the log2 ratio (Y-axis). The colour key and the X-axis suggests the colour used for different stages of development. The violin plot on the right represents the distribution of antisense in positively correlated category in the nuclear and cytosolic fraction.

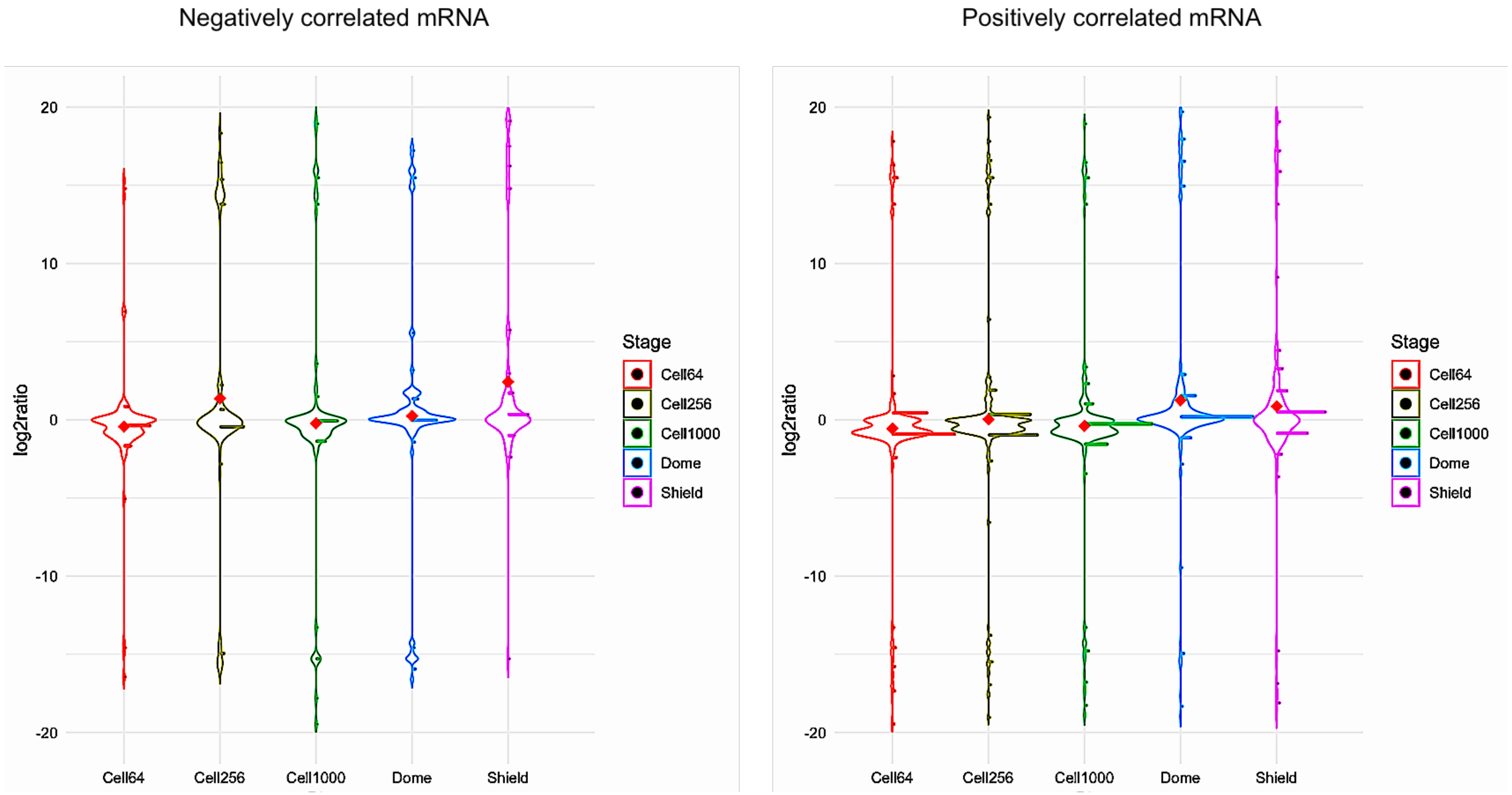


Figure 4.2.6 Localisation of overlapping protein-coding in the two categories.
The violin plot on the left shows the distribution of mRNAs in negatively correlated group in the nuclear and cytosolic fraction through development based on the log2 ratio (Y-axis). The colour keyand the X-axis suggests the colour used for different stages of development. The violin plot on the right represents the distribution of mRNAs in positively correlated category in the nuclear and cytosolic fraction.

4.2.3.2 Maternal RNA dynamics and their cellular localization

It is expected that maternal RNAs are deposited mainly in cytoplasm and their levels drop as zygotic transcription is initiated. Before zygotic genome activation, the growth of the embryo is completely dependent on housekeeping functions of maternal RNAs. The maternal RNAs, being mostly housekeeping, are required to be replaced by same RNAs transcribed from zygotic genome. Therefore, the anticipation would be that after zygotic transcription initiates the levels of housekeeping RNAs in nuclear fractions will go up. We examined the RNA-seq data from nuclear and cytosolic fractions to verify these changes in maternal RNAs (Fig 4.2.7). RNA levels of predicted 1719 maternal genes (Harvey et al., 2013) was studied. To define an RNA as nuclear or cytosolic, we used the log2 ratio of nuclear and cytosolic RNA levels (Nuclear/Cytosolic) as explained in the previous section (section 4.2.3.1) but combined the cytosolic and both categories.

As expected, majority of maternal RNAs (99%) are cytosolic in the early stages. The percentage of cytosolic RNAs decreased (92%) at 3hpf where zygotic genome activation normally takes place (Fig 4.2.7). In contrast, the percentage of maternal RNAs in nuclear fraction considerably increased upon activation of zygotic transcription (from 0.95% to ~18%).

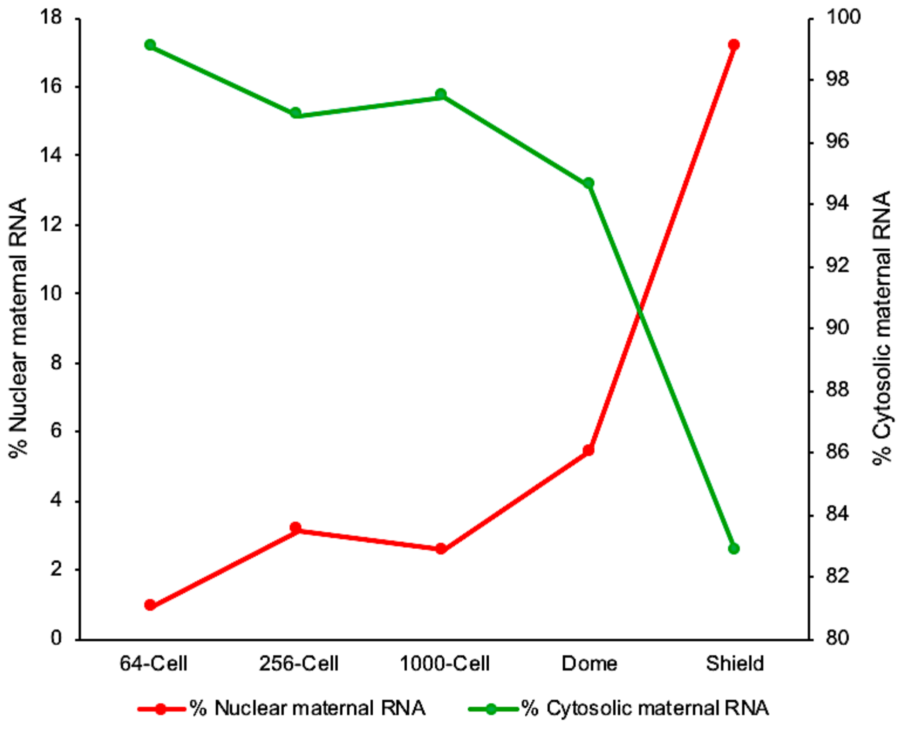


***Figure 4.2.7 Localisation of maternal RNAs with MZT**.*
A line plot showing the levels of cytosolic maternal genes in green (represented by right-hand Y-axis) and nuclear maternal genes in red (represented by left-hand Y-axis). The X-axis shows the stages of development. The percentage was calculated as the number of genes which were either cytosolic or nuclear, based on the log2 ratio, divided by the total number of genes.

4.2.3.3 mRNAs show distinct levels in the nuclear and cytosolic fractions

The cytosolic enrichment of maternal RNAs pre-MZT and the increase in the percentage of nuclear RNAs post-MZT supports that our RNA-seq data and cellular fractions reflect known MZT related behaviour of maternal RNAs. To investigate if other mRNAs show dynamics similar to maternal RNAs, analysis was carried out to study distribution of coding mRNAs with zebrafish development. This would give us an additional knowledge about the types of RNA that are present pre-MZT (maternal) and post-MZT (mostly zygotic).

On plotting the percentage of mRNAs which were nuclear or cytosolic, based on the log2 ratio, with the different stages of development (Fig 4.2.8), we found that the majority (95%) of mRNAs were cytosolic during the early stages suggesting they were maternal RNAs. Similar to maternal RNAs, with time a decrease in the percentage of cytosolic mRNAs (65%) can be observed. This decrease could be due to the degradation of majority of maternal RNAs both during MZT and post-MZT. Also, as seen in case of known maternal RNAs (Fig 4.2.7) the percentage of nuclear mRNAs increased (by 30%) most likely contributed by zygotic transcription due to zygotic genome activation.

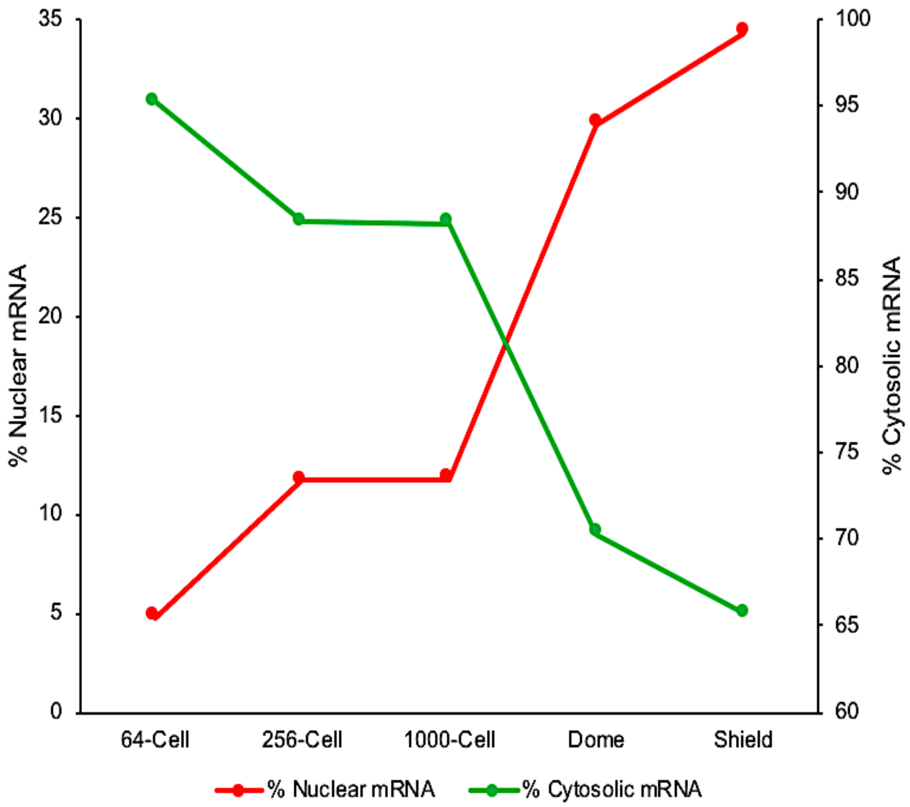


Figure 4.2.8 Localisation of mRNAs with MZT

Line plot showing the percentage of mRNA that are nuclear or cytosolic with development. Right-hand Y-axis denotes percentage of cytosolic mRNAs while left hand Y-axis represents percentage of nuclear mRNAs. X-axis shows stages of development. The percentage was calculated as the number of genes which were either cytosolic or nuclear, based on the log2 ratio, divided by the total number of genes. There is a decrease in the percentage of cytosolic mRNA from 95% to 65% by the shield stage while an increase from 5% to almost 35% in the nuclear mRNA levels.

4.2.3.4 Changes in the distribution of RNA-seq reads across the maternal and zygotic TSSs during development

The read distribution across individual RNAs in nuclear fraction before and after zygotic genome activation was also plotted in the form of a heatmap. This can give a good idea about changes in transcriptional and RNA levels during early embryonic development. Nuclear and cytosolic reads were mapped and centered on transcriptional start site of previously identified maternal (1719) and zygotic genes (257) (Harvey et al., 2013).

The heatmap showed that nuclear read density at maternal TSS (Fig 4.2.9) decreased with progress in developmental stages. The shield stage showed much fewer reads at the TSS with an increase in downstream reads possibly indicating transcription (ZGA). In comparison, in the cytosolic fraction (Fig 4.2.10), the reads were highly abundant (intensity greater than 140) but showed a similar pattern of distribution. However, the shield stage revealed fewer reads overall suggesting maternal RNA degradation which was also observed in Fig 4.2.10. On the other hand, the zygotic RNAs are not expressed in the early stages (64-cell, 256-cell and 1000-cell), barely any RNA-seq reads were observed in the nuclear fractions. However, in the later developmental stages (dome and shield) reads were observed downstream of the zygotic TSS suggesting their transcriptional activation. In the cytosolic fraction we see a similar pattern of reads distribution downstream and at TSS for zygotic genes with MZT.

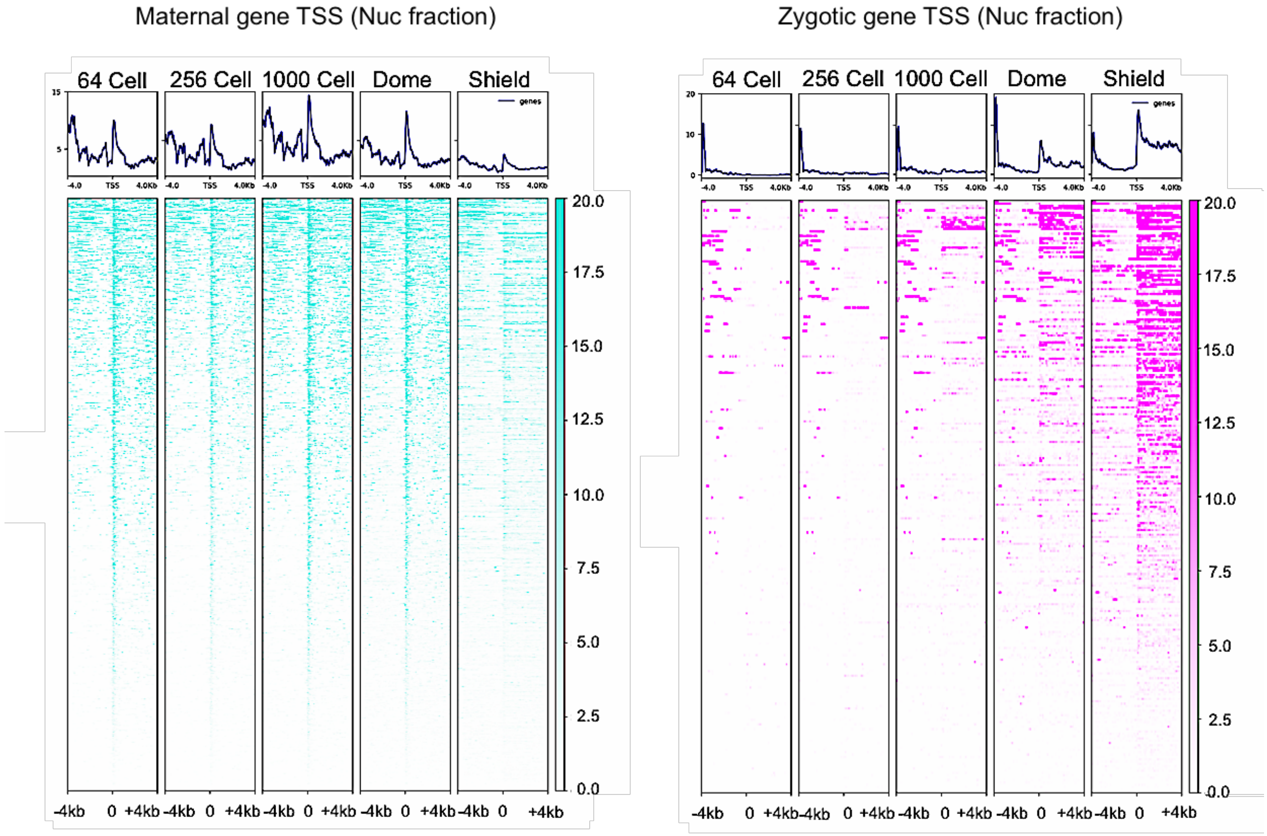


Figure 4.2.9 RNA-seq read distribution at the TSS in the nuclear fraction.
The heatmap shows the distribution of reads across the maternal (on the left, teal) and zygotic TSS (on the right, magenta) during zebrafish development. Each panel in the heatmap represents a developmental stage and the width is -4kb on the left to +4kb on the right with the TSS in the centre (X-axis). The y-axis represents all the maternal and zygotic RNAs and the z-axis is the intensity of the read (0-20).

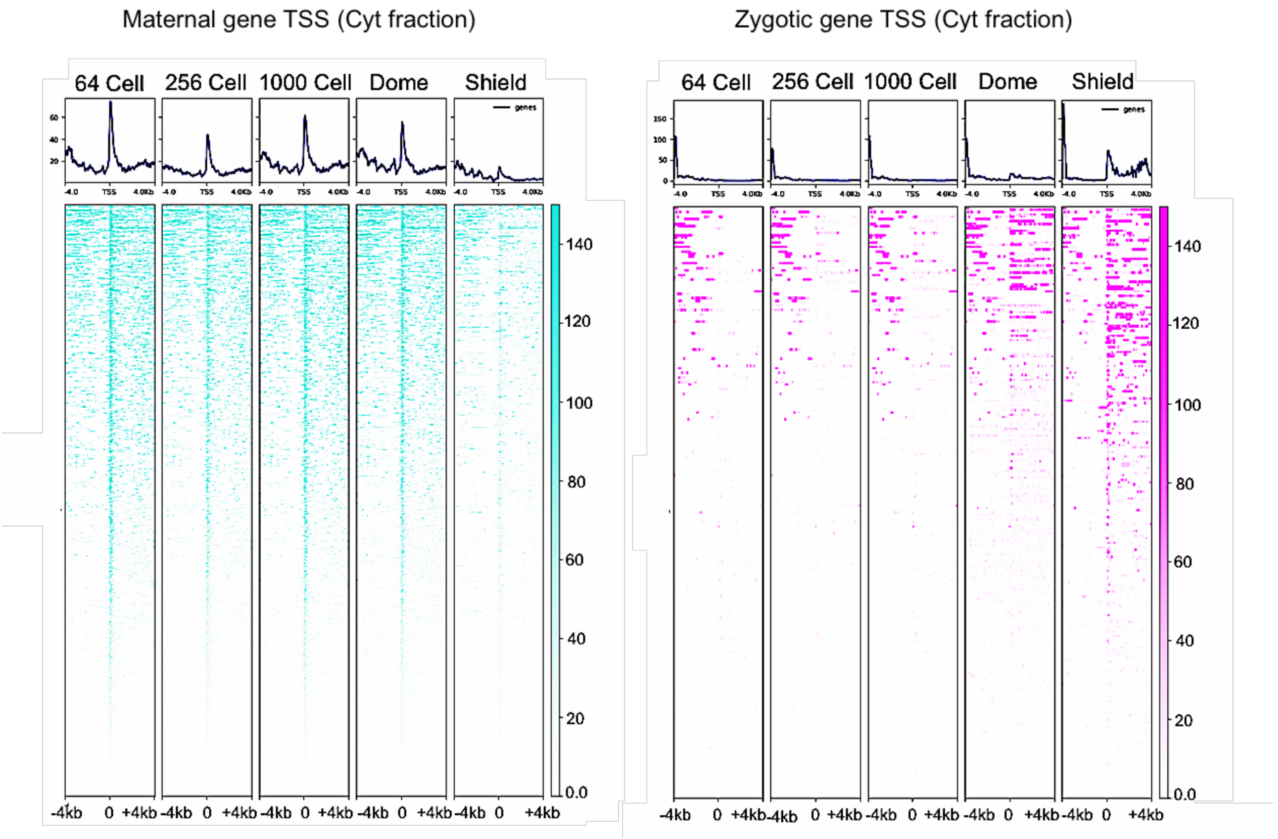
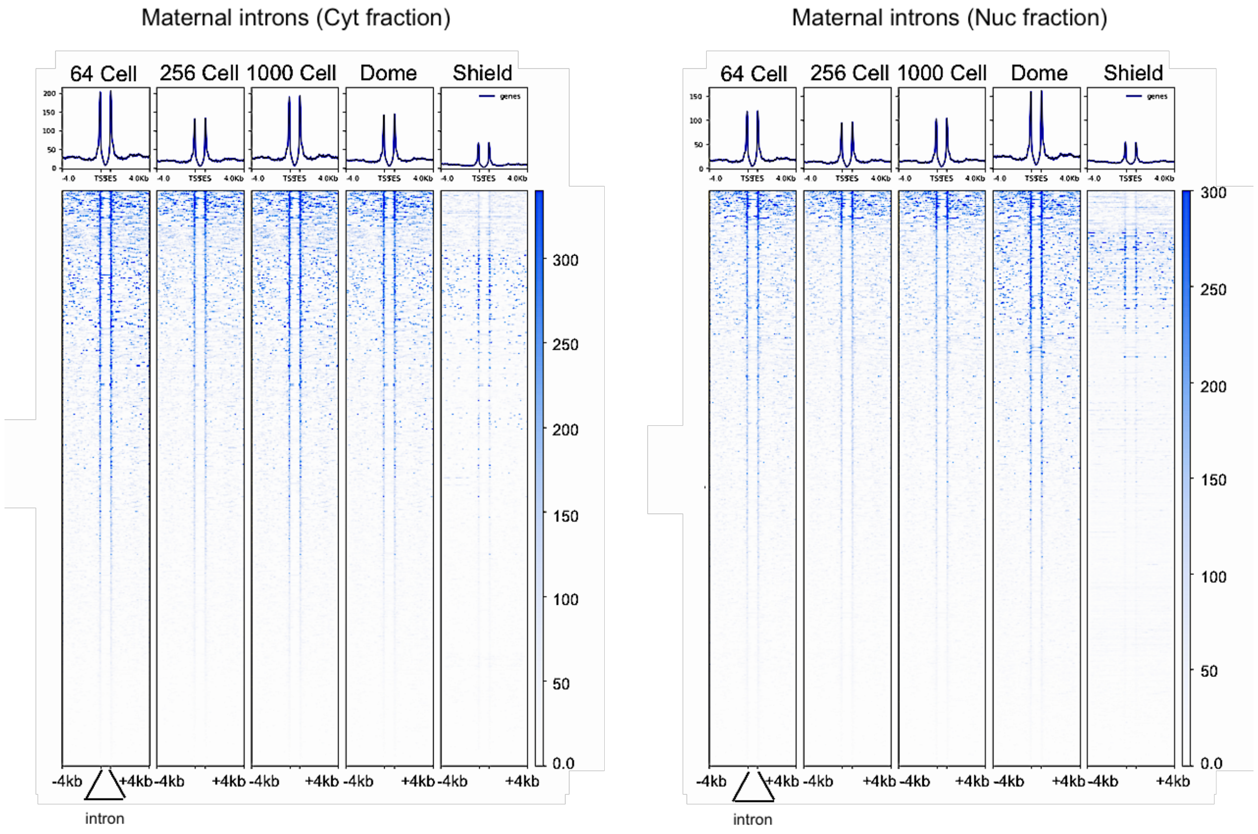


Figure 4.2.10 RNA-seq read distribution at the TSS in the cytosolic fraction.
The heatmap shows the distribution of reads across the maternal (on the left, teal) and zygotic TSS (on the right, magenta) during zebrafish development. Each panel in the heatmap represents a developmental stage and the width is -4kb on the left to +4kb on the right with the TSS in the centre (X-axis). The y-axis represents all the maternal and zygotic RNAs and the z-axis is the intensity of the read (0-140).

4.2.3.5 Zygotic introns showed presence of reads in the nuclear fraction with zygotic genome activation

Studies suggest that maternally deposited RNAs are post-transcriptionally processed mRNA products or spliced with a lot of them (50-60%) showing splice variant (Aanes et al., 2011). Given there is no or very low level of RNA synthesis before the zygotic genome activation (at 3hpf) we should see little or no intronic reads. Another advantage of investigating intronic reads in the nucleus in genes during embryogenesis is that it would be useful in detecting early zygotic transcription, if any, and also in validating that our fractions are clean. We mapped both nuclear and cytosolic reads to the annotated introns of maternal and zygotic RNAs.

As expected, the maternal introns in both the fractions showed a complete lack of reads in all the five stages considered (Fig 4.2.11). The intron-exon junction, however, showed presence of reads suggesting the presence of spliced maternal RNAs. The reads also showed an increase in the dome stage in the nuclear fraction, absent in cytosolic fraction, possibly implying their expression with genome activation. The zygotic introns also did not show any enrichment for reads in both the nuclear and cytosolic fractions pre-MZT (Fig 4.2.12). However, we did see the presence of reads at the introns post-MZT (dome and Shield) with an increase in expression. The results are also a reflection of the quality of our cytosolic and nuclear fractions.



Maternal introns (Cyt fraction)

Maternal introns (Nuc fraction)

Figure 4.2.11 RNA-seq read distribution across the maternal introns.

The heatmap shows the distribution of reads across the maternal introns in the cytosolic (on the left) and nuclear fraction (on the right) during zebrafish development. Each panel in the heatmap represents a developmental stage and the width is ±4kb from the intron (TSS to TES) represented by the schematics in the centre (X-axis). The y-axis represents all the maternal introns and the z-axis is the intensity of the read (0-300).

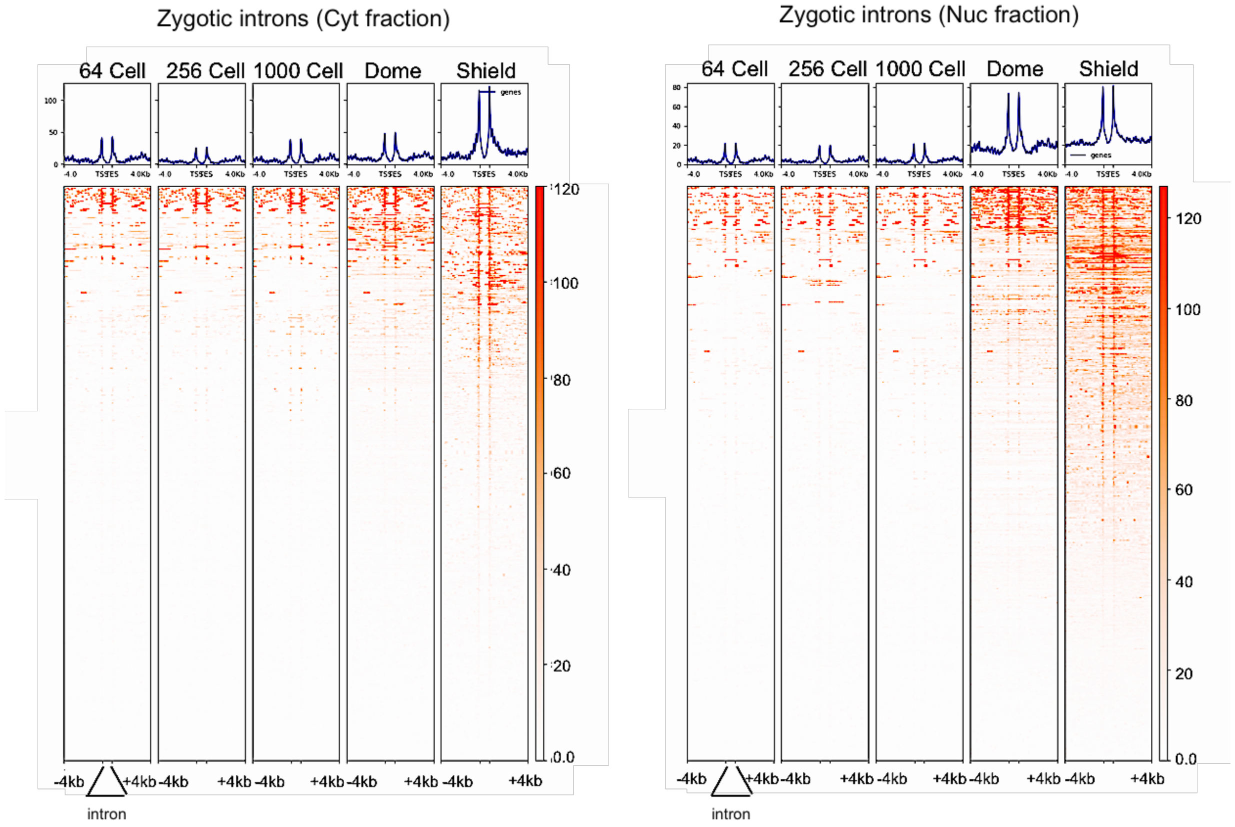


Figure 4.2.12 RNA-seq read distribution across the zygotic introns.
The heatmap shows the distribution of reads across the zygotic introns in the cytosolic (on the left) and nuclear fraction (on the right) during zebrafish development. Each panel in the heatmap represents a developmental stage and the width is ±4kb from the intron (TSS to TES) represented by the schematics in the centre (X-axis). The y-axis represents all the zygotic introns and the z-axis is the intensity of the read (0-120).

4.2.3.6 NcRNAs are more enriched in the cytosol but become nuclear with zygotic genome activation (or MZT)

It will be interesting to see if non-coding RNAs show similar patterns of localisation as in case of coding RNAs. Non-coding RNAs are known to be associated with transcription or post-transcription regulation of gene expression and are expected to be mostly localized in the nucleus (Djebali et al., 2012). Localisation of ncRNAs, both small and long, in RNA-seq data was analysed and their expression pattern was checked during maternal-to-zygotic transition (Fig 4.2.13). Similar to mRNAs, in early stages, ncRNAs are more enriched in the cytosol (59%) as compared to only 21% in the nucleus, suggestive of being maternally deposited. However, with the maternal-to-zygotic transition and subsequent transcription activation, ncRNAs eventually become much more localized in the nucleus (74%) in comparison to mRNAs (13%).

This pattern was further represented using a heatmap of all ncRNAs (RNA levels above 1 TPM) in the nuclear and cytosolic fractions (APPENDIX, Fig 2). The analysis also revealed the stage specificity of these ncRNAs. Both, mRNA and ncRNA levels in the nucleus, increased more than two-fold during the course of development contributed by zygotic transcription activation. During all stages, it can be seen that the nuclear ncRNA levels are higher as compared to the mRNA levels (Fig 4.2.14). In contrast, the cytosolic fraction shows higher level of mRNAs relative to ncRNA. Levels of cytosolic RNAs, ncRNA as well as mRNAs, decreased with maternal-to-zygotic transition most likely due to active degradation. Interestingly, however, cytosolic ncRNAs showed a much rapid decrease following maternal-to-zygotic transition as compared to mRNAs.

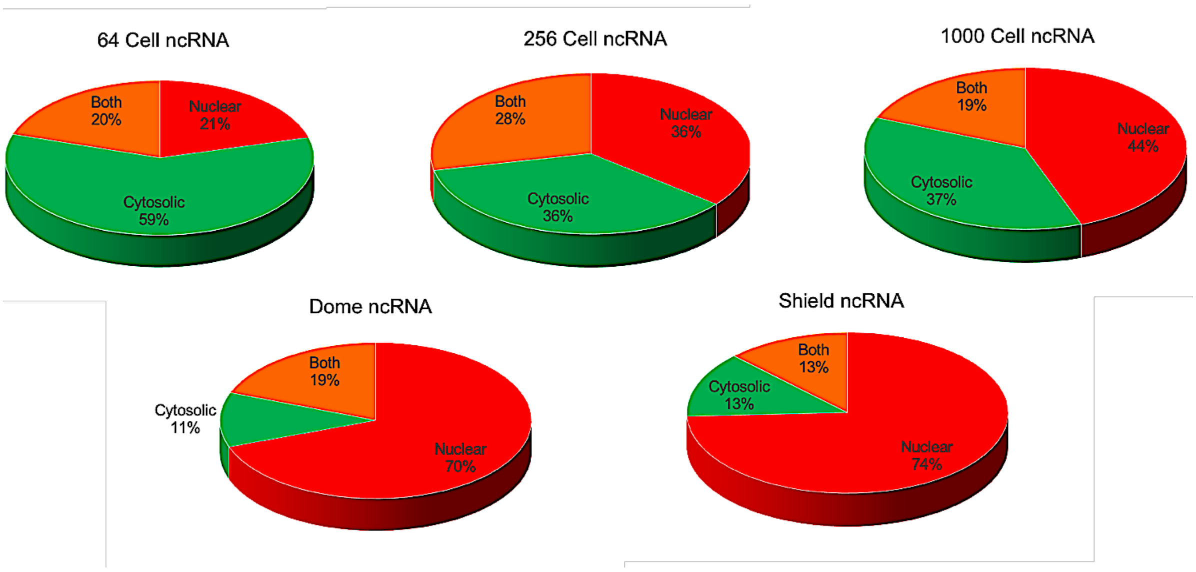


Figure 4.2.13 NcRNA localization with MZT
Non-coding RNA localisation. It shows the distribution of ncRNAs in the nuclear and cytosolic compartments during zebrafish embryogenesis. The red colour means the RNAs are nuclear, green means they are cytosolic, and orange means they are distributed equally in both the fractions.

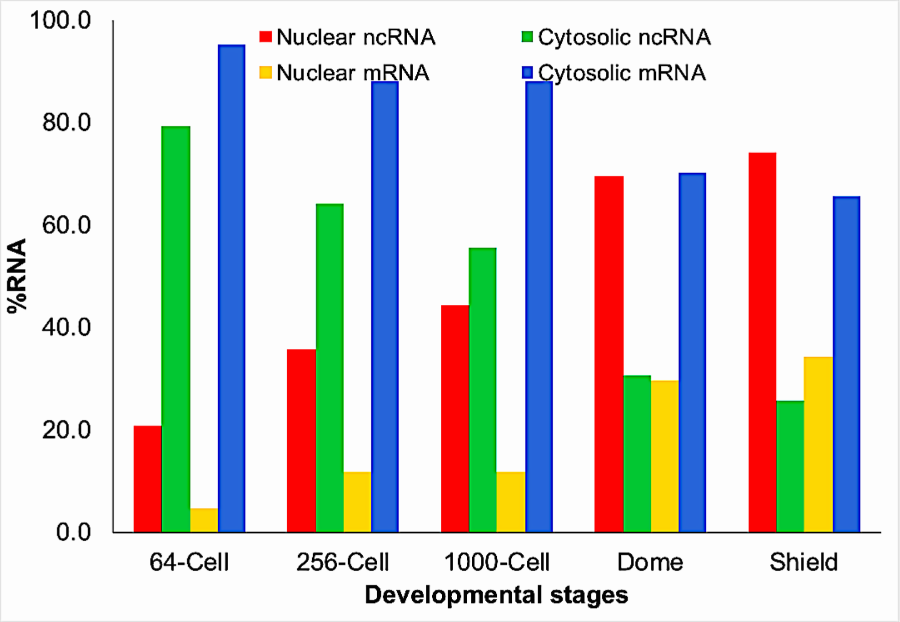


Figure 4.2.14 Comparison between ncRNA and mRNA

Bar plot comparison between mRNAs and ncRNAs levels in the two fractions. X-axis shows the percentage of nuclear and cytosolic ncRNAs (cytosolic + both) and mRNA levels while the Y-axis represents the stages of development. red bars represent nuclear ncRNA, green for cytosolic ncRNA, yellow for nuclear mRNA and blue shows cytosolic mRNA.

4.2.3.6.1 Distribution of reads across the TSS of ncRNAs in the nuclear fraction shows a more intense bidirectional pattern in the later stages

From the previous sections we have seen how with embryonic development the localisation as well as expression of both coding and non-coding RNAs change. Now observing the distribution of reads upstream and downstream across the TSS for these coding and non-coding regions could give us a general idea of the transcription dynamics that are occurring. Reads from the nuclear fraction for the different stages of development, were mapped and centered at the TSS of all annotated mRNAs and ncRNAs (Fig 4.2.15).

On analysing the coverage of reads across the TSS for the mRNAs we found that they show more reads downstream of TSS, in the direction of transcription after the activation of zygotic genome (Dome and Shield). This is what we had observed in the previous section and was expected with newer transcripts being synthesised. In contrast, the distribution of reads across the TSS of ncRNAs shows a bidirectional pattern as we see a lot of reads both upstream and downstream which becomes prominent with maternal zygotic transition. This could be due to the presence of an adjacent transcription start site upstream or downstream of the non-coding RNAs TSS.

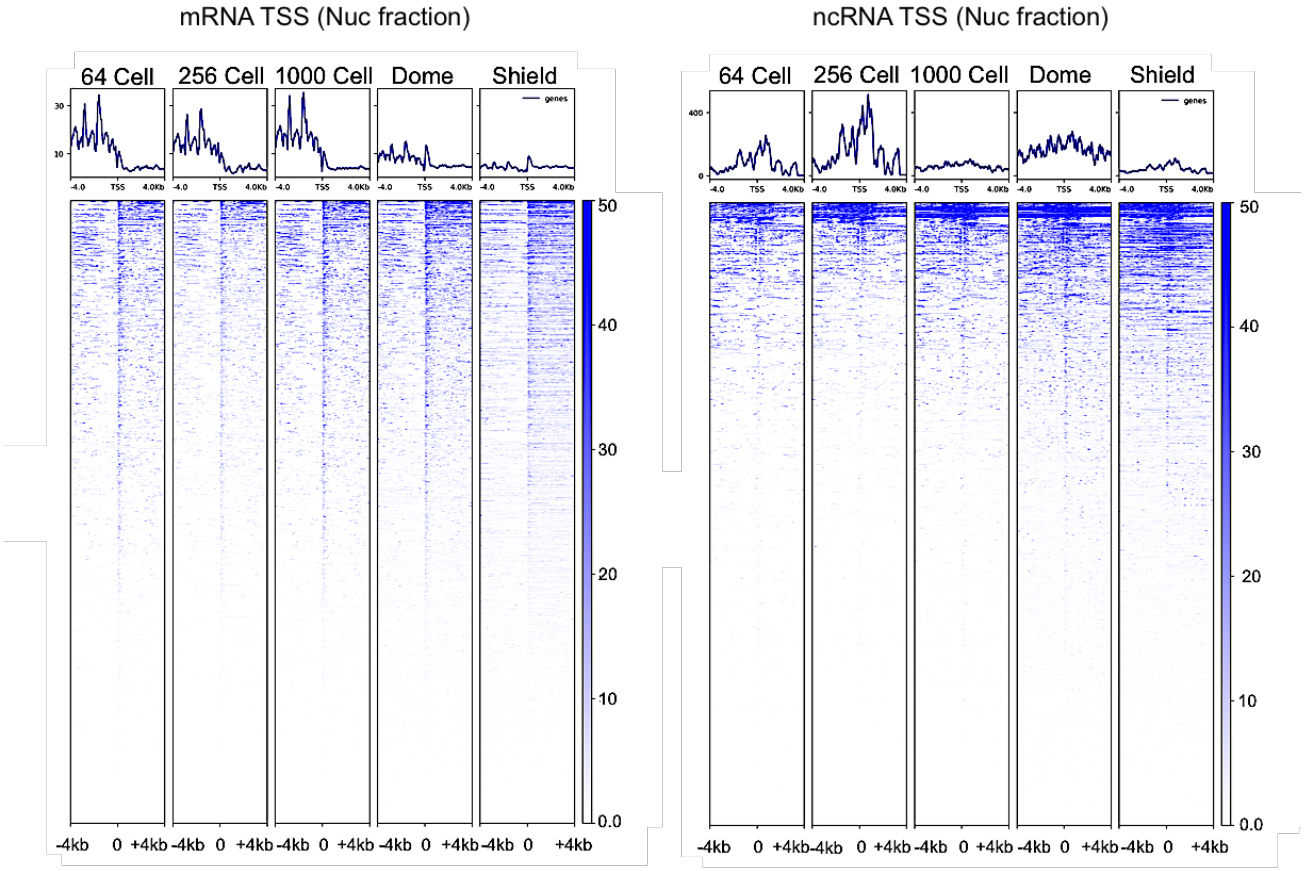


Figure 4.2.15 Nuclear RNA-seq read distribution at the TSS of ncRNAs and mRNAs
Read distribution across TSS. Heatmap showing read coverage across the TSS of both mRNA (on the left) and ncRNA (on the right). Each panel in the heatmap represents a developmental stage and the width is ±4kb from the TSS in the centre (X-axis). The y-axis represents all the ncRNA and mRNAs and the z-axis is the intensity of the reads (0-50).

4.2.3.6.2 Our dataset detects ncRNAs with two to three-fold higher depth than a previous zebrafish RNA-seq data

To estimate the quality and coverage of our fractionation data in detecting ncRNAs, particularly, we compared the expression of ncRNAs detected in both the nuclear and cytosolic fraction to total RNA-seq data (Pauli et al., 2012).

Compared to the study (Pauli et al., 2012) both of our fractions detect early stage ncRNAs with a significantly higher depth (Fig 4.2.16-4.2.18). This is due to the nuclear enrichment that we carried out for this study and also because our study includes sequencing for both polyA+ and polyA- RNAs. Moreover, in our study we identified several unspliced RNAs in the nuclear fraction after zygotic genome activation (Dome and Shield) which is either absent or very lowly detected in the Pauli et al. study (Pauli et al., 2012). Therefore, our dataset is highly comprehensive in detecting both early stage ncRNAs (64-Cell) and unspliced ncRNAs that are transcribed prior to MZT.

The examples highlighted below showed a higher coverage of reads for ncRNAs in all the stages considered when compared to the Pauli et al. RNA-seq (Pauli et al., 2012). The *si:dkey-81p22.11* antisense to *Hoxc3a, Hoxc4a, Hoxc5a* and *Hoxc6a* genes, essential in embryonic development, showed a three-fold higher coverage in our dataset (Fig 4.2.17). Our dataset also detected a higher coverage for a protein-coding gene, *mafbb* (Fig 4.2.16)*,* in the early stages and to some extent the stages after MZT (Shield).

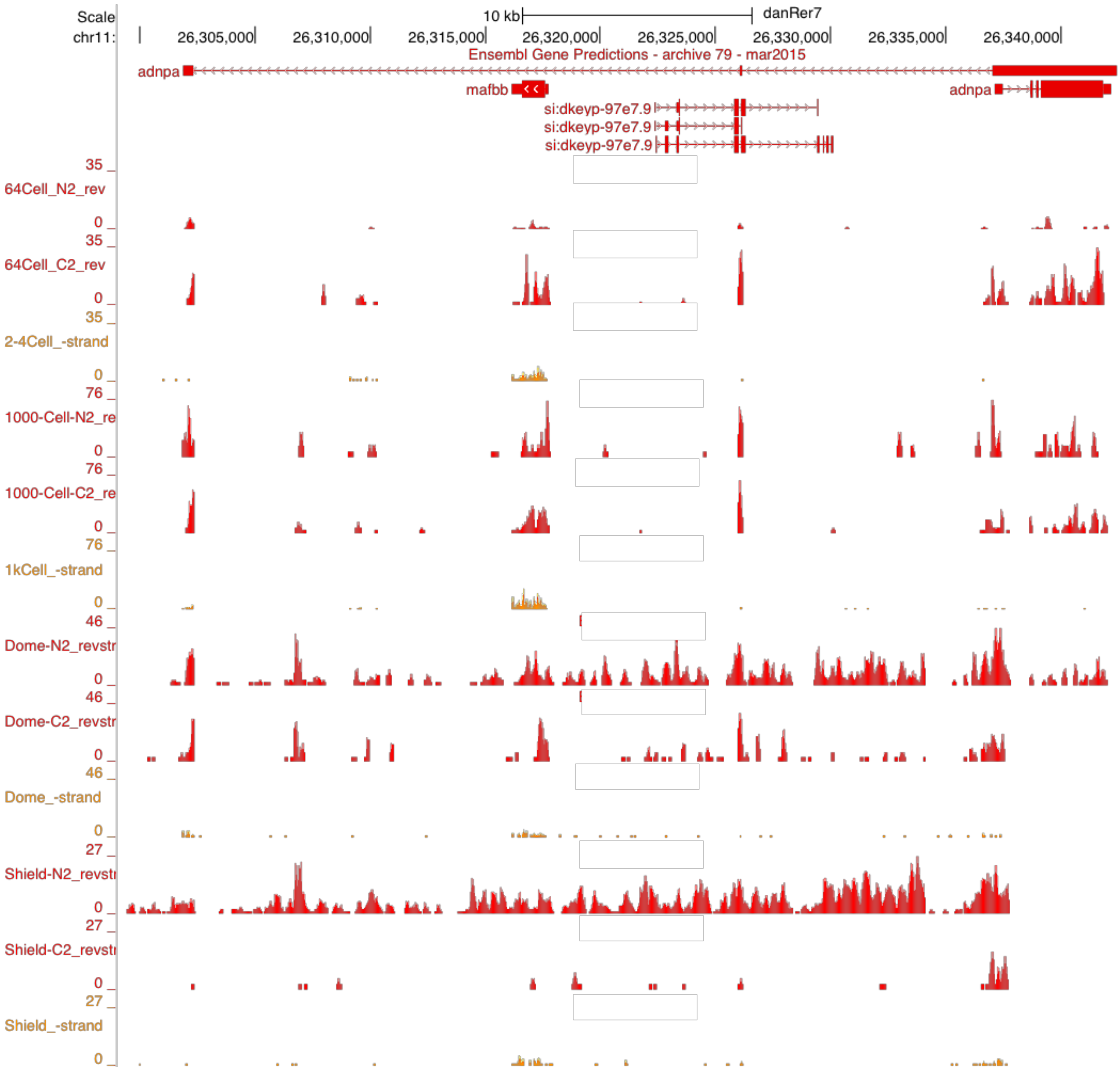


Figure 4.2.16 Genome browser view of RNA-seq coverage across adnpa antisense.

The expression profile in red represents our data (nuclear, N2 and cytosolic, C2) and in orange is the RNA-seq coverage for Pauli et al. (Pauli et al., 2012) for comparable stages. The adnpa antisense ncRNA gene (minus strand) shows a higher presence of reads (red) in all the stages when compared to Pauli et al. RNA-seq [108].

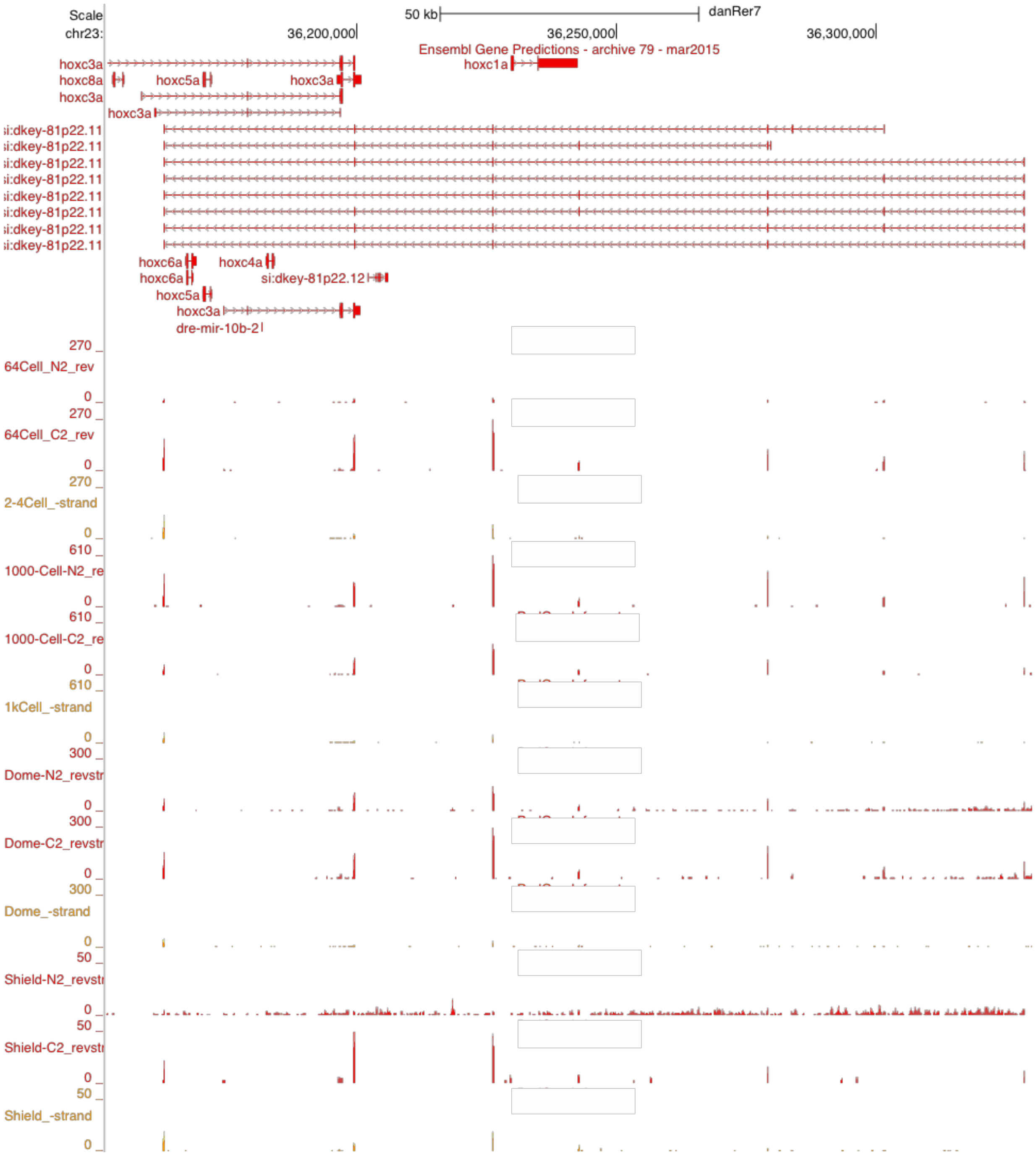


Figure 4.2.17 Genome browser view of RNA-seq coverage across si:dkey-81p22.11.

The expression profile in red represents our data (nuclear, N2 and cytosolic, C2) and in orange is the RNA-seq coverage for Pauli et al. (Pauli et al., 2012) for comparable stages. The si:dkey-81p22.11 antisense ncRNA gene (minus strand) shows a higher presence of reads (red) in all the stages when compared to Pauli et al. RNA-seq (Pauli et al., 2012).

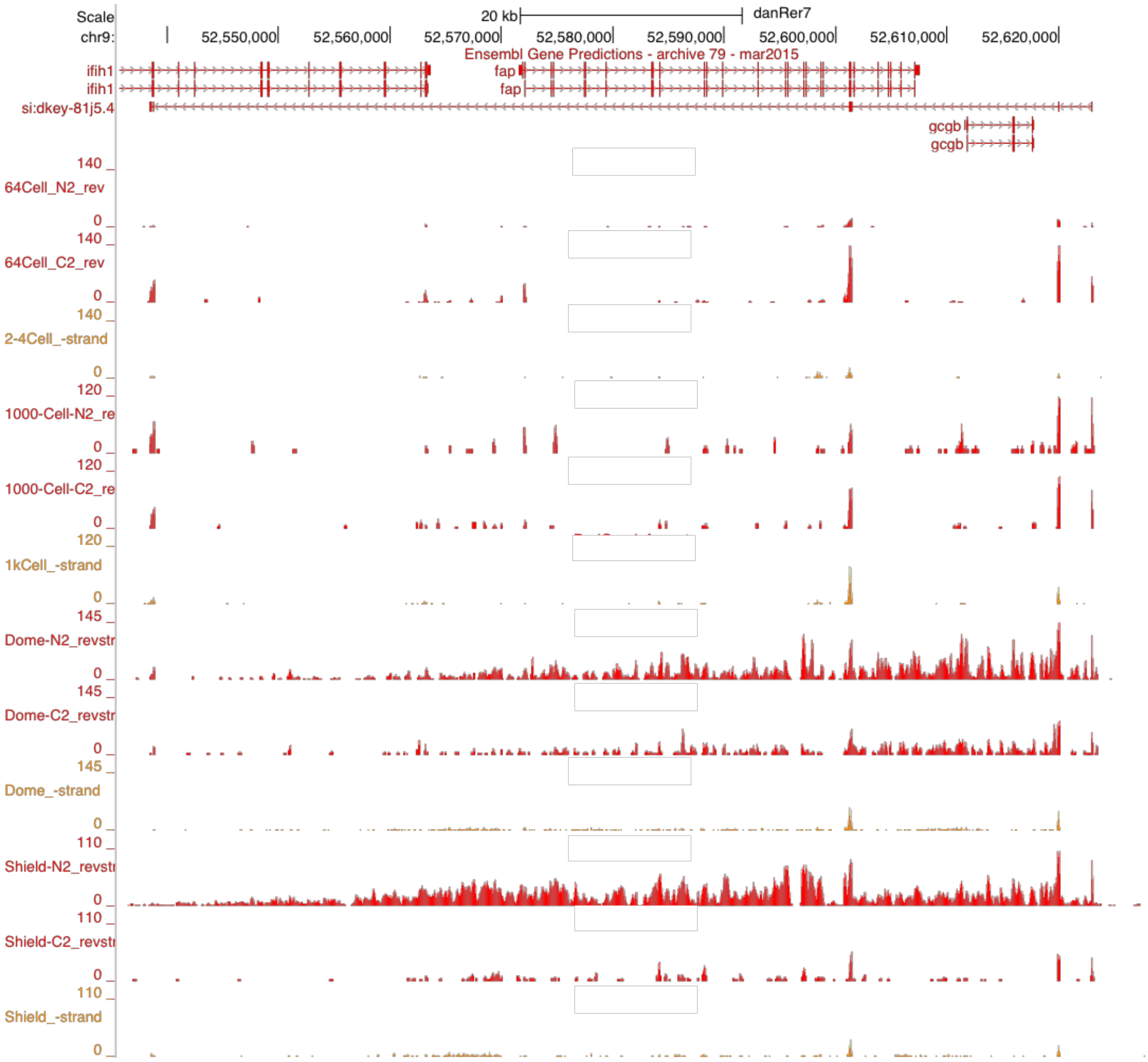


Figure 4.2.18 Genome browser view of RNA-seq coverage across si:dkey-81j5.4.

The expression profile in red represents our data (nuclear, N2 and cytosolic, C2) and in orange is the RNA-seq coverage for Pauli et al. for comparable stages (Pauli et al., 2012). The si:dkey-81j5.4 antisense ncRNA gene (minus strand) shows a higher presence of reads (red) in all the stages when compared to Pauli et al. RNA-seq (Pauli et al., 2012).

4.2.3.6.3 Majority of the nuclear ncRNAs in the early stage were either lincRNAs or antisense ncRNA category

Several examples show that long and short ncRNAs are involved in regulation of gene expression. Long ncRNAs such as lincRNAs and antisense RNAs have been shown to be involved in the regulation of neighbouring or overlapping protein-coding gene expression (in *cis*) or genes elsewhere on the chromosome (in *trans*) (Pauli et al., 2011). miRNAs, on the other hand, processed by Dicer complex and transported to the complimentary RNA, through binding with argonaute proteins, leads to degradation of the complimentary RNA and thus, silencing of gene expression (Pauli et al., 2011). On the other hand, snoRNAs and snRNAs have been shown to be important in gene activation via their association with the open chromatin structure (Matera et al., 2007). To understand the regulation of early developmental events, it is therefore necessary to understand what types of ncRNAs are abundant in the stages before and after zygotic genome activation (MZT). We first investigated the types of small ncRNAs (miRNA, snRNA and snoRNA) and long ncRNAs (lincRNA, antisense) that are present during early cytosolic and nuclear compartments. Study on human cell lines have suggested that miRNAs are predominantly present in the cytosol, snoRNAs in the nucleus and snRNAs seemed to equally present in both the fractions (Djebali et al., 2012). However, studying dynamic ncRNAs during development especially through maternal-to-zygotic transition could give us insight into functions of these essential ncRNAs in regulating development.

All the ncRNAs with levels greater than 10 TPM (transcript per million) were classified according to their type and were analysed during the 5 stages of development (Fig 4.2.19). The most predominant forms of ncRNAs in nuclear fractions was, lincRNAs and antisense ncRNAs. This was especially true for early pre-MZT stages (64c cell and 256 cell) This could be due to the role of lncRNAs in post-transcriptional regulation of gene expression. They might be deposited maternally to control the early stages of development prior to ZGA. While small nuclear RNAs (snRNAs) and miRNAs are the predominantly ncRNA class enriched in the nuclear fraction of the dome and shield stages. This can be explained by the fact that transcription of miRNAs is induced post-MZT as they are needed for clearance of maternal RNAs. Also, the role of snRNAs has been very well established in alternative splicing making their transcription a priority owing to their function in post-transcriptional processing.

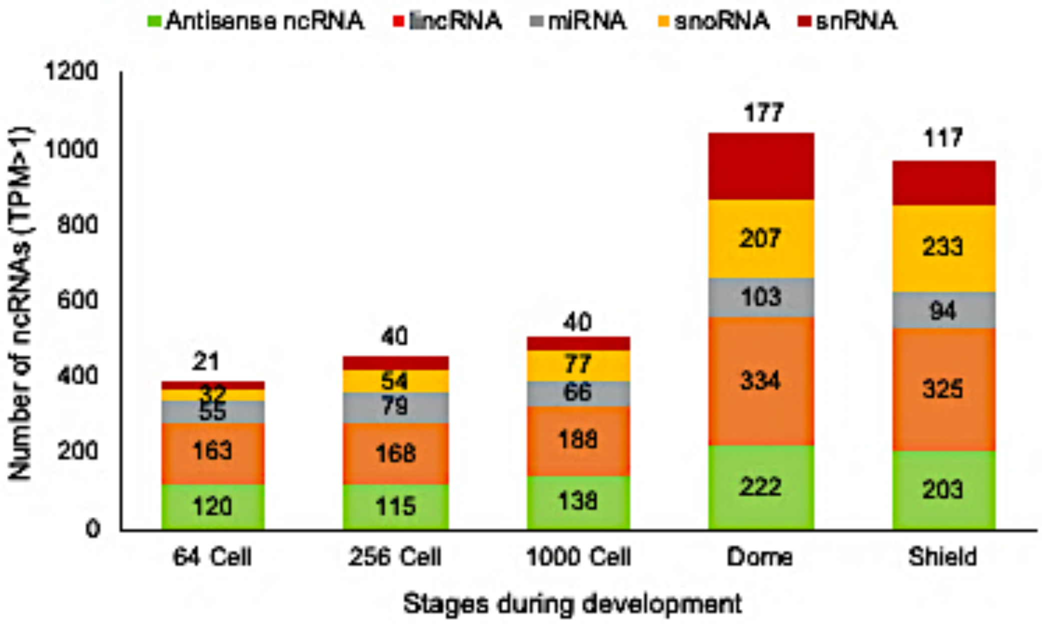
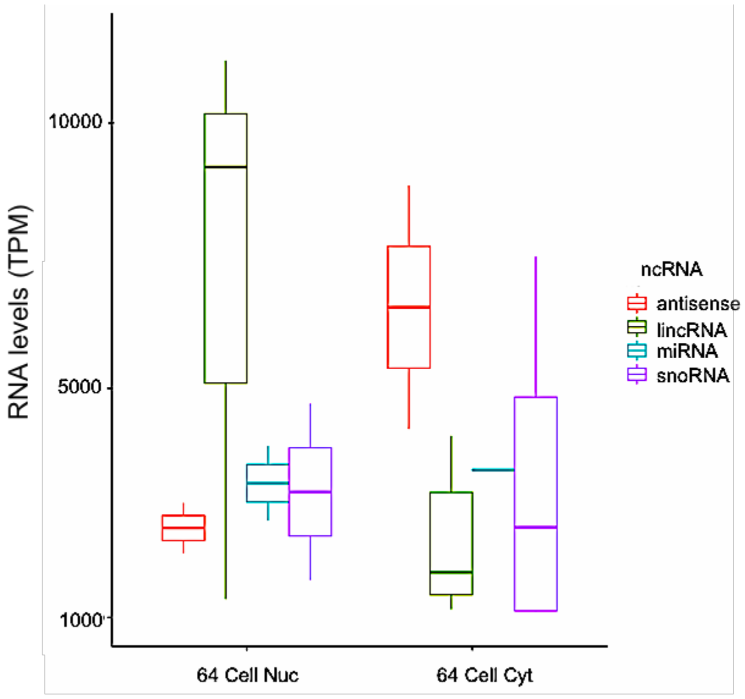


Figure 4.2.19 Nuclear localisation of all ncRNAs.
The bar chart above shows the distribution of the different types of ncRNAs in the nuclear fraction with zebrafish development. The X-axis represents the stages of development considered and the Y-axis shows the number of ncRNAs detected (TPM >1). The red shows the number of snRNAs, yellow is for snoRNAs, grey for miRNAs, orange for lincRNAs and green for antisense.

In addition to the type of ncRNAs, abundance of ncRNAs was also studied. For simplicity, two stages of development, one pre-MZT (64-Cell) and a post-MZT stage (Shield), were studied. ncRNAs with levels greater than or equal to 1000 TPMs were considered for this analysis. This analysis not only reflected how types of ncRNAs vary before and after MZT but also showed that their localization in cytosolic and nuclear compartment is quite distinct.

Investigation of the classes of RNAs in the two stages showed that antisense ncRNAs are the most abundant RNAs in the 64 Cell cytosolic fraction whilst in the nuclear fraction it is the lincRNAs that had much higher levels compared to other types of ncRNA (Fig 4.2.20). On the contrary, in the shield stage or post-MZT, the small nuclear RNAs (snRNAs) are the highly expressed ncRNAs in both the nuclear and cytosolic fractions with TPMs greater than 5000. The change in ncRNA repertoire between the two stages was also very striking. The lincRNA levels (light green) in the nucleus dropped dramatically from before MZT to post MZT while the antisense levels (red) remained almost same with MZT. The number of high expressing miRNAs (blue pre-MZT and dark green post-MZT) in the nucleus increased post-MZT. The snoRNAs levels also increased in the nuclear fraction with zygotic genome activation (purple pre-MZT and blue post-MZT). We did not detect any snRNAs (above 1000 TPM) pre-MZT but they were highly abundant in both fractions post-MZT (in purple).



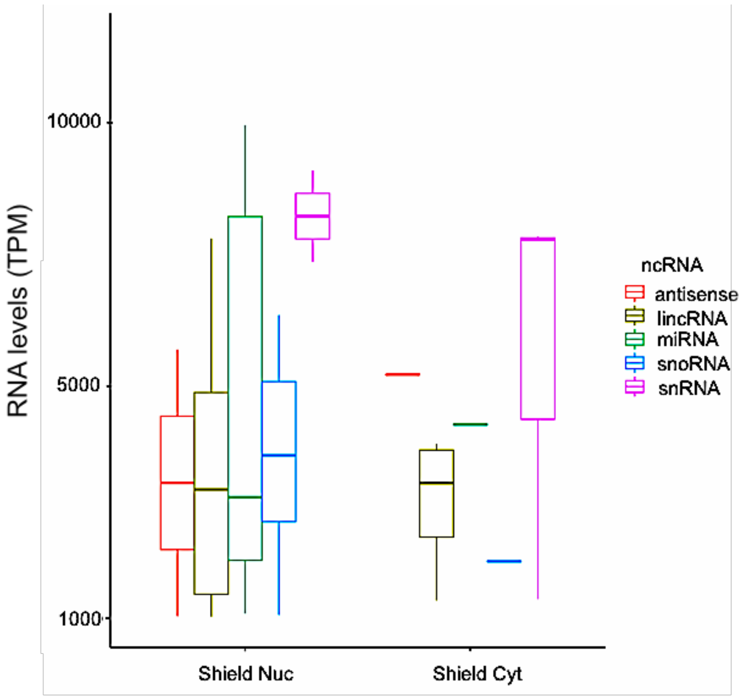


Figure 4.2.20 ncRNA levels in the nuclear and cytosolic fraction

RNA levels of ncRNAs. The plot shows the highly abundant RNAs in the 64-cell nuclear and cytosolic fraction and the shield nuclear and cytosolic fractions. The X-axis represents the fractions (nuclear on left and cytosolic on right) and the Y-axis shows the RNA levels, minimum 1000 TPM and max. going up to 10,000 TPM.

4.2.3.6.4 Analysis of all the nuclear and cytosolic antisense present across the five stages showed time and space dependent expression

Analysis of RNA-seq data showed that antisense RNAs form the highly abundant ncRNA class in the 64-cell cytosolic fraction (Fig 4.2.20), therefore, it can be hypothesized that the majority of these antisense RNAs must be amongst the maternally deposited RNAs. As one of the aims of this thesis was to identify and analyse antisense ncRNAs during the early stages of zebrafish development, this class of ncRNAs was investigated in detail. Antisense ncRNAs were divided according to their predominance in nuclear (log2 (Nuclear/Cytosolic) > 0.65) and cytosolic (log2 (Nuclear/Cytosolic) < -0.65) fraction across the five stages. The number of nuclear enriched antisense ncRNAs increase as development progressed (21 antisense in 64-Cell, 44 in 256-Cell, 48 in 1000-Cell, 142 in Dome and 196 in shield stage). In contrast, the number of cytosolic antisense ncRNAs decreased as development progressed. The number of cytosolic antisense ncRNAs found were 123 in 64-Cell, 64 in 256-Cell, 79 in 1000-Cell, 28 in Dome and 38 in the shield stage. Next the expression of both the nuclear and cytosolic antisense in the five stages was compared with their overlapping protein-coding partner genes.

The nuclear antisense in the 64-Cell stage were abundant with RNA levels greater than 60 TPM, however, their levels decreased (1000 Cell) and then increased again post zygotic genome activation (Dome and Shield) (Fig 4.2.21). The pattern observed is more indicative of both maternal and zygotic contribution for the nuclear antisense. The comparison of RNA levels between the cytosolic antisense and its overlapping mRNA showed anti-correlation with the cytosolic antisense being present in abundance in the early stages and then eventually lost in the shield (Fig

4.2.22). Thus, suggesting that majority of the cytosolic antisense are maternally deposited which are degraded with zygotic genome activation. The corresponding mRNAs, on the other hand, were either lowly expressed or absent in the stages. On comparison, the RNA expression pattern observed for cytosolic and nuclear antisense resembles the manner that the antisense in the negatively and positively correlated group (discussed in CHAPTER 4) express, respectively.

We further plotted these results using a violin plot for the expression of nuclear and cytosolic antisense (APPENDIX, Fig 3). We observed that the levels of nuclear antisense is higher in the 64-Cell stage (median of 75TPM) which then decreased by 1000-Cell stage but increased again in the shield stage. Thus, representing the results using a heatmap. The cytosolic antisense, on the other hand, showed the highest RNA levels in the 64-Cell (median of 75TPM, with majority of them being 100TPM) and then decreased in the shield stage (median of 25TPM) with MZT.

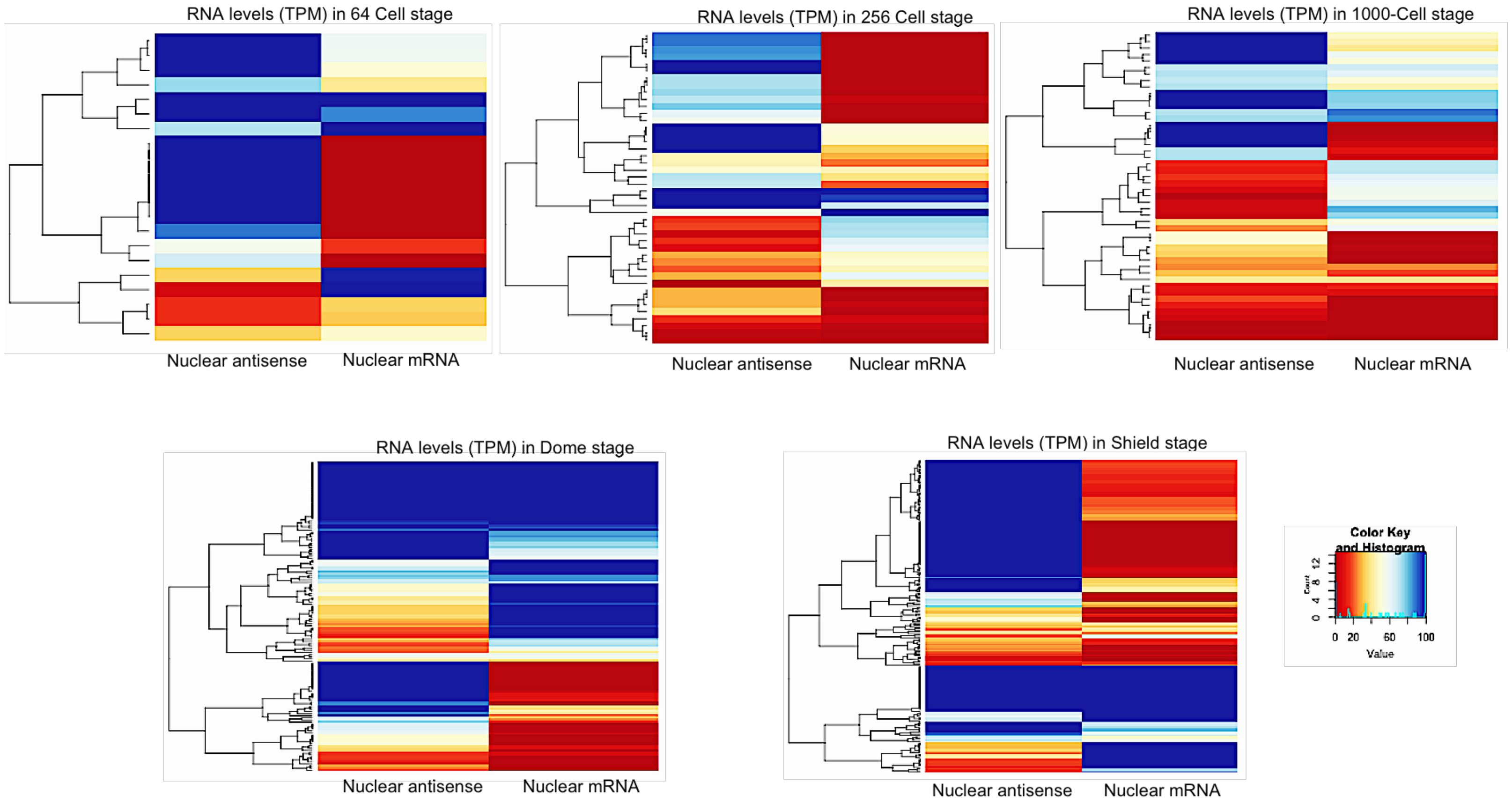


Figure 4.2.21 Nuclear antisense – mRNA pair RNA levels in the different stages of development.
heatmap showing the expression of nuclear antisense with respect to their overlapping mRNA genes in the nuclear fraction during embryogenesis. The colour key indicates the RNA levels with red being low or not present (0-20) and blue being maximum levels (60-100). Each row represents an antisense and its corresponding mRNA expression in the stages.

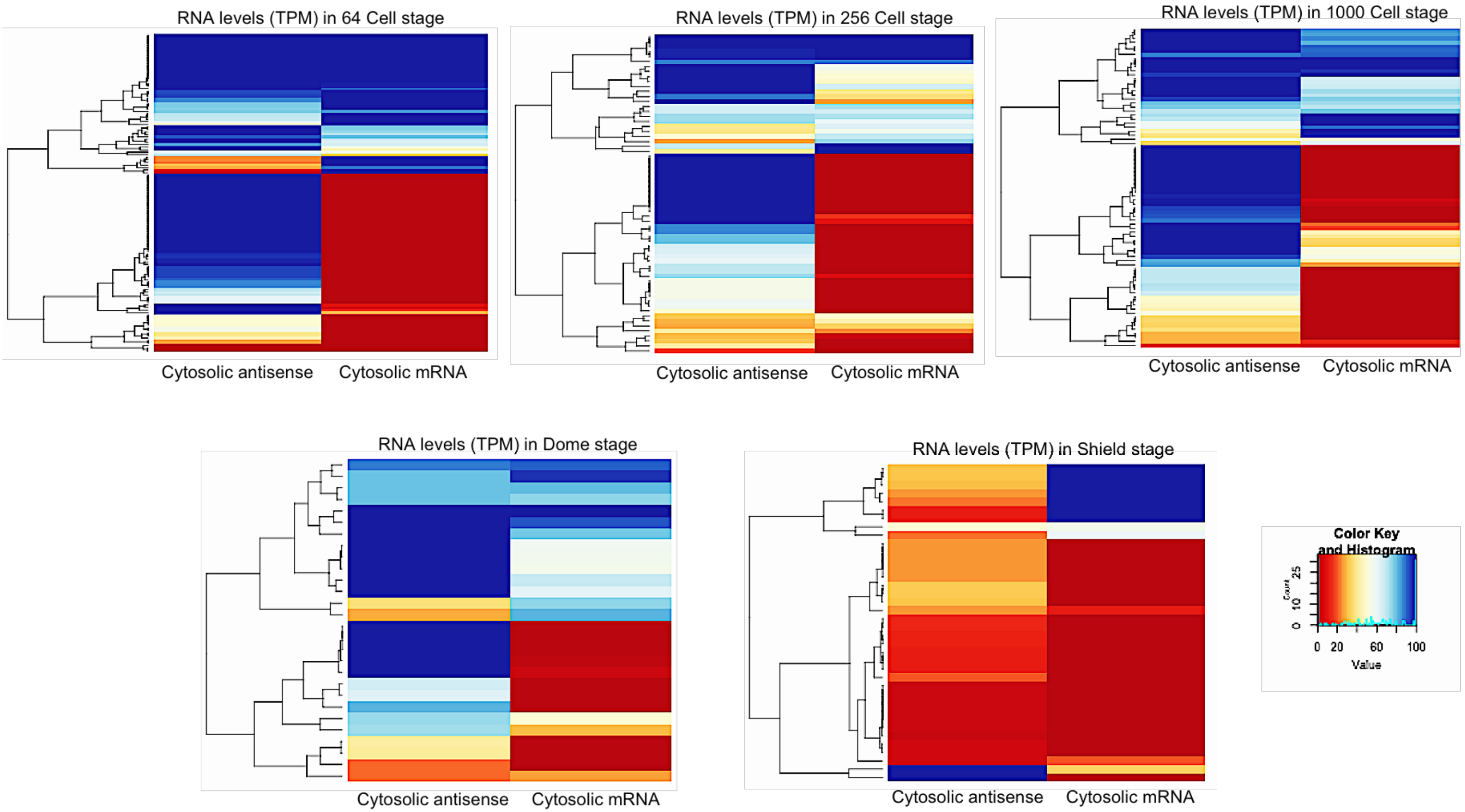


Figure 4.2.22 Cytosolic antisense – mRNA pair RNA levels in the different stages of development.
Heatmap showing the expression of cytosolic antisense with respect to their overlapping mRNA genes in the cytosolic fraction during embryogenesis. The colour key indicates the RNA levels with red being low or not present (0-20) and blue being maximum levels (60-100). Each row represents an antisense and its corresponding mRNA expression in the stages.

4.2.3.6.5 The average overlap region between the cytosolic antisense and their overlapping mRNA gene was comparatively higher in all the stages

To further investigate the extent of similarity between the nuclear and cytosolic antisense category with the class of antisense characterised (negatively and positively correlated) in the previous chapter, we calculated the overlap region amongst the two antisense ncRNA gene with their corresponding mRNA gene (Fig 4.2.23).

In general, all the cytosolic antisense had a higher overlap region with their protein-coding gene when compared to the nuclear antisense. The cytosolic antisense in the 1000-Cell stage showed the highest overlap region of ~17kbp. The overlap region between the cytosolic antisense and their mRNA genes showed an increase at MZT (~16kbp) but then decreased post-MZT (~10kbp) followed by a decrease in the number of antisense by the shield stage (Fig 4.2.23). The nuclear antisense in the shield stage, however, showed an overlap region of 12.5kbp which is the highest for nuclear antisense with an increase in the number of antisense (196) detected for this stage. On comparison the results hints towards the similarity between the antisense identified based on the expression correlation, negatively and positively correlated antisense (CHAPTER 3, Fig 3.2.9) and the antisense identified in the different subcellular compartment of the cell in this chapter.

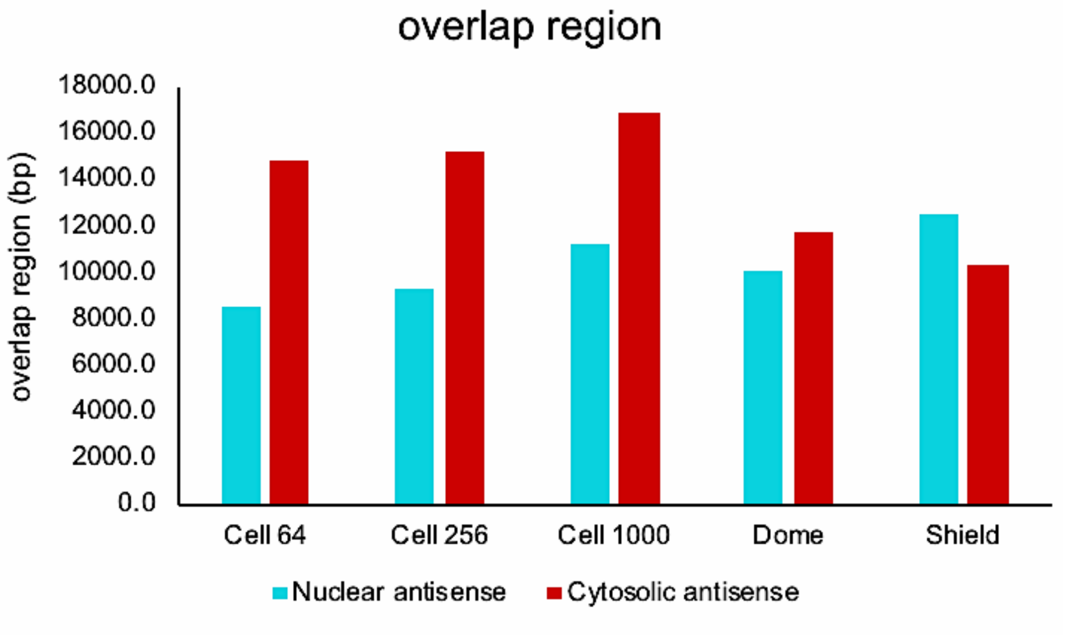


Figure 4.2.23 Average Overlap region.

The bar plot above shows the average overlap region between the cytosolic (blue) or nuclear (red) antisense and their overlapping protein-coding genes. The X-axis represents the overlap region in bp while the Y-axis shows the stages of development.

4.2.3.6.6 Majority of cytosolic antisense ncRNAs were found to be unique in the 64-Cell stage whereas most of the nuclear antisense were exclusive to the shield stage

Another pertinent question is whether the cellular localisation of antisense ncRNAs changes as development progresses. To address this question, nuclear and cytosolic RNA levels were used to identify the antisense ncRNAs that are either localised in nuclear or cytosolic compartments in the five developmental stages. This helped in understanding how antisense RNAs dynamics change in a time and space dependent manner (Fig 4.2.24).

64 antisense RNAs were exclusively identified in the 64-cell stage in both fractions. A large number (56) of these RNAs were cytosolic and only 8 were localised in the nucleus. The cytosolic localisation of pre-MZT antisense RNAs suggest maternal inheritance as there is no transcription going on at this stage. Whereas about 92 nuclear antisense were alone detected in the shield stage, proposing zygotic contribution after activation of zygotic genome (Fig 4.2.24). The high number of unique nuclear antisense and cytosolic antisense observed in the shield and 64-Cell stages, respectively, shows the spatial and temporal specificity of antisense expression.

We further looked at the gene ontology terms associated with the protein-coding genes overlapping the unique cytosolic antisense in the 64-Cell stage (56) and the mutually exclusive nuclear antisense in the shield stage (92) (APPENDIX, Fig 4).

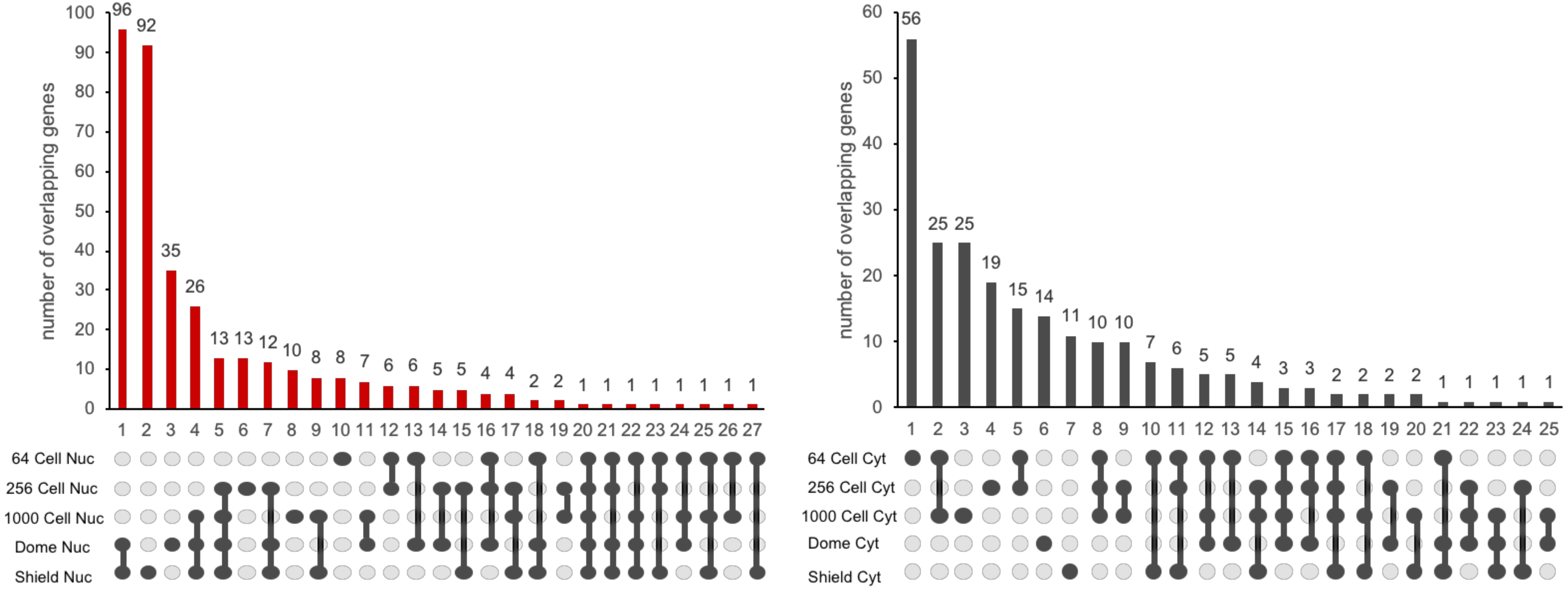


Figure 4.2.24 Antisense ncRNAs expressing in a space and time dependent manner.
The figure shows an overlap amongst the antisense present in the five stages (64Cell, 256Cell, 1000Cell, Dome and Shield) of development in the nuclear fraction (left) and the cytosolic fraction (right). Each bar on the plot represents the number of common or unique antisense (mentioned above the bar). The interactions shown between the dots indicates the stages that have common antisense and dots with no interaction suggest unique antisense (using UpSet visualisation).

4.2.3.6.7 Gene Ontology (GO) terms associated with the overlapping protein-coding gene in the 64-cell stage showed enrichment in developmental proteins

Antisense RNAs are known to regulate their expression of overlapping protein-coding genes (Faghihi and Wahlestedt, 2009). However, which kind of genes are regulated by antisense is not understood. To address this the functional categories of protein-coding genes overlapping cytosolic and nuclear antisense RNAs was analysed (Fig 4.2.25-4.2.26) using DAVID or Database for Annotation, Visualisation and Integrated Discovery tools (Huang da et al., 2009a, Huang da et al., 2009b).

The 123 protein-coding genes overlapping the cytosolic antisense in the 64-cell stage (Fig 4.2.25) showed a significant enrichment in transcription factor activity (p-value, 10^{-13}), regulation of transcription (p-value, 10^{-12}) and regulation of RNA metabolic process (p-value, 10^{-12}). The functions associated with protein-coding genes overlapping cytosolic antisense in the subsequent stages (256-Cell, 1000-Cell and Dome) remained same, albeit less significantly than 64-cell stage. However, in the shield stage we again observed a significant increase in the terms associated with transcription factor activity (10^{-14}) and regulation of transcription (10^{-12}).

The protein-coding genes overlapping nuclear antisense RNAs (Fig 4.2.26) showed enrichment for housekeeping genes phosphorous metabolic process (p-value, 10^{-2}) and protein kinase activity (with p-value, $10^{-1.5}$) in the 64-Cell and 1000-Cell stages. However, in the later stages like Dome and Shield we saw an enrichment for functions associated with transcription regulation (with p-value ranging from $10^{-1.5}$ to 10^{-3}).

This differences in the functions of genes overlapping cytosolic vs nuclear antisense RNAs suggests that antisense RNAs might use different regulatory pathways depending on the functions of genes.

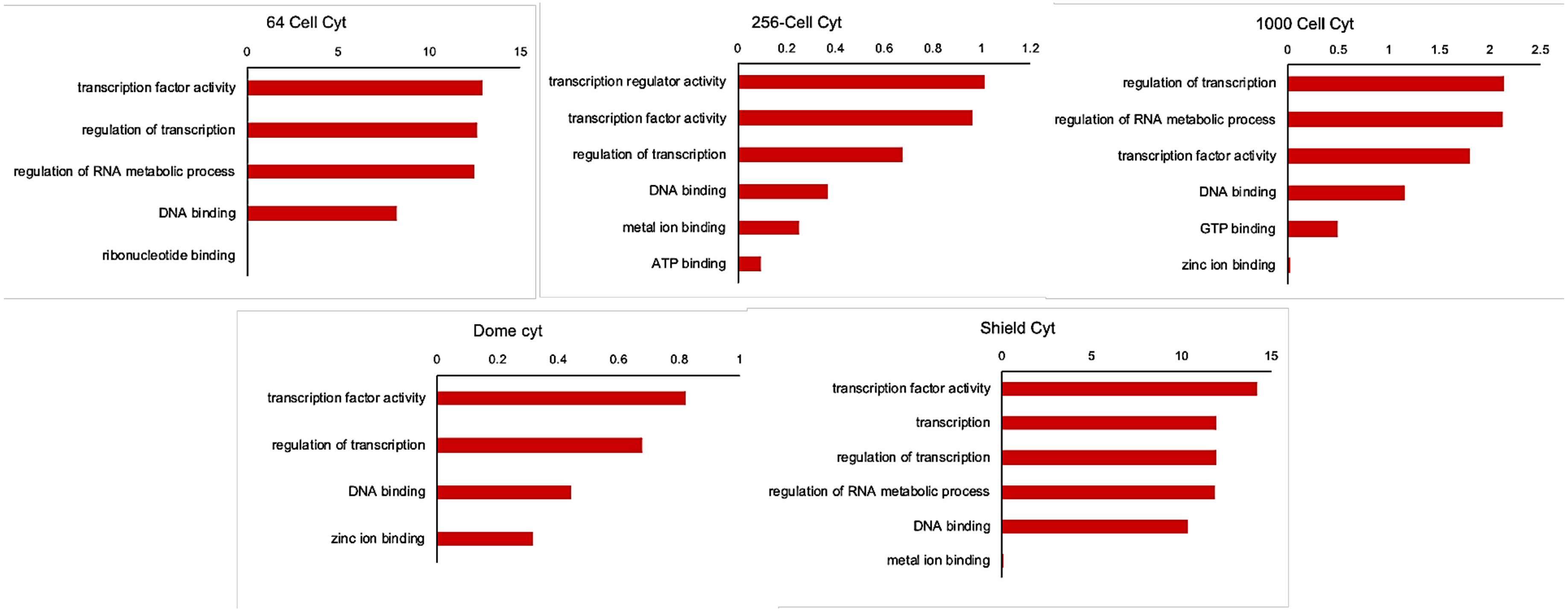


Figure 4.2.25 Gene ontology terms associated with mRNAs overlapping cytosolic antisense
The schematics shos the Gene Ontology terms associated with the protein-coding genes overlapping the cytosolic antisense in the 64-cell, 256-Cell, 1000-Cell, Dome and shield stages of development. The y-axis represents the different GO terms associated with the genes and the X-axis shows the negative log of p-value, the higher the number the more significant the term associated with the gene.

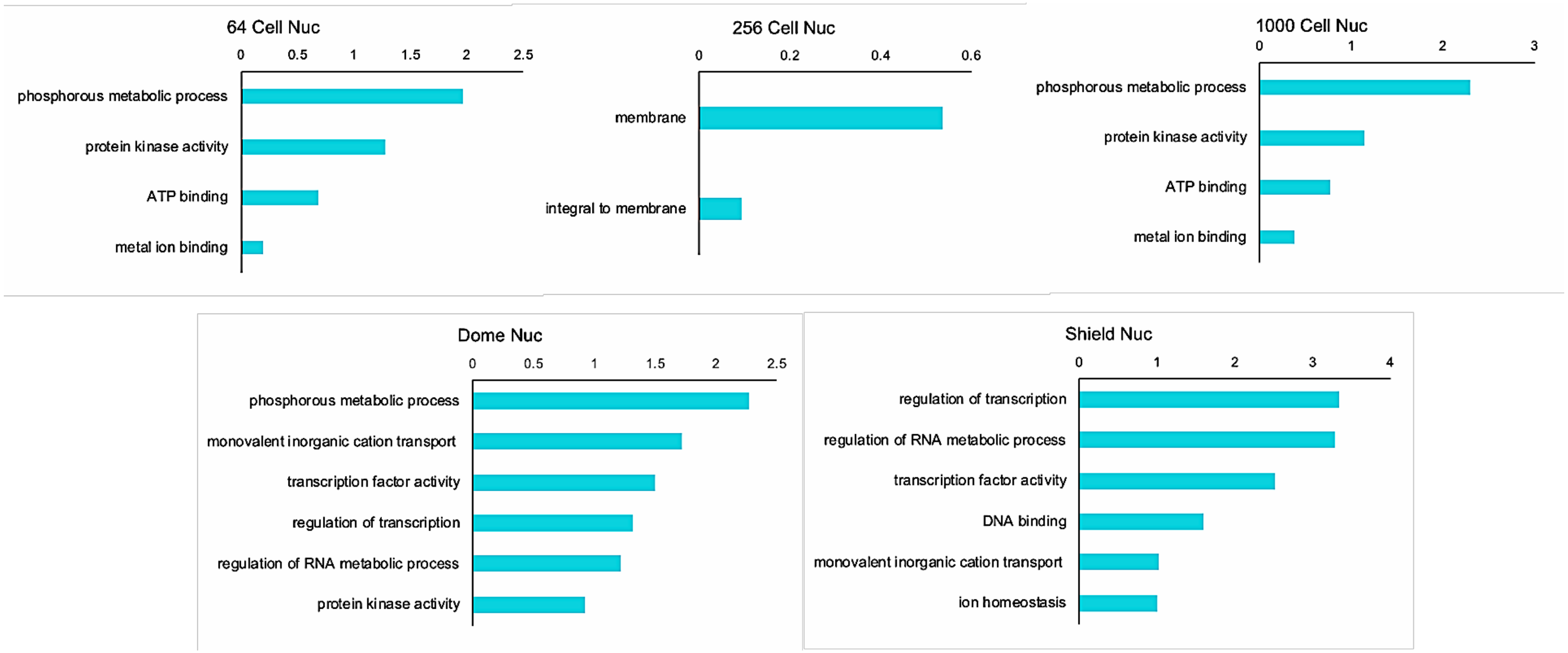


Figure 4.2.26 Gene ontology terms associated with mRNAs overlapping nuclear antisense
The schematics shows the Gene Ontology terms associated with different protein-coding genes overlapping the nuclear antisense in five stages of development. The y-axis represents the different GO terms associated with the genes and the X-axis shows the negative log of p-value, the higher the number the more significant the term associated with the gene.

4.3 Discussion

The aim of this study was to examine the cellular localisation of RNAs through maternal-to-zygotic transition during zebrafish development. We successfully fractionated the nucleus and cytosol from zebrafish embryonic cells at different stages of development (Fig 4.2.1). We established that our RNA-seq detects early stage ncRNAs with a higher depth (2-3 times) and quality (Fig 4.2.16 -4.2.18) than a previous study by Pauli et al. (Pauli et al., 2012). In addition, our nuclear fractions showed presence of unspliced or heterogenous transcripts with zygotic genome activation, thus, detecting early zygotic transcription.

The percentage of nuclear maternal RNA increased post-MZT suggesting zygotic contribution for these maternal RNAs because of their importance later in development (Fig 4.2.7). A similar set of analysis on mRNAs (Fig 4.2.8) also showed a rapid degradation of cytosolic mRNAs, suggesting maternal deposition and a rapid increase in the nuclear mRNA levels with ZGA proposing zygotic RNA expression (similar to (White et al., 2017)). In the analysis of the distribution of nuclear reads across the TSS of maternal and zygotic RNAs suggests an increase in the abundance of both maternal and zygotic RNAs after the ZGA (Fig 4.2.9-4.2.10). This is in accordance to what have been observed previously for maternal and zygotic RNAs (Mathavan et al., 2005). We did not observe any reads for the introns of maternal RNAs, both pre-MZT and post-MZT. However, we do see coverage for the introns of zygotic RNAs after maternal-to-zygotic transition suggesting their transcription.

Non-coding RNA dynamics. ncRNAs were found to be more enriched in the cytosol (maternally deposited) in the early stages of development but as zygotic transcription starts, they became more localized in the nucleus, with their percentage being much higher than nuclear mRNAs (Fig 4.2.13 – 4.2.14). The higher abundance of lincRNAs and antisense RNAs (Fig 4.2.20) both in terms of number and transcript levels in 64-cell stage indicates the importance of them in early zebrafish development, probably in the regulation of neighbouring protein-coding gene (Derrien et al., 2012). The higher abundance of snRNAs later on in development is expected owing to the importance of these ncRNAs in splicing the newly transcribed zygotic RNAs (Matera et al., 2007). Further analysis of the expression pattern of the antisense and overlapping mRNA genes in the two subcellular compartments showed pattern similar to the antisense class characterised (negatively and positively correlated) in the previous chapter. Majority of the cytosolic antisense in all the stages behave like maternally deposited RNAs while the nuclear antisense showed patterns of both maternal and zygotic contribution (Fig 4.2.21 – 4.2.22). The results from the GO analysis (Fig 4.2.25) showed the importance of maternal antisense as majority of them overlap protein-coding genes involved in development or transcription regulation. It has been previously observed in both mouse and human ESCs that divergent lncRNAs tends to be associated with protein-coding genes essential in transcription or development (Luo et al., 2016). Therefore, the maternal antisense ncRNAs identified in our study might be essential in keeping in check the expression of these important developmental genes.

Our nuclear RNA-seq data provides crucial information about the distribution of major RNAs before and after MZT. The data also raise many questions as to the function of 5% of mRNAs detected in the nucleus during the early stage when there is no transcriptional activation. As dramatic changes occur in the localisation of RNAs at each time course it also stresses further on the importance of these stages. We also observe a very high abundance of maternally deposited antisense ncRNAs in the cytosol overlapping major developmental genes and genes involved in transcription regulation (Fig 4.2.25 – 4.2.26). This further validates the theory we predicted in our previous chapter, a post-transcriptional mechanism of regulation of mRNA expression by overlapping antisense which would further explain why we see a higher overlap region amongst the cytosolic antisense compared to the nuclear antisense.

CHAPTER 5: IDENTIFICATION OF DIFFERENTIAL TSS USAGE AND PROMOTER ORGANISATION IN THE NUCLEAR AND CYTOSOLIC FRACTION DURING ZEBRAFISH EMBRYOGENESIS

5.1 Introduction

Transcription initiation is a very complex process which requires the presence of a core promoter in proximity of the transcription start sites (TSS). This core promoter helps in the recruitment of the general transcription factors (GTFs) which, in turn, recruit's RNA Pol II to the TSS. Recent studies have shown the presence of unconventional promoter usage, apart from TATA-box, in genes essential for zebrafish development (Haberle et al., 2014). This diversity in the usage of promoters may result in differential regulation of transcription in the context of development. Cap Analysis of Gene Expression (CAGE) is a method for transcriptome analysis about changes in TSS and its relative usage at single nucleotide resolution (Kodzius et al., 2006). CAGE gives us the information for the start sites of capped RNAs which in turn can be used as an indicator of promoter organisation. Apart from the differential TSS usage, CAGE also gives evidence about the expression or RNA levels of a gene. A recent study on zebrafish have identified differential promoter usage as well as the presence of a novel initiator sequence during development. The study showed the presence of several classes of post-transcriptionally cleaved RNAs essential during development (Nepal et al., 2013).

In the previous chapter we analysed nuclear and cytosolic RNA-seq data to get insight into the localization of protein-coding (maternal & zygotic) and non-coding genes with zebrafish development. We wanted to further use this nuclear and cytosolic fractions, from zebrafish embryonic cells, to help understand the differences in TSS as well as promoter organisation with regard to RNA localisation in cellular compartments and how it changes with development. Another objective was to accurately annotate nuclear localised non-coding RNAs such as enhancer RNAs by combining the nuclear RNA-seq and CAGE-seq datasets. We expected that sequencing nuclear fraction will enrich these RNAs and make it easier to detect them. In addition, sequencing of CAGE-tags in addition to RNA-seq data will help in providing robust annotation of many lowly expressed RNAs as well as reaffirming the results we obtained from RNA-seq.

5.2 Results

5.2.1 Detection of capped transcripts in the nuclear fraction across the zebrafish developmental stages

The conventional method used for 5'-cap analysis, also called cap trapping method (Carninci et al., 1996), targets all capped mRNAs by chemical modification of the 5' end (cap) followed by its biotinylation and then a pulldown using streptavidin-coated beads. We used a modified version of cap trapping called Low Quantity CAGE or LQ-CAGE for our CAGE library preparation. This method allowed use of limited quantities of RNAs as is generally obtained in early stages of development. Therefore, more useful over the canonical cap trapping method. We prepared libraries for 6 different stages of zebrafish development, 64 Cell, 256 Cell, 1000 Cell, Dome, Shield and Prim5. In this chapter I will be discussing the results we got from this LQ-CAGE sequencing. A total of ~211 million reads per replicate was generated from Illumina sequencing out of which on an average ~8.5 million reads per stage were mapped uniquely to the Zv9 genome (Table 5.2.1).

Table 5.2.1 CAGE-seq alignment summary
Showing the raw tag count after quality control (trimming) and the uniquely mapped reads for the above 6 stages in the nuclear and cytosolic fractions.

Fraction	Stage	Raw tags	Uniquely aligned	% aligned tags
Nuclear	64 Cell	2,602,673	1,711,474	65.7583185
	256 Cell	9,331,397	5,949,490	63.7577632
	1000 Cell	11,901,032	7,053,997	59.2721455
	Dome	10,462,935	6,612,202	63.1964358
	Shield	17,301,622	11,250,967	65.0283944
	Prim5	3,613,362	2,704,757	74.8543047
Cytosolic	64 Cell	1,704,017	1,228,792	72.1114872
	256 Cell	7,439,214	5,247,970	70.5446839
	1000 Cell	22,639,185	15,369,134	67.8873113
	Dome	29,823,212	21,458,222	71.9514115
	Shield	88,335,616	67,068,983	75.9251885
	Prim5	6,577,256	4,764,835	72.4441165

5.2.2 CAGE-seq dataset suggested a decrease in the correlation among the fractions in the same stage after MZT

To get an idea about the relation of samples with one another vis-à-vis zebrafish development, a correlation plot was generated. This would help us in understanding how the fractions from the same stage correlate with each other as compared to the other stages. Using CAGEr software, the expression of samples were compared to each other using scatter plots of tag counts per TSS and extent of correlations between all sample pairs was calculated with the help of correlation coefficient. We observed (Fig 5.2.1) that initially the correlations between the nuclear and cytosolic fractions from the same stage show high correlation (greater than 0.90). However, after shield we see a decrease in the correlation coefficient suggesting differences in the expression of transcripts or a more precise distribution of RNAs according to their biological functions. This can also be due to the fact that later on in development the transcripts are more stage specific and do not show any correlation with the early expressing transcripts.

This is well reflected in both the fractions in the Prim5 (24hpf) showed very low correlation coefficient (less than 0.5) with all the other fractions in the different stages.

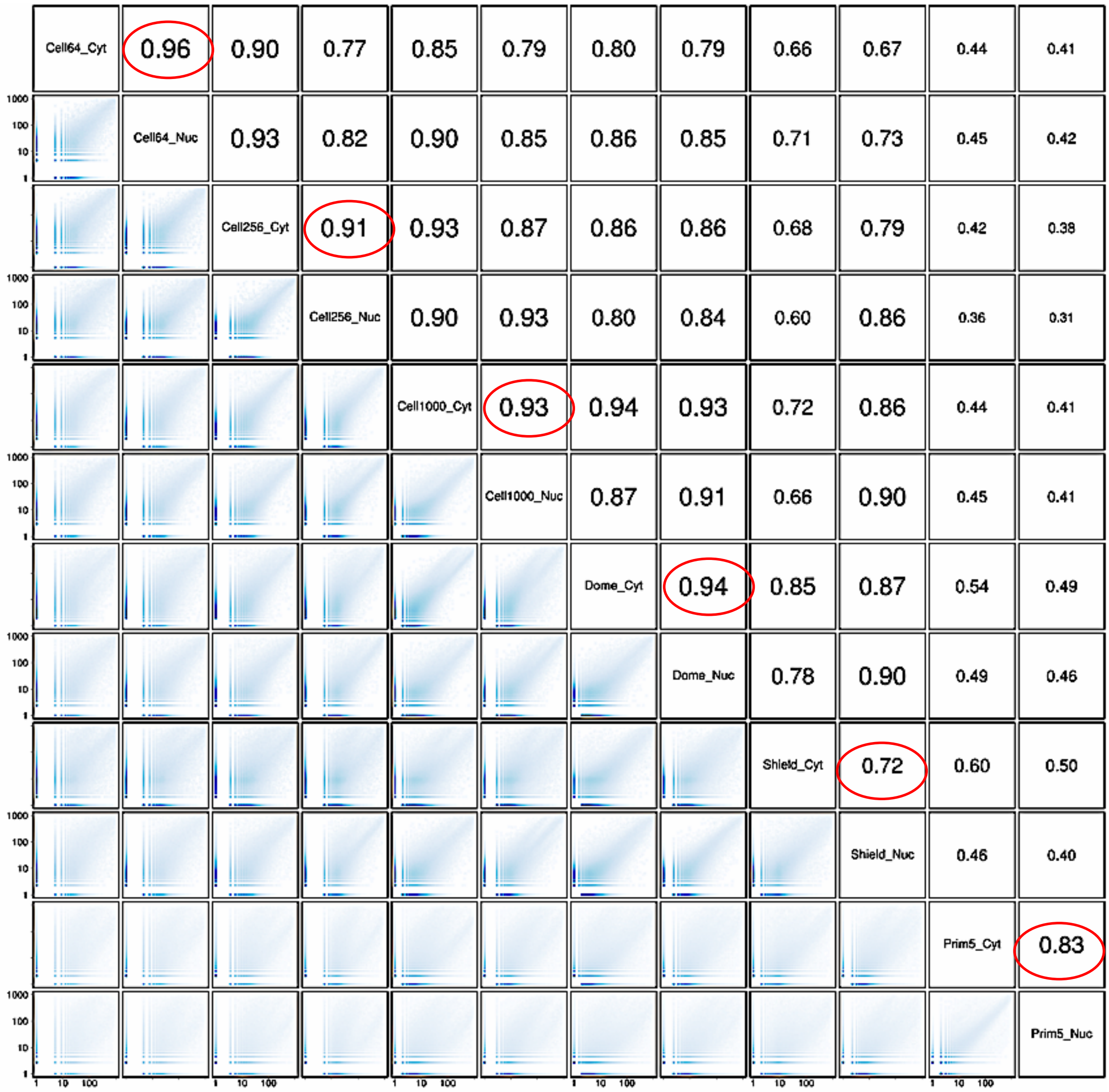


Figure 5.2.1 Pearson correlation coefficient.

The lower triangle shows scatter plots among the different stages based on the CAGE tag counts. The diagonal represents the names of different stages and fractions. The numbers in the upper triangle denotes the correlation coefficient. The red circles indicate the correlation between the fractions from the same stage. Both the X-axis and Y-axis represent the 12 samples in the study (12 X 12 box).

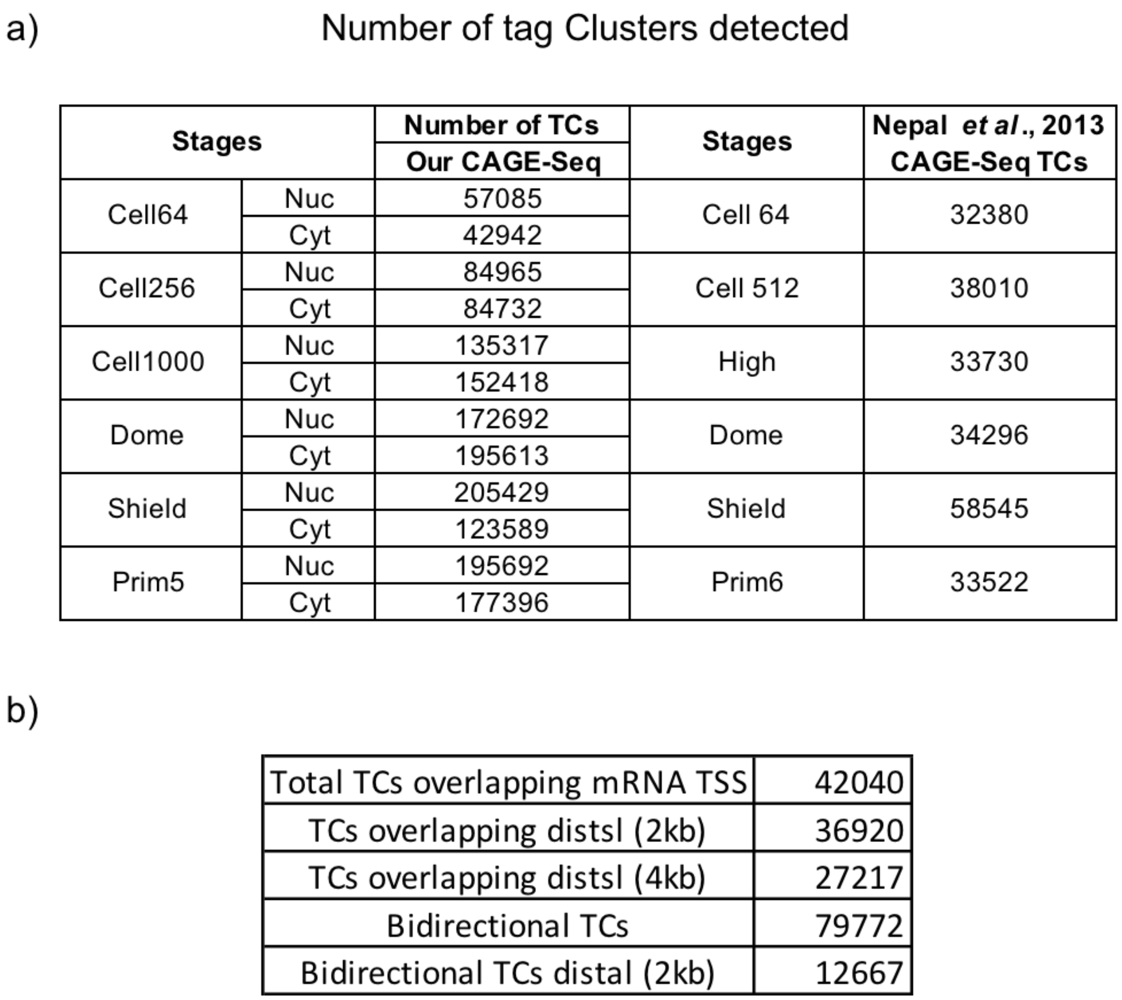
5.2.3 Normalization of tag counts and CTSS clustering into tag clusters

Since the raw tag counts per sample vary in number in order to compare them, the tag counts were normalized (Fig 2.7.1). With the help of CAGEr package, a reverse cumulative distribution was plotted where number of CAGE TSSs (Y-axis) that have equal or greater than the number of tags (X-axis). This further assists us in choosing the appropriate parameters for normalization of dataset.

To estimate the promoter width, individual CAGE transcription start sites (CTSS), in a close proximity, are clustered together. Neighbouring TSSs in close proximity to each other are most likely associated with the same promoter element. Multiple CAGE-seq tags that are within 20bp are clustered together into tag clusters (TCs). For each cluster genomic coordinates, the number of CTSSs in the cluster and in the dominant CTSS in tags per million (tpm) was noted. We further compared the number of TCs detected in our study with a previous CAGE-seq study (Nepal et al., 2013) on zebrafish development (Fig 5.2.2). The CAGE TCs detected in either the nuclear or cytosolic fractions were higher than the TCs detected in Nepal et al. study (Nepal et al., 2013) in all the stages. The number of TCs identified in our CAGE-seq was more than three-fold higher than the Nepal, C CAGE-seq (Nepal et al., 2013) in the stages after zygotic genome activation. We also looked for tag clusters distal to mRNA TSS (2kb and 4kb) in our study suggesting the presence of enhancer RNAs identified by nuclear CAGE. This has been further discussed in section 5.2.9.

5.2.4 Differences in the distribution of CTSSs within a tag cluster

An example showing the differences observed in the TSS position from the nuclear to cytosolic fraction across the stages is shown in Fig 5.2.3. It shows changes in the distribution of CTSSs within a tag cluster at a particular gene as development progresses. This example (Fig 5.2.3) clearly showed that the interquantile width or the peak width as well as the position of the dominant TSS changes with progress embryonic development. The shift in the position of dominant TSS in relation to RNA localization was also obvious in the different fractions of the cell for all the six stages of development. The *nvl* gene considered here revealed variations in the promoter width as well as TSS position in the nuclear and cytosolic fraction in the 64-Cell and 256-Cell stages, with the 0_2 CTSS cluster being dominant in the nuclear fraction. However, the cytosolic fraction in the two stages showed very different arrangement of promoter and TSS. Although the promoter width was similar within the fractions in the same stage after zygotic genome activation, the position of the dominant TSS varied.



a) Number of tag Clusters detected

Stages		Number of TCs Our CAGE-Seq	Stages	Nepal *et al*., 2013 CAGE-Seq TCs
Cell64	Nuc	57085	Cell 64	32380
	Cyt	42942		
Cell256	Nuc	84965	Cell 512	38010
	Cyt	84732		
Cell1000	Nuc	135317	High	33730
	Cyt	152418		
Dome	Nuc	172692	Dome	34296
	Cyt	195613		
Shield	Nuc	205429	Shield	58545
	Cyt	123589		
Prim5	Nuc	195692	Prim6	33522
	Cyt	177396		

b)

Total TCs overlapping mRNA TSS	42040
TCs overlapping distsl (2kb)	36920
TCs overlapping distsl (4kb)	27217
Bidirectional TCs	79772
Bidirectional TCs distal (2kb)	12667

Figure 5.2.2 CAGE sequencing statistics.

***a)** The table above outlines the number of CAGE TCs detected in our CAGE compared to a previously existing CAGE-seq dataset during zebrafish development (Nepal et al., 2013). In cases where we could not find a matching stage we chose a stage before or after. **b)** the table represents the number of tag clusters observed overlapping mRNA TSS, 2kb and 4kb distal of mRNA TSS, bidirectional clusters (or enhancers) and TCs 2kb distal to bidirectional clusters.*

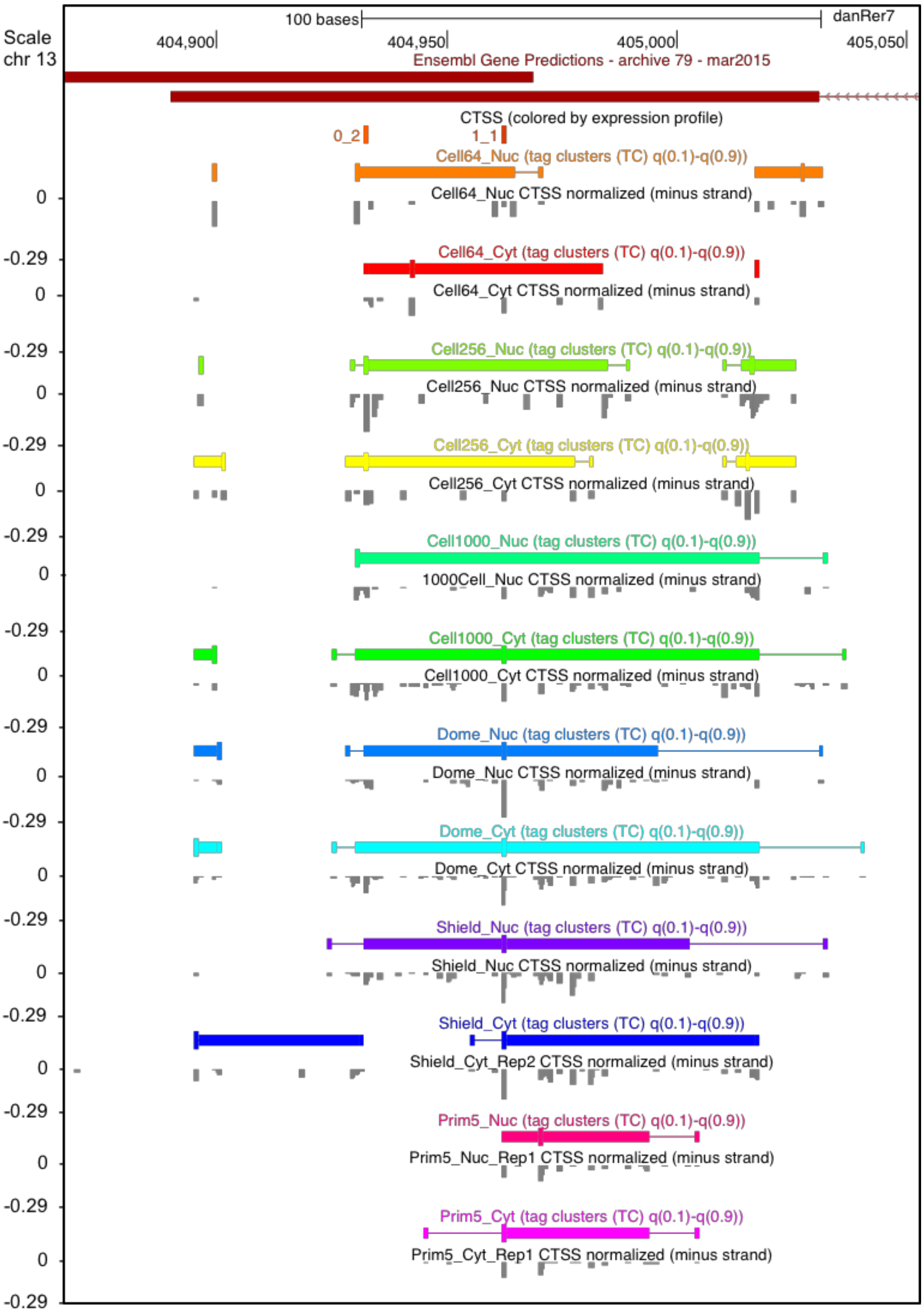

Scale
chr 13
100 bases
danRer7
404,900
404,950
405,000
405,050
Ensembl Gene Predictions - archive 79 - mar2015
CTSS (colored by expression profile)
0_2
1_1
Cell64_Nuc (tag clusters (TC) q(0.1)-q(0.9))
Cell64_Nuc CTSS normalized (minus strand)
Cell64_Cyt (tag clusters (TC) q(0.1)-q(0.9))
Cell64_Cyt CTSS normalized (minus strand)
Cell256_Nuc (tag clusters (TC) q(0.1)-q(0.9))
Cell256_Nuc CTSS normalized (minus strand)
Cell256_Cyt (tag clusters (TC) q(0.1)-q(0.9))
Cell256_Cyt CTSS normalized (minus strand)
Cell1000_Nuc (tag clusters (TC) q(0.1)-q(0.9))
1000Cell_Nuc CTSS normalized (minus strand)
Cell1000_Cyt (tag clusters (TC) q(0.1)-q(0.9))
Cell1000_Cyt CTSS normalized (minus strand)
Dome_Nuc (tag clusters (TC) q(0.1)-q(0.9))
Dome_Nuc CTSS normalized (minus strand)
Dome_Cyt (tag clusters (TC) q(0.1)-q(0.9))
Dome_Cyt CTSS normalized (minus strand)
Shield_Nuc (tag clusters (TC) q(0.1)-q(0.9))
Shield_Nuc CTSS normalized (minus strand)
Shield_Cyt (tag clusters (TC) q(0.1)-q(0.9))
Shield_Cyt_Rep2 CTSS normalized (minus strand)
Prim5_Nuc (tag clusters (TC) q(0.1)-q(0.9))
Prim5_Nuc_Rep1 CTSS normalized (minus strand)
Prim5_Cyt (tag clusters (TC) q(0.1)-q(0.9))
Prim5_Cyt_Rep1 CTSS normalized (minus strand)
0
-0.29

Figure 5.2.3 Distribution of CTSSs within a tag cluster in nvl gene.
Individual peaks represent CTSSs clustered together into TCs based on the distance (less than 20bp apart). The coloured lines above the peak tracks indicate the interquantile width (q_{up} -q_{low}) and also the position of the dominant TSS. The interquantile width is an estimate of promoter width and represents a single promoter element regulating multiple TCs.

5.2.5 Promoter width increased after zygotic genome activation

Previous analysis of CAGE TSS distribution (Carninci et al., 2006) revealed the presence of different types of promoters based on the cumulative distribution and intensity of tag clusters. Broad promoters with dispersed TSS show presence of several CTSSs distributed over a broad region and are usually associated with developmentally regulated genes and a high GC content or CpG islands. On the other hand, sharp promoters have majority of their CTSSs concentrated at one dominant TSS and are usually associated with TATA-box motifs (Haberle et al., 2015). Thus, promoter width is a useful feature in understanding the regulation of a gene by its promoter. Interquantile width is used to estimate the promoter width and is calculated by measuring the space between two quantiles of the total CAGE signal (shown in Fig 2.7.2). To study the distribution of promoter width across the six nuclear and cytosolic fractions with zebrafish development, interquantile width for each stage was plotted (Fig 5.2.4-5.2.5). The width distribution showed that, prior to MZT, promoters were sharper with no differences in the nuclear and cytosolic fractions while during and after MZT i.e. in the 1000-Cell, Dome and Prim5 cytosolic fraction the promoters were broader relative to the nuclear fraction. However, in the Shield stage both nuclear and cytosolic fractions showed similar promoters.

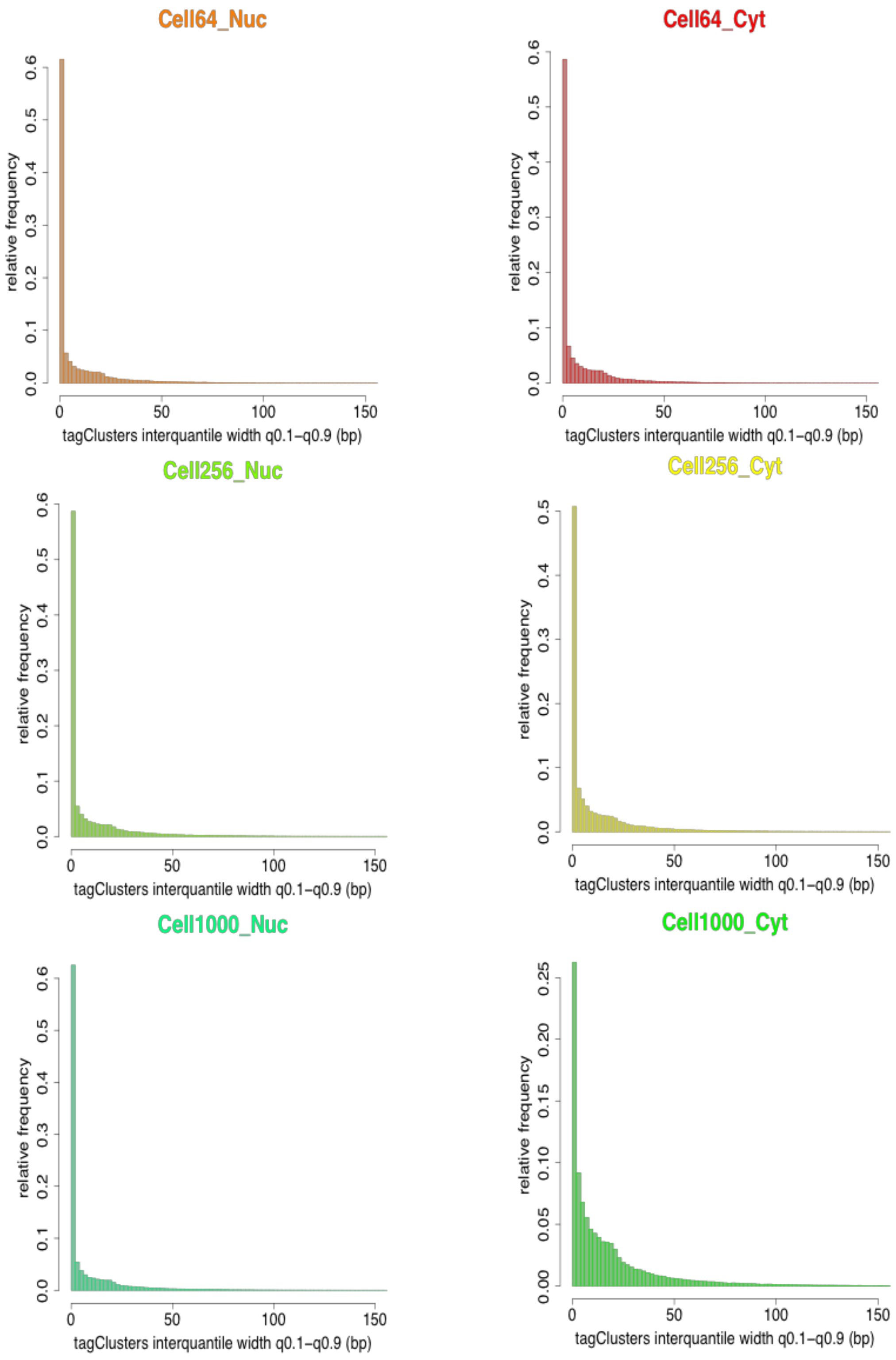


Figure 5.2.4 Promoter width across the stages in the fractions.
Histograms showing promoter width in the nuclear and cytosolic fractions across three stages of development (64Cell, 256Cell and 1000Cell). The X-axis represents the interquantile width (q0.1 -q0.9) in bp and the Y-axis is the relative frequency of CTSSs within 150bp of start sites of genes.

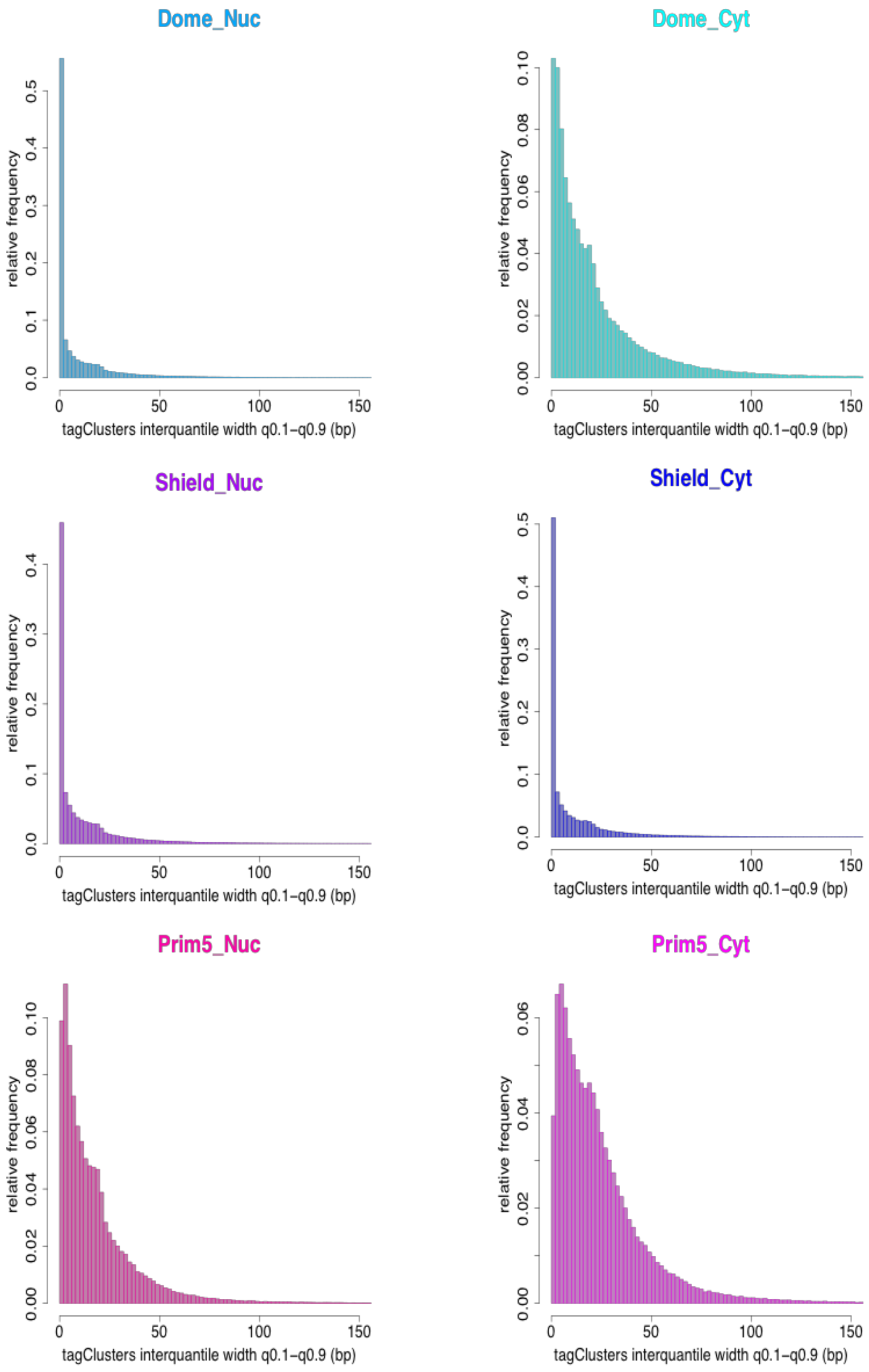


Figure 5.2.5 Promoter width across the stages in the fractions (post-MZT). *Histograms showing promoter width in the nuclear and cytosolic fractions across three stages of development (Dome, Shield and Prim5). The X-axis represents the interquantile width (q0.1 -q0.9) in bp and the Y-axis is the relative frequency of CTSSs within 150bp of start sites of genes.*

5.2.6 Expression profiles of CTSSs with zygotic genome activation

CAGE can also be used to measure the expression from individual CTSS or promoters. Expression profiling can be done either at the level of individual CTSSs or on entire promoters. The former is done by considering CAGE signal at individual CTSSs across the different stages while in the latter case CAGE signal from an entire consensus cluster (promoter) is used. A consensus cluster is constructed by aggregating all the overlapping TSSs and any neighbouring TSSs within defined proximity of each other. Self-organising maps or SOMs were used to cluster individual CTSSs across the different stages of development (shown in Fig 5.2.6).

The changes in the expression of CTSSs from the pre-MZT stages (64 Cell and 256 Cell) to the post-MZT stages (Dome, Shield and Prim5) was unique to each cluster. Clusters 1_1, 2_2, 4_1, 4_2, 4_3 and 4_4 showed difference in the nuclear and cytosolic fractions within the same stage (post-MZT). Clusters 1_1 and 4_3 contains CTSSs with different expression levels in the nuclear and cytosolic fractions across all the six stages under study. However, the differences observed between the fractions were only prominent after the 256-Cell stage, in other words with MZT. We also observed CTSSs specific for maternal transcripts (0_4, 1_4 and 2_4), highest expression before MZT, or zygotic CTSSs (4_0, 3_0 and 2_0) with higher expression in the later stages. Clusters 0_0 and 1_0 peaked during the maternal-to-zygotic transition while cluster 2_3 contained CTSSs expressing throughout.

These clusters can be visualized on the genome browser by colouring the CTSSs according to the expression class they belong to (shown in Fig 5.2.7). The figure showed a genome browser view of a gene, *galnt2*, with two distinct start sites from the 64 Cell stage (pre-MZT) to the prim-5 stage (post-MZT) belonging to 4 different expression cluster. Further (Fig 5.2.7) showed an example of *pprc1* gene having TSSs belonging to 14 different expression clusters (coloured and labelled) but part of the same promoter.

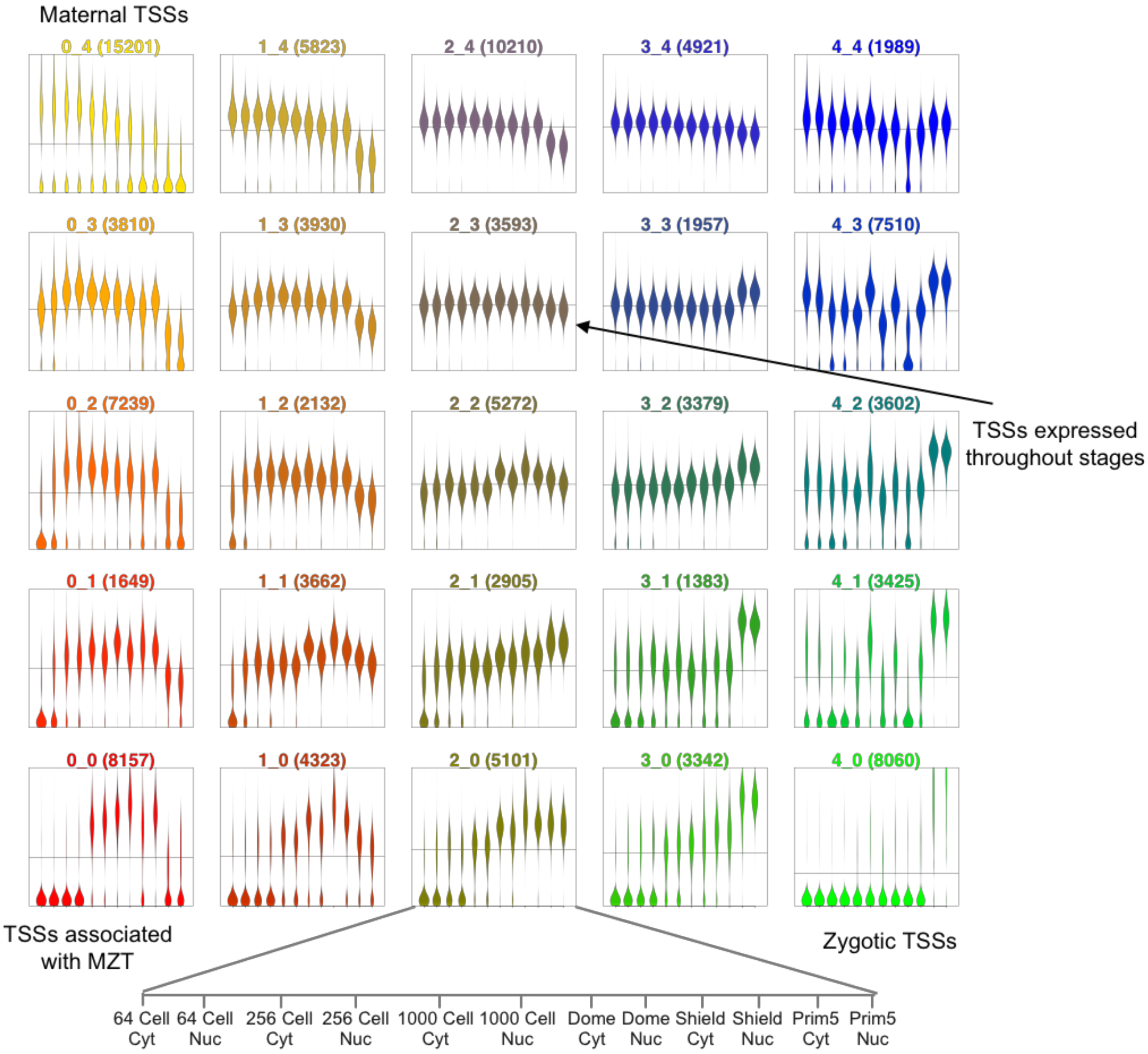


Figure 5.2.6 CTSSs Expression profiling with MZT.

Each box represents a different expression cluster and the figure denoted in bracket above each box is the number of TSSs plotted. Each beanplot within a cluster represents scaled normalized expression of CTSSs across the six stages in the different fractions (total 12) as denoted in the X-axis.

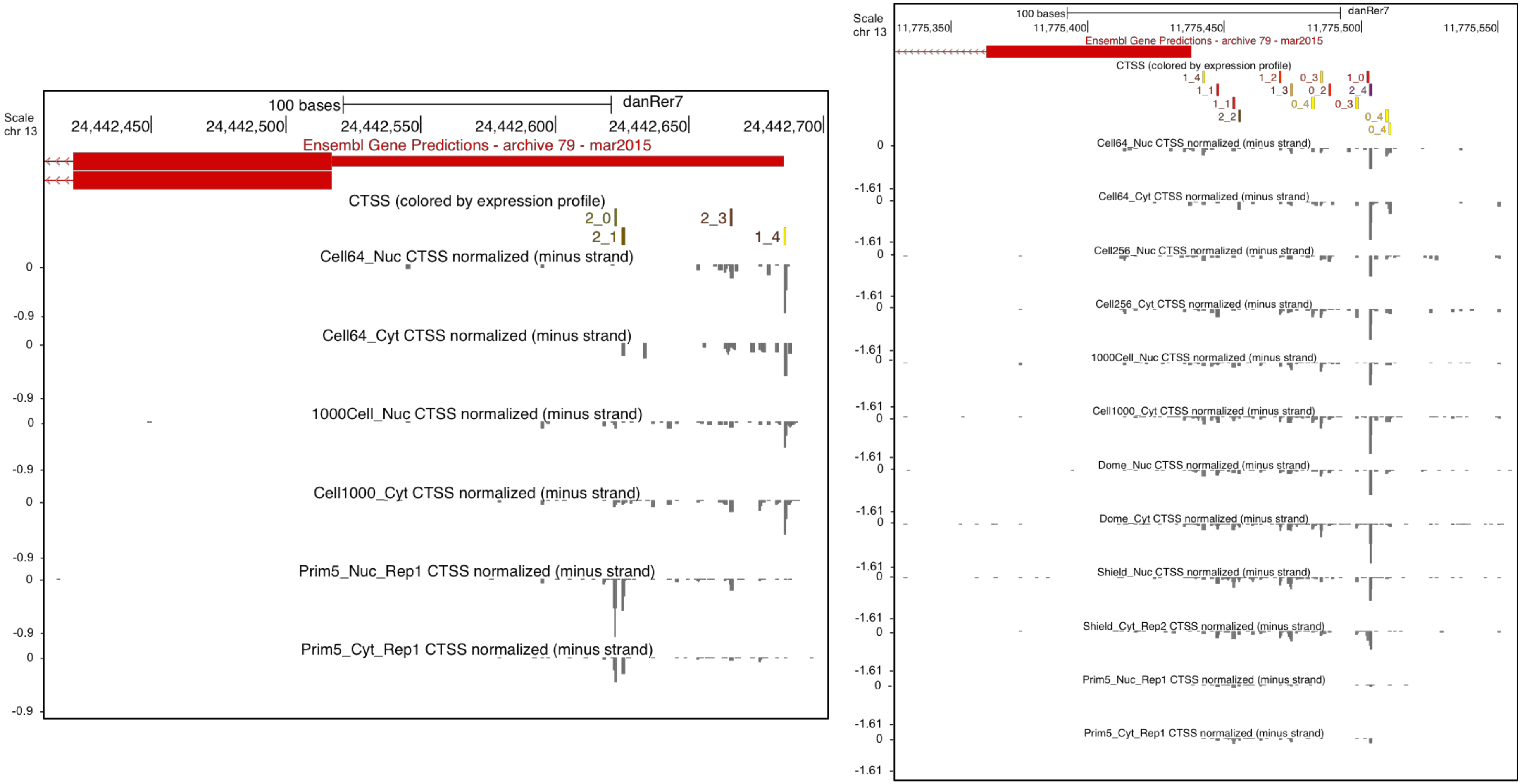

Scale
chr 13
100 bases
danRer7
24,442,450
24,442,500
24,442,550
24,442,600
24,442,650
24,442,700
Ensembl Gene Predictions - archive 79 - mar2015
CTSS (colored by expression profile)
2_0
2_1
2_3
1_4
Cell64_Nuc CTSS normalized (minus strand)
Cell64_Cyt CTSS normalized (minus strand)
1000Cell_Nuc CTSS normalized (minus strand)
Cell1000_Cyt CTSS normalized (minus strand)
Prim5_Nuc_Rep1 CTSS normalized (minus strand)
Prim5_Cyt_Rep1 CTSS normalized (minus strand)
0
-0.9
11,775,350
11,775,400
11,775,450
11,775,500
11,775,550
1_0
1_1
1_2
1_3
0_2
0_3
0_4
2_2
2_4
Cell256_Nuc CTSS normalized (minus strand)
Cell256_Cyt CTSS normalized (minus strand)
Dome_Nuc CTSS normalized (minus strand)
Dome_Cyt CTSS normalized (minus strand)
Shield_Nuc CTSS normalized (minus strand)
Shield_Cyt_Rep2 CTSS normalized (minus strand)
-1.61

Figure 5.2.7 Genome browser view of expression profiles.
The *galnt2 gene* (left) shows a partial shift in the position of the TSS with each TSS belonging to a different cluster based on expression levels. The *pprc1* gene (right) shows the presence of 14 different expression clusters within the same promoter region. The track above the CAGE expression in different stages is the track for CTSSs coloured by expression.

5.2.7 Differential promoter usage detected by CAGE during zebrafish embryonic development

During early embryogenesis of Zebrafish, until the tenth cell cycle, development of the embryo is dependent on maternally provided RNAs and proteins. With the activation of the zygotic genome, as transcription of zygotic RNAs is initiated, majority of maternal RNAs undergo degradation. To understand the changes in the usage of promoter between maternal RNAs and zygotic RNAs, ‘shifting’ promoter patterns were detected in our CAGE dataset. Although the overall expression levels of a promoter may not change between the different stages, the usage of a different TSS may suggest differences in regulation of transcription which cannot be detected in expression profiling. Using CAGEr package, shifting score of promoters was calculated by comparing the cumulative distribution of CAGE signal within a consensus promoter among the different stages (highlighted in Fig 2.7.3). For example, a shifting score of 0.6 indicates that 60% of the CAGE signal detected in a stage is occurring outside the region that was used as TSS (within the same promoter) in another stage. A shifting score of 0.6 was used to detect differential TSS usage between the 64-Cell to 1000-Cell stage (Pre-MZT) and the 1000-Cell to the Prim5 stage (Post-MZT).

On observing the genes on genome browser, there was a significant difference between the promoter usage from 64-Cell to Prim5 stage (also shown in Fig 5.2.7). A number of CTSSs revealed either a partial loss of a broad promoter (*chchd1* gene) with a single dominant TSS at the Prim-5 stage or a partial gain of a different dominant TSS in the Prim5 stage (*eif3s10* gene) (Fig 5.2.8a). We also found differences in the usage of dominant TSS within the same promoter between the nuclear and cytosolic fractions (*trmt61a, mcph1, kif14, cuedc2,* etc) which are further discussed in the next section 5.2.7.1. We observed a complete shift in TSS usage (*dag1* gene) in the later stages of development (Prim5) with the transcript expressing more than 2kb upstream of the annotated TSS (Fig 5.2.9). These shifts in the TSS position with zygotic genome activation suggests the presence of differences in promoter usage from maternal-to-zygotic transcripts. The organisation of the promoter in the early stages reflects maternal inheritance. With zygotic genome activation the appearance of broad promoters shows zygote specific promoter organisation. The analysis also revealed the presence of intragenic TSS, e.g., within the second intron of *cpne5* gene, instead of the canonical promoter lying upstream of gene start site (Fig 5.2.9).

We also discovered different types of promoter usage before MZT (64-cell to 1000-Cell stage) and after MZT (1000-Cell to Prim5 stage). A total of 8927 incidents of shifting promoters was observed post-MZT and 3827 incidents of promoter shifting before the zygotic genome activation.

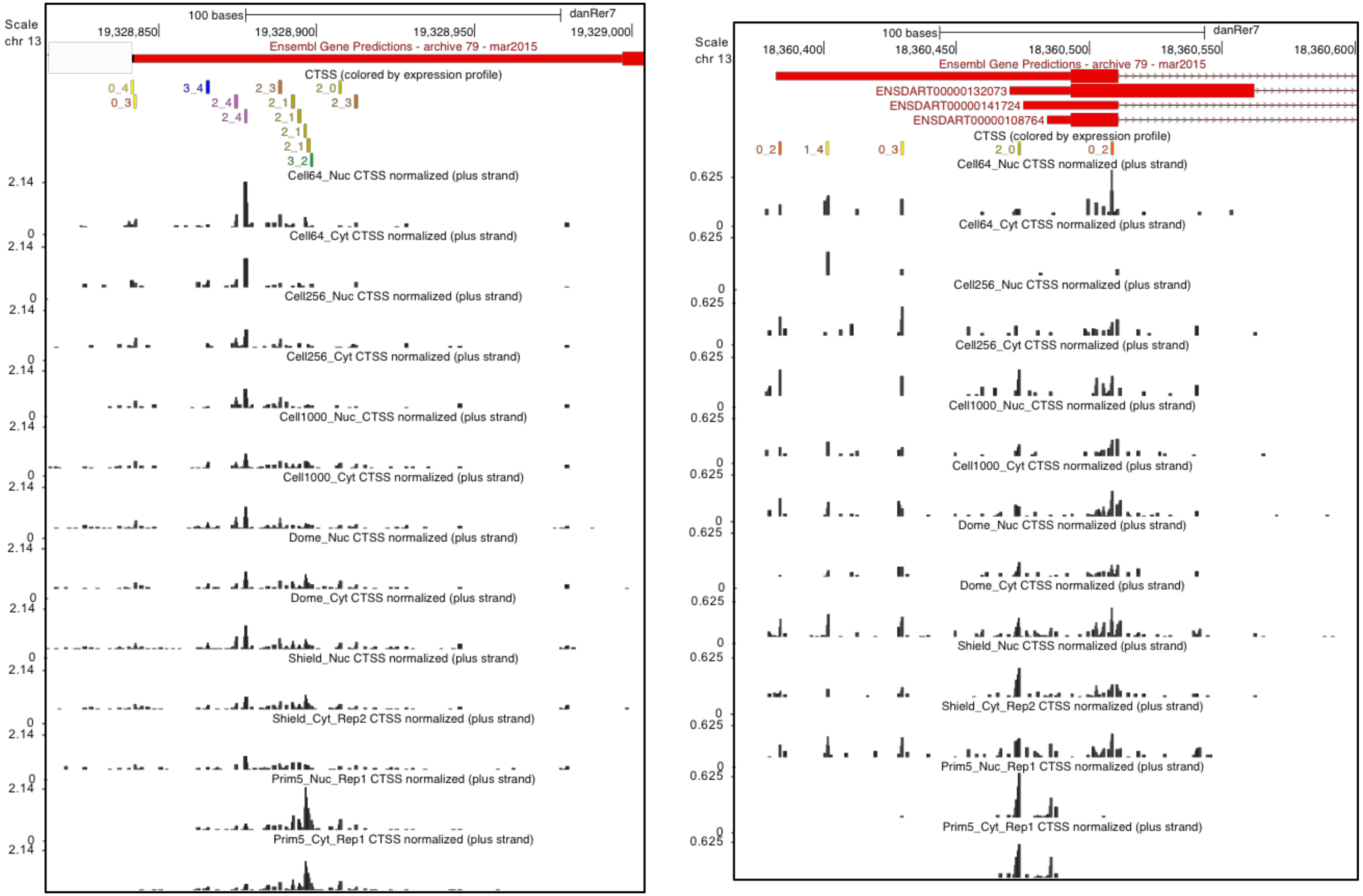


Figure 5.2.8 Differential promoter usage detected by CAGE-seq.
It shows on the left-hand side panel a partial gain of TSS within the same promoter region of *eif3s10* gene. The right-hand side panel shows a partial loss of TSS with narrowing of the promoter for *chchd1* gene.

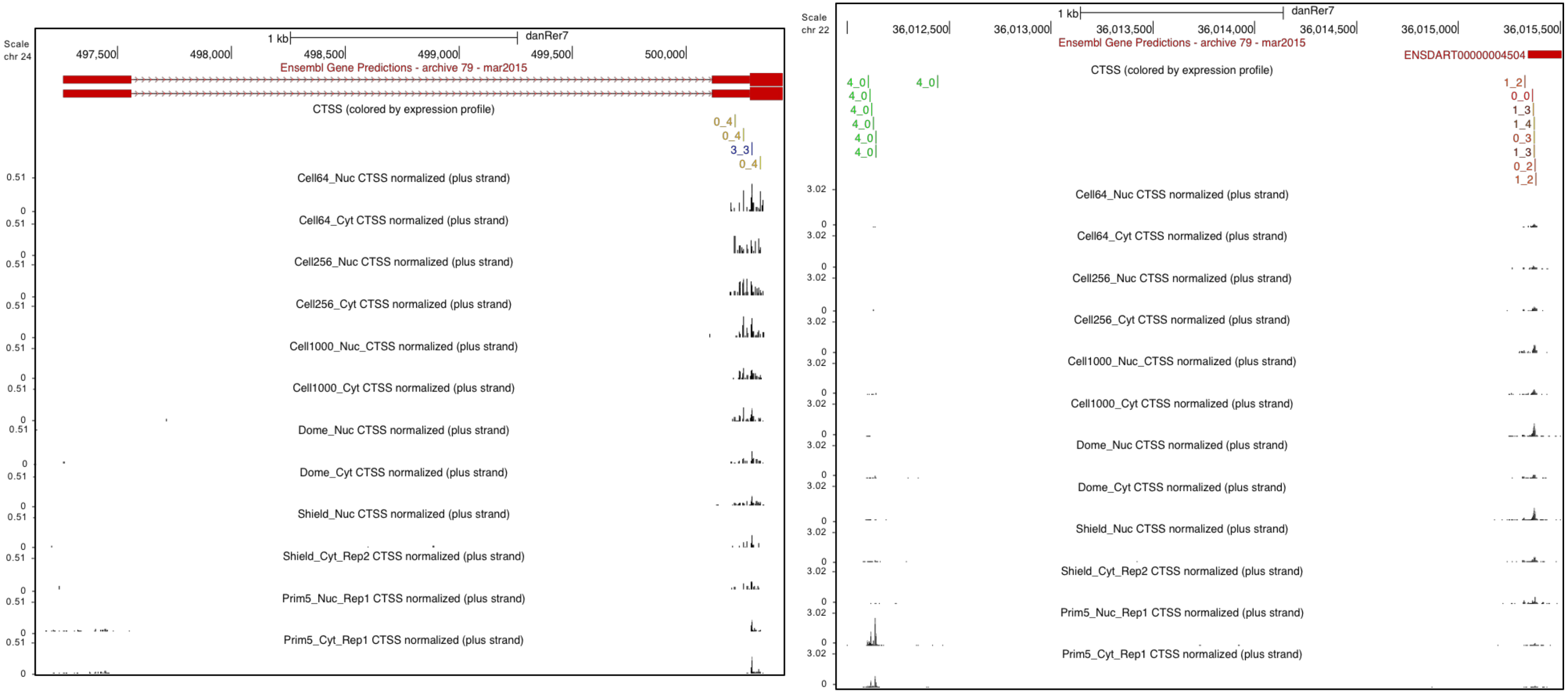


Figure 5.2.9 Differential promoter usage detected by CAGE-seq.
The left-hand side panel shows an example of a gene (*cpne5* gene) with intragenic TSS starting at the second exon. On the right-hand side we have *dag1* gene with a complete shift in the promoter region in the later stages of embryonic development.

5.2.7.1 Differences in promoter width and position of dominant TSS between the nuclear and cytosolic fractions among the stages.

The main purpose of this study, which distinguishes it from other CAGE-seq studies, was in understanding promoter dynamics in the different subcellular compartments. Hence, the next analysis was to focus on examples that show different promoter organization and even differences in the transcripts that are expressed in the nuclear and cytosolic fractions. We found genes (*mettl16*, *med13b, mxi1, pkig*) that showed differences in promoter width, the position of the dominant TSS within the same cluster, differences in the levels of the same transcript as well as the presence of a different transcript between the nuclear and cytosolic fraction (Fig 5.2.10-5.2.11). ENSEMBL predicts two different transcripts for the *mettl16* gene our nuclear CAGE detects expression for the longer isoform in the stages prior to MZT (Cell 64) (Zerbino et al., 2018). An increase in promoter width with multiple CTSSs is observed after MZT in both the fractions with higher expression levels in the cytosolic fraction (Fig 5.2.10, left). CAGE detected an alternative TSS, 500bp downstream of annotated TSS, for *med13b* gene (Fig 5.2.10, right) in the nuclear fraction in the dome and shield stages. The occurrence of alternative promoters in the two compartments of the cell with development suggests spatial and temporal regulation of the same gene. *mxi1* gene example shows the presence of two different transcripts (ENSDART00000104751 and ENSDART00000059923) for the same gene. On comparing the nuclear and cytosolic fractions the expression levels of the two transcripts showed differences. CAGE revealed the presence of another upstream TSS (> 1kb) for *pkig* gene with higher expression in the nuclear fraction in the dome and shield stages (Fig 5.2.11).

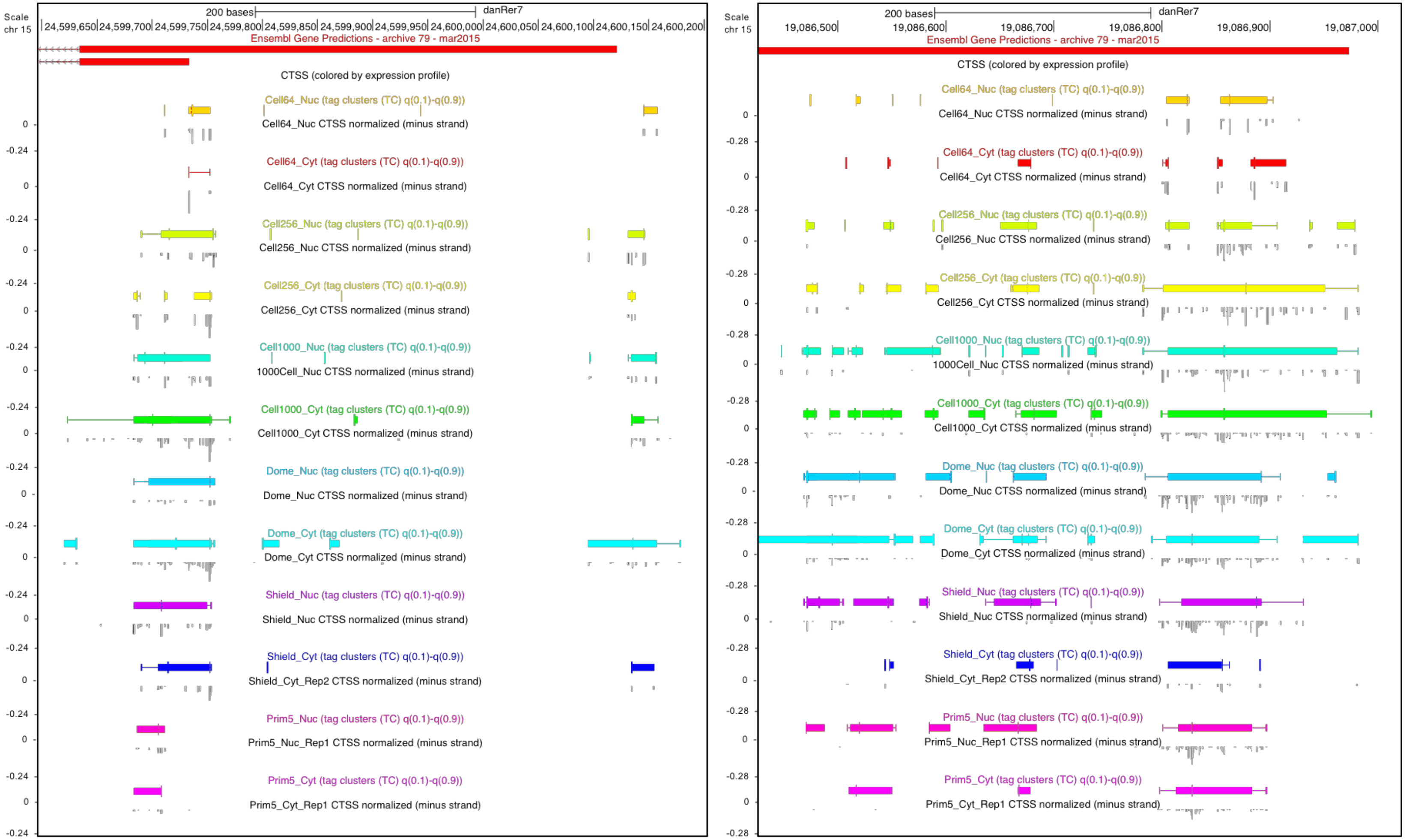


Figure 5.2.10 Differential TSS usage between the nuclear and cytosolic fractions.
mettl16 (left) and *med13b* (right) are examples of genes showing differences in TSS utilization between the fractions as well as changes in promoter width indicated by the coloured expression profiles above each track.

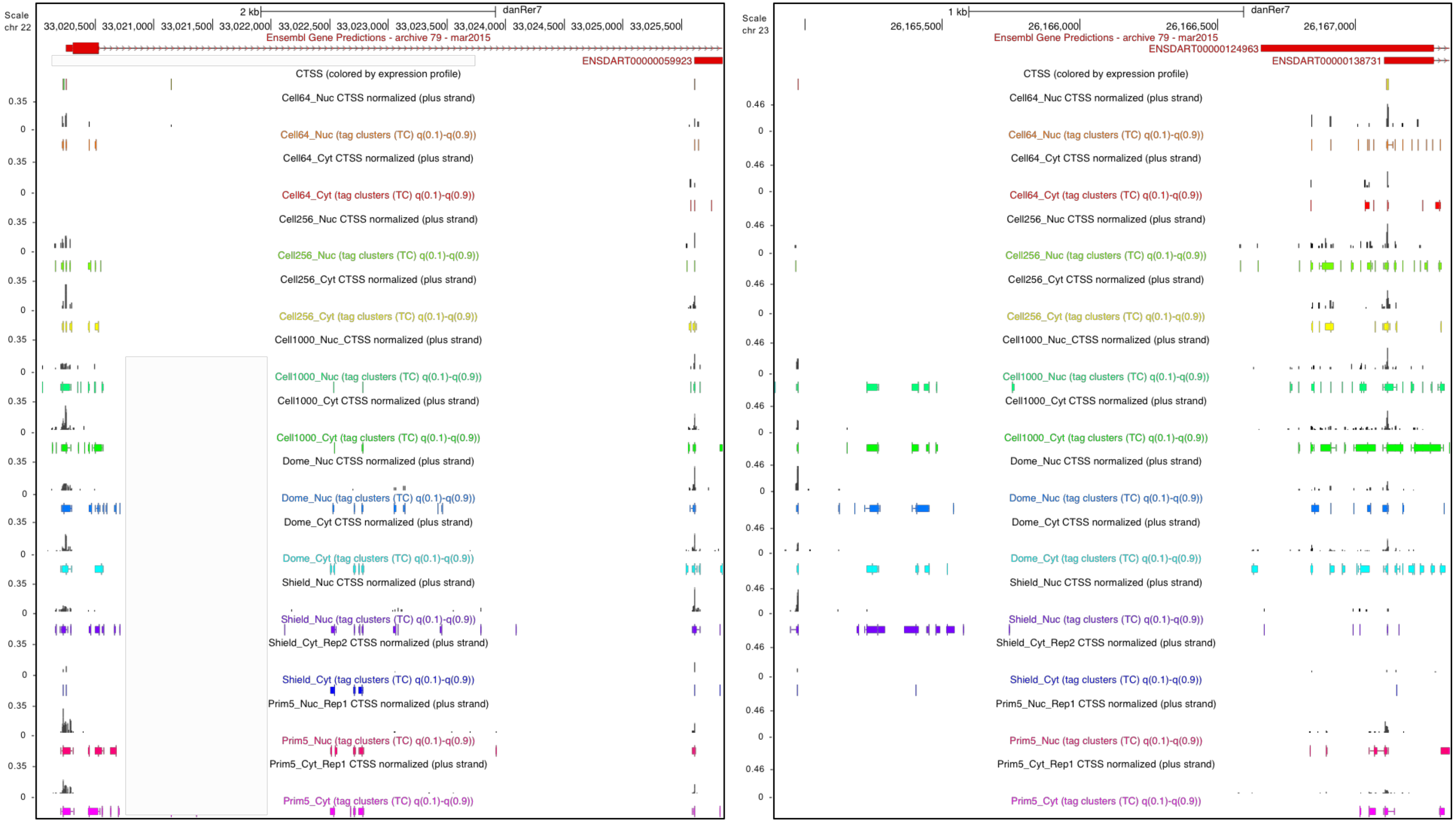
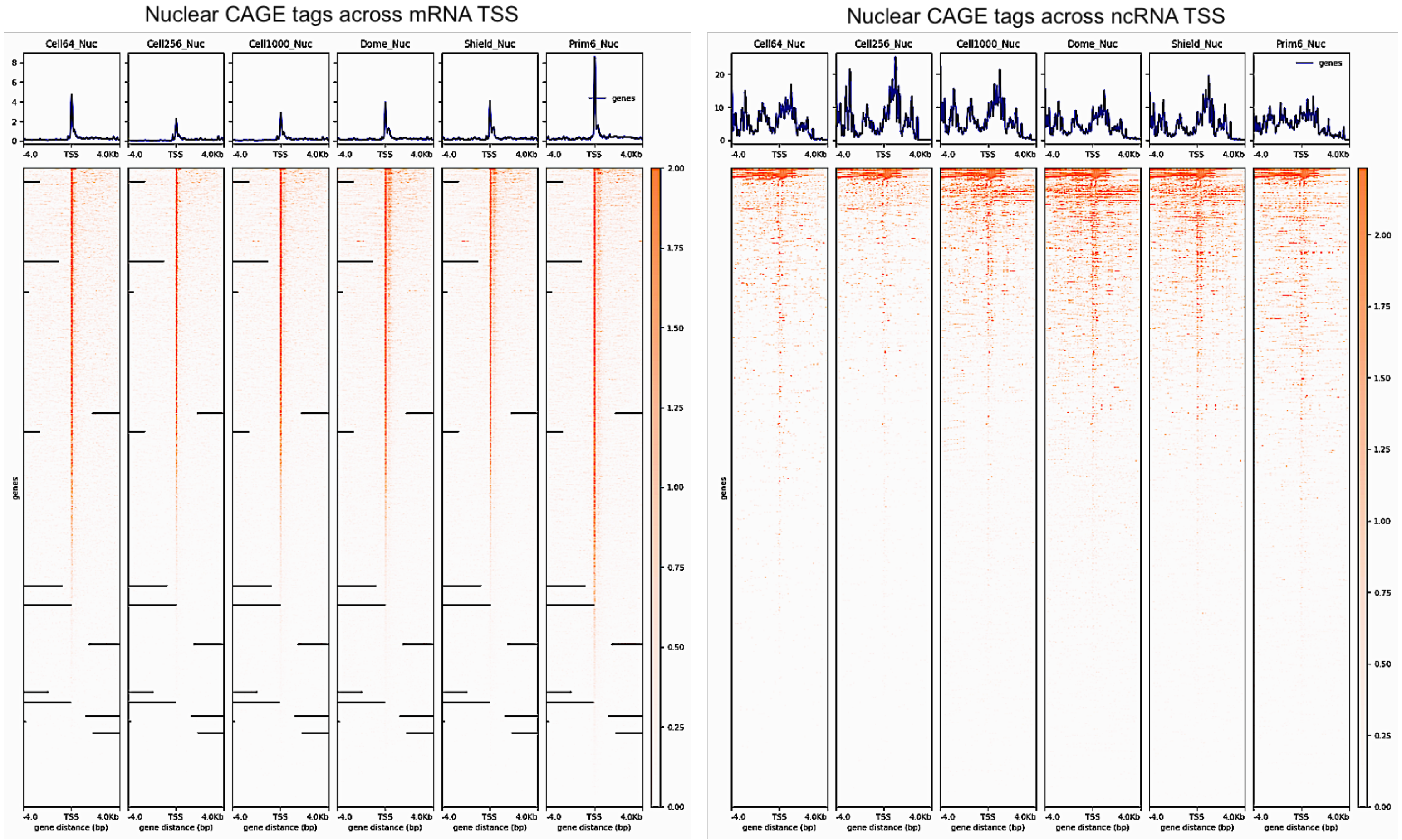


Figure 5.2.11 Differential TSS usage between the nuclear and cytosolic fractions.
mxi1 (left) and *pkig* (right) genes show differences in the levels of the different transcripts being expressed and a complete shifting of promoter which is more prominent in the nuclear fraction, respectively.

5.2.8 Distribution of CAGE-seq tags across the TSS of all ncRNAs in the nuclear fraction shows a bidirectional pattern

During the nuclear RNA-seq analysis we observed differences in the distribution of RNA reads at the TSS between the mRNA and ncRNA in the nuclear fractions. CAGE tag distribution analysis, across the annotated TSS (upstream and downstream of TSS) of ncRNAs and their comparison with TSS of mRNAs was carried out. Heatmaps of nuclear tags from different stages centered onto the annotated TSS were plotted (Fig 5.2.12).

As expected, for mRNAs, heatmaps showed that CAGE-seq tags mapped onto the annotated TSS with an increase in CAGE-tag intensity. Since CAGE is also a measure of expression of RNAs an increase in the transcript level with zygotic genome activation was also observed for mRNAs. For ncRNAs the intensity of the CAGE-seq tags increased with MZT but then reduced again by Prim5 stage. Majority of CAGE-seq tags observed mapped mostly upstream of or at the annotated TSS for ncRNAs. Some ncRNAs also show a bidirectional pattern with reads distributed both upstream and downstream of TSS after zygotic genome activation. This peculiar pattern detected for ncRNAs could be due to differences in the ENSEMBL annotation of ncRNAs. This result also validates our observations from the RNA-seq (Fig 4.2.14) which showed similar expression changes and a bidirectional pattern of read distribution both upstream and downstream of the ncRNA TSS.



Nuclear CAGE tags across mRNA TSS

Nuclear CAGE tags across ncRNA TSS

Figure 5.2.12 Read distribution across the TSS of mRNA and ncRNA.
Each panel in the heatmap represents a developmental stage and covers the 4kb upstream and downstream of TSS of genes (X-axis). The y-axis represents all the ncRNA and mRNAs and the z-axis is the intensity of the CAGE-seq tags (0-2 tpm).

5.2.8.1 CAGE and RNA-seq both detect promoters of antisense ncRNAs during zebrafish development

We further validated the expression levels observed for CAGE predicted TSSs by correlating the differences in the usage of TSS with the dynamics of gene activity detected by nuclear RNA-seq generated. For this study three of the antisense ncRNA examples from the negatively correlated group, one of the class of antisense characterised in the first results chapter (Fig 5.2.13-5.2.15) were observed on the genome browser. We also included the H3K27me3 data from the dome and shield stages as these antisense are polycomb targets. The antisense examples considered overlap important homeobox proteins which are essential for establishment of anterior/posterior body patterning during early vertebrate development (Corsetti et al., 1992). *Hoxb* cluster on chromosome 3 consists of 15 genes which all have transcription factor activity and binding specificity for RNA polymerase II promoter region and involved in embryonic skeletal system morphogenesis (Waxman et al., 2008). *lhx5* or the LIM homeobox 5 is a DNA binding protein which is responsible for eye and forebrain development (Peng and Westerfield, 2006). The *Hoxc* cluster on chromosome 23 consists of 15 genes which have RNA polymerase II distal enhancer activity and are involved in transcription regulation. Most of the genes in the cluster are essential for body patterning and skeletal morphogenesis (Stoll et al., 2013).

We observed that the antisense overlapping *hoxb* genes (Fig 5.2.13) showed a higher nuclear RNA level in the 64-Cell stage which further increased with zygotic genome activation. However, in the shield stage we observed an alternative start for the antisense transcript which was further validated by the RNA-seq and is only expressed in the nuclear fraction. The antisense overlapping the *lhx5* gene (Fig

5.2.14), on the other hand, showed higher cytosolic RNA levels and then switched to nucleus after the 1000-Cell stage. In the shield stage we again see a higher nuclear RNA levels and the presence of reads in the introns as well suggesting an unprocessed antisense in the nuclear fraction which is absent in the cytosol. We also observe correlation between the CAGE-seq tags, RNA reads and the H3K27me3 mark for *lhx5* antisense in the shield stage which is shown by the presence of both reads and tags close to the H3K27me3 histone mark. The antisense overlapping the *hoxc* genes (Fig 5.2.15) showed a similar increase in the nuclear level with zygotic genome activation. The CAGE, however, showed the presence of only two transcripts and the longer transcript in the 1000-Cell stage (MZT). It also detected an intragenic CAGE signal in the nuclear fractions for almost all stages of development considered.

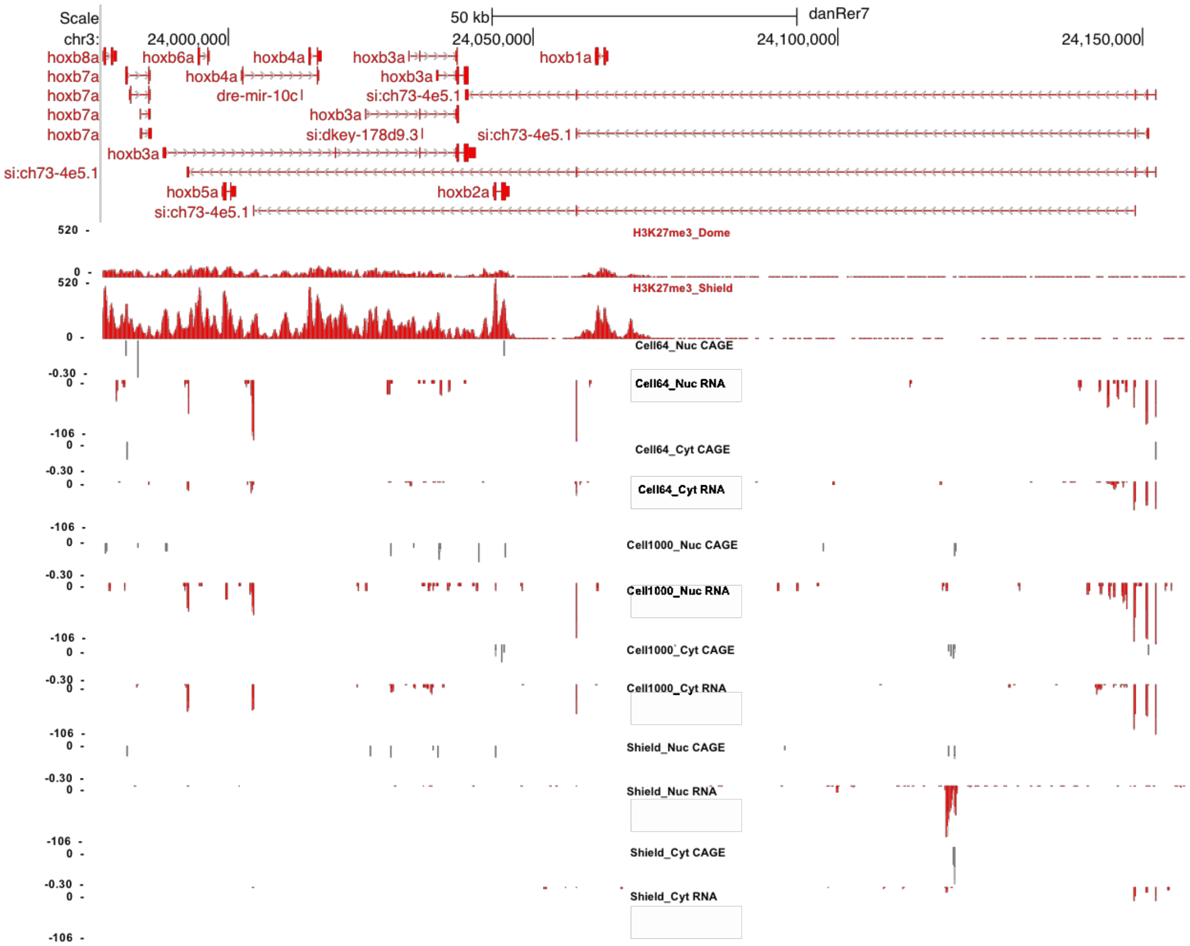


Figure 5.2.13 Antisense to* hoxb *genes.
The figure above shows a genome browser view of expression detected by both CAGE and RNA-seq for the antisense overlapping hoxb genes on chromosome 3. For simplicity we have considered only three stages 64-Cell, 1000-Cell and Shield. The left-hand panel shows the RNA levels (TPM) and CAGE-seq tags (tpm).

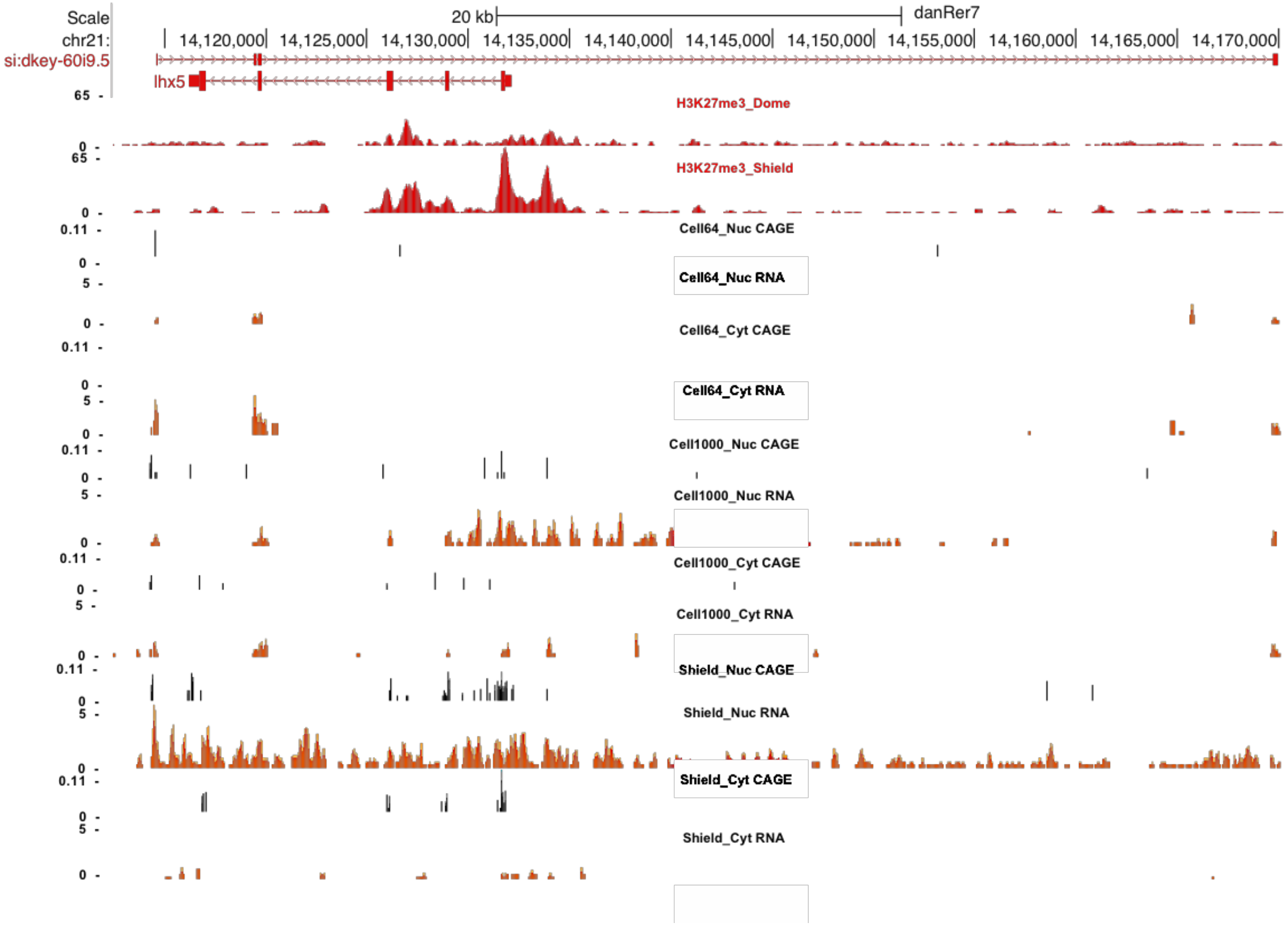


Figure 5.2.14 Antisense to* lhx5 *gene.
The figure above shows a genome browser view of expression detected by both CAGE and RNA-seq for the antisense overlapping lhx5 gene on chromosome 21. For simplicity we have considered only three stages 64-Cell, 1000-Cell and Shield. The left-hand panel shows the RNA levels (TPM) and CAGE-seq tags (tpm).

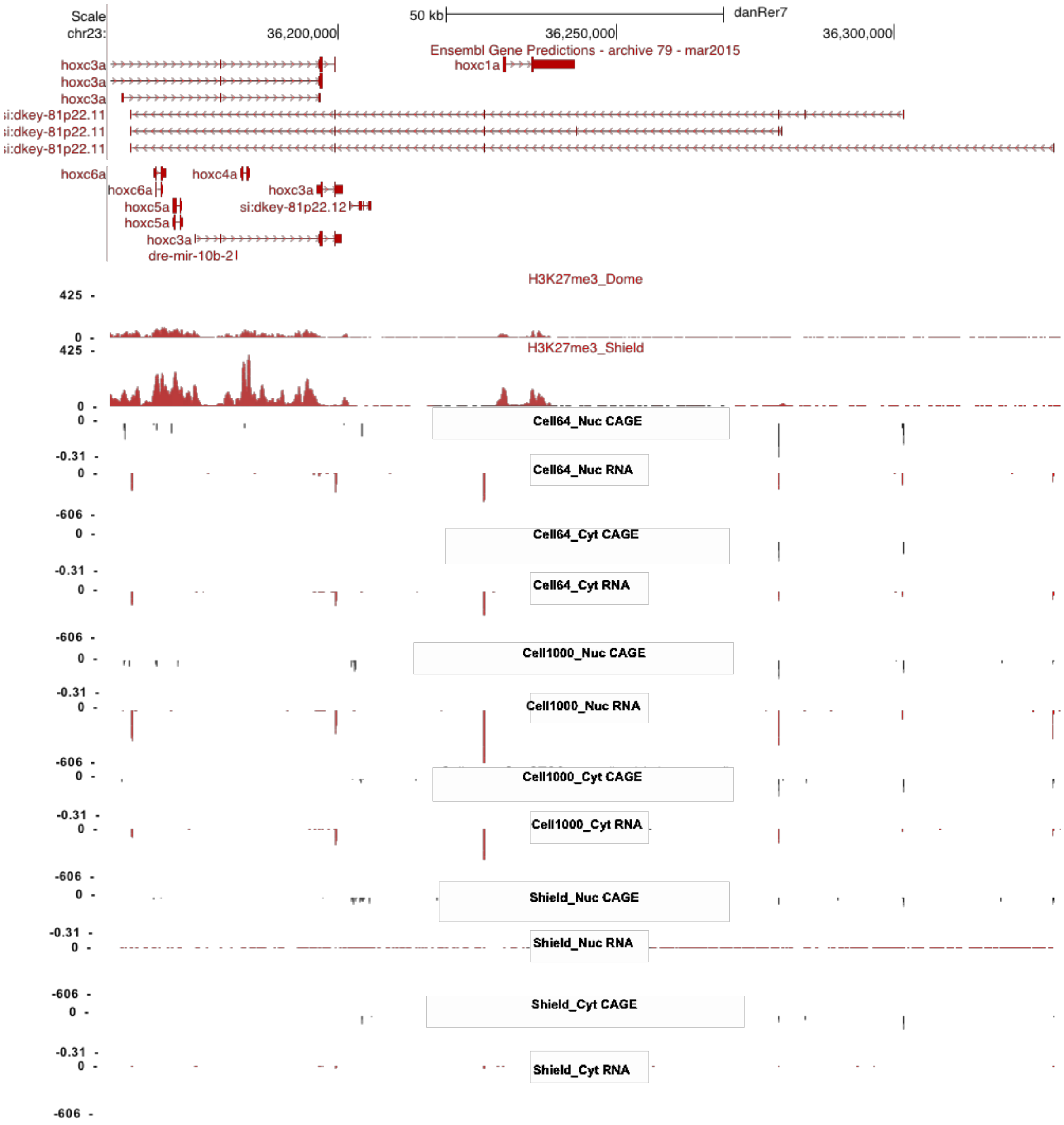


Figure 5.2.15 Antisense to* hoxc *genes.
The figure above shows a genome browser view of expression detected by both CAGE and RNA-seq for the antisense overlapping hoxc genes on chromosome 23. For simplicity we have considered only three stages 64-Cell, 1000-Cell and Shield. The left-hand panel shows the RNA levels (TPM) and CAGE-seq tags (tpm).

5.2.9 CAGE-seq showed absence of tags across distal peaks for repressive histone modification mark

Post-translational histone modifications generally affect gene expression by epigenetic changes which results in closed or open chromatin structures. Several studies in the past have demonstrated how histone modifications change with zygotic genome activation during zebrafish embryogenesis (Vastenhouw et al., 2010). To observe the association between the nuclear TSS selection and the chromatin configuration we mapped all the nuclear reads upstream and downstream of the peaks for H3K27ac (activation mark), H3K4me3 (poised promoters mark) and H3K27me3 (repressive mark) in the dome stage.

The nuclear CAGE-seq tags increased with zygotic genome activation on mapping at the H3K27ac and H3K4me3 peaks especially in the Dome stage. However, the intensity of CAGE-seq tags decreased in both the dome and shield stages for the H3K27me3 peaks but then increased in the Prim5 stage suggesting the removal of these repressive marks (APPENDIX, Fig 5). In addition, the CAGE-seq tags for all stages were also mapped on H3K27me3 peaks which were more than 2 kb distal to the annotated TSS of any protein-coding genes (Fig 5.2.16). We observed that H3K27me3 peaks were mutually exclusive with H3K27ac and H3K4me3 modifications as reflected by an absence of both H3K27ac (in red) and H3K4me3 ChIP-seq reads (in green) at H3K27me3 enriched regions. The presence of CAGE-seq tags at these distal peaks indicates towards the presence of distal regulatory elements, such as enhancers, that bind chromatin modifying complexes. In addition, CAGE-seq tags did not show any enrichment at H3K27me3 (in gold) modified regions. As expected, an increase in enrichment of both CAGE-seq tags and H3K27ac and H3K4me3 ChIP-seq reads was seen as the intensity of

H3K27me3 marks decreased. However, the Prim5 stage showed enrichment for CAGE-seq tags even when the H3K27me3 mark was higher proposing removal of these marks as observed at the H3K27me3 peaks.

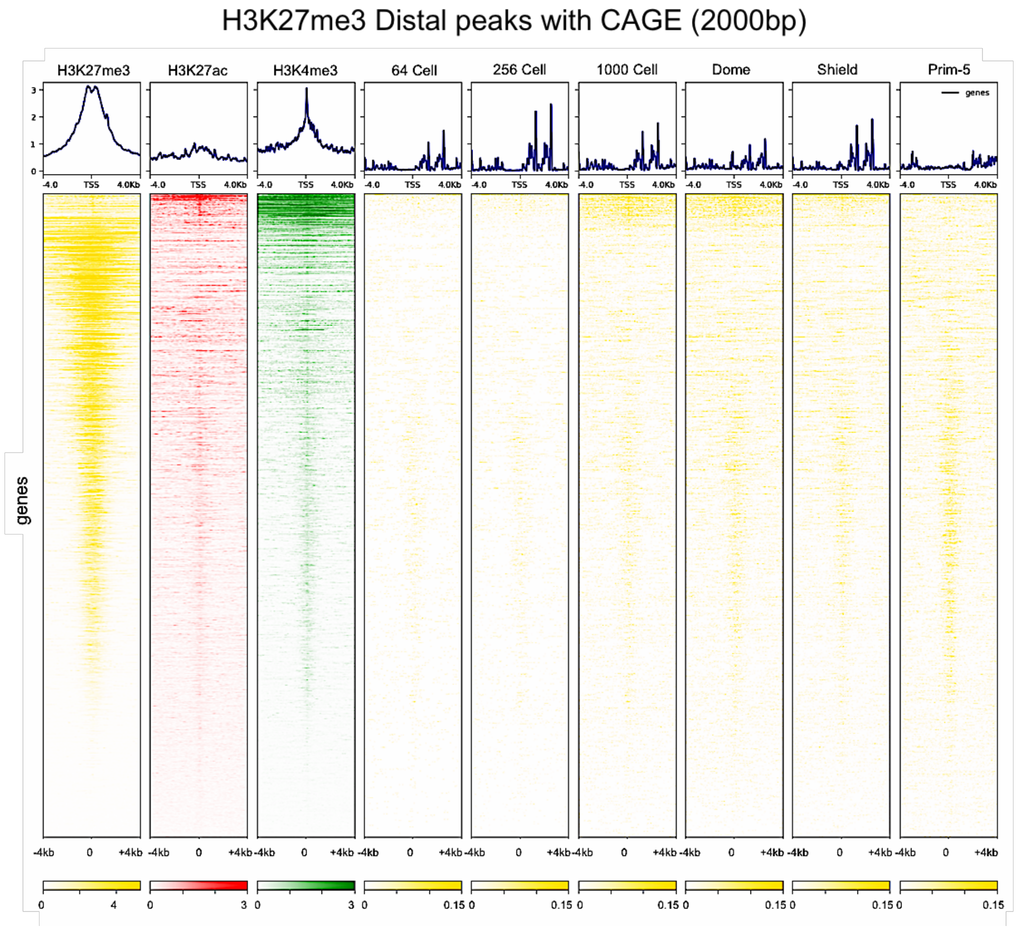


Figure 5.2.16 Chromatin modifications at peaks distal (2kb) of H3K27me3 peaks.
The heatmaps show the different histone modifications in the dome stage of development. The X-axis of each window ranges from -4kb to +4kb of the peaks (labelled as TSS and TES). Red shows the distribution of nuclear reads across the H3K27ac peaks, green is for H3K4me3 peaks and yellow represents the H3K27me3 peaks.

5.3 Discussion

Maternal-to-zygotic transition is one of the most essential phenomena that occurs in the development of a vertebrate embryo (Lee et al., 2013).It is accompanied by a change in the embryonic transcriptome from maternal-to-zygotic and is, therefore, of importance in understanding the variations that occur in the promoter and TSS usage. The analysis of the CAGE-seq data in the present chapter showed differences in promoter width with embryonic development (similar to (Nepal et al., 2013)). The switch in the presence of sharp promoters (associated with TATA-box) in the early pre-MZT stages to broad promoters (that overlap with CpG islands) in later post-MZT stages suggests changes in the mode of regulation of a gene by the promoter. The sharp promoters being associated with housekeeping and developmentally regulated genes and broad promoters being common in tissue specific genes (Carninci, 2006). Our study also revealed differences in promoter width within the same stage (1000-Cell, Dome and Prim5) in different compartments of the embryonic cell proposing a spatial and temporal regulation of gene expression.

The clustering of individual CTSSs based on their expression levels in the different stages with development revealed the presence of several transcript classes. CTSSs that were present in the maternal stages (64-Cell, 256-Cell and 1000-Cell) such as cluster 0_4 suggesting maternally deposited RNAs. Then the cluster 0_0 showing CTSSs that peaked during the maternal-to-zygotic transition proposing transcripts important in zygotic genome activation. Zygotic CTSSs that are active during the later stages such as revealed by cluster 4_0. Cluster 2_3 denotes CTSSs that are present throughout the embryonic development suggestive of housekeeping genes. This has already been observed by Haberle et al. (Haberle et al., 2014) in their study. However, our CAGE-seq also revealed another class of CTSSs, cluster 4_3, that showed differences in expression between the nuclear and cytosolic fractions in the MZT and post-MZT stages. Thus, indicating the localisation and potential function of the transcript in different subcellular compartments.

The analysis of alternative promoter usage showed several examples where the TSS position within a single promoter were different between the maternal and zygotic transcripts. We also see the presence of a broader promoter in the 1000-Cell and dome stages of development, in comparison to the 64-cell stage, possibly due to changes in the regulation of gene with zygotic genome activation. The presence of intragenic CTSSs suggests the synthesis of cleaved RNAs as a by-product of post-transcriptional processing which are possibly important during development (also shown in (Nepal et al., 2013)). The differences in both the RNA levels and the position of the dominant TSS within the same promoter for the two fractions perhaps implies the variations in the usage of the transcripts. The nuclear transcript might have regulatory roles and remains in the nucleus while the cytosolic transcript is utilised in the process of translation and thus, the TSS is also an indication of the translation start site. The translation start site selection is also quite possibly the reason for differential TSS usage that we see amongst the maternal and zygotic transcripts.

Non-coding RNA dynamics. The bidirectional pattern observed for ncRNAs in the nuclear fraction with zygotic genome activation (Fig 5.2.12) further validates the arrangement seen with nuclear RNA reads for ncRNAs (Fig 4.2.14). This suggests the presence of alternative transcription start sites as opposed to ENSEMBL annotated for ncRNAs. We see a significant correlation between the nuclear RNA-seq and nuclear CAGE-seq data when observed on the genome browser (Fig

5.2.13-5.2.15). In addition, the CAGE also detects a lot of intragenic CTSSs for the antisense post-MZT, also confirmed by nuclear RNA-seq, indicating towards the presence of post-transcriptional processing. Thus, giving further insights into the mechanism of regulation of overlapping protein-coding genes by these antisense.

Histone modification. On mapping the nuclear CAGE-seq tags 2 kb distal of H3K27me3 peaks at dome stage we observed an absence of tags in loci where the ChIP-seq reads were enriched. However, the genomic loci which showed lack of H3K27me3 ChIP-seq reads showed presence of CAGE-seq tags which was also true for the activating chromatin marks, H3K27ac and H3K4me3. The intensity of the CAGE-seq tags increased in the Prim5 stage quite possibly due to the removal of H3K27me3 mark by this stage. Thus, our nuclear CAGE-seq revealed the presence of distal regulatory elements that bind chromatin modifying complexes.

CHAPTER 6: GENERAL DISCUSSION

Considering the low expression levels and specific localisation of lncRNAs, availability of deep sequencing studies such as RNA-seq or CAGE-seq would help in identification of novel lncRNAs. In zebrafish so far, there have been only a few studies which have systematically identified lncRNAs during development based on RNA-seq or CAGE-seq datasets (Pauli et al., 2012, Haque et al., 2014, Dhiman et al., 2015, Hu et al., 2018). However, these studies have not categorised these lncRNAs and their regulation of protein-coding genes during zebrafish development. Past studies in human cell lines have also shown the importance of lncRNAs in development. In humans, studies on stem cell differentiation have shown the importance of lncRNAs such as *hoxBlinc* in hematopoietic differentiation, *ALIEN* in endodermal differentiation, SENCR in endothelial differentiation and *Tunar* in neuronal differentiation (Perry and Ulitsky, 2016). Several studies during vertebrate development have also identified the presence of conserved non-coding sequences predominantly in close proximity to developmental genes (Woolfe et al., 2005, Ponjavic et al., 2009, Luo et al., 2016). Thus, establishing the significance of lncRNAs expressed during zebrafish development in the context of maternal-to-zygotic transition will help in understanding the development process better.

6.1 Two distinct class of antisense with differences in expression pattern, localisation and chromatin features

In humans several approaches such as localisation within a cell, genomic positioning, expression and association with chromatin features have been used to characterise lncRNAs (Ponjavic et al., 2009, Mondal et al., 2010, Cabili et al., 2011, Derrien et al., 2012, Derrien et al., 2011). Our aim was to identify coding and antisense pairs present during the early stages of zebrafish embryogenesis and characterise them into a class based on their expression correlation. Using an existing RNA-seq dataset (Pauli et al., 2012) we discovered the presence of two distinct classes of antisense: negatively correlated and positively correlated group. The two class of antisense showed peculiarities in inheritance with negatively correlated antisense being exclusively maternally deposited and antisense in the positively correlated group showed both maternal and zygotic contributions (Fig 3.2.2). Our analysis also revealed differences in the function of coding genes overlapping these antisense, with maternally deposited antisense overlapping important developmental genes (Fig 3.2.6). We are proposing the importance of these maternally deposited antisense in early zebrafish development. The coding genes in the negatively correlated category were also associated with repressive histone modification mark, H3K27me3. In the positively correlated group, however, both the antisense and the coding genes had H3K27ac and H3K4me3 but no H3K27me3 marks at their promoter (Fig 3.2.10). We also observed a higher overlap region (~14kb) between the maternal antisense and their overlapping coding gene compared to positively correlated category of antisense and mRNA gene (Fig 3.2.8). Since the two class of antisense showed such distinct features we further investigated for differences in their subcellular localisation. This will also help us in better understanding the mechanism of regulation of protein-coding expression by these antisense. The negatively correlated antisense showed a more stage specific localisation with predominant cytosolic presence during the early stage (64-Cell) and then becoming more abundant in nucleus with MZT. There was, however, a few antisense which become cytosolic (purple) in the shield stage and a few that become nuclear in the 256-Cell probably indicating the importance of these antisense in regulation of their overlapping protein-coding genes (Fig 4.2.3-4.2.4). The coding genes did not show specific localisation in the early stages possibly because they were not being expressed. But with MZT we saw a very specific nuclear localisation suggesting their transcription (Fig 4.2.5). In the positively correlated antisense, on the contrary, majority of antisense were not specifically or exclusively localised (green or yellow) into compartments until after MZT when they became nuclear (red) (Fig 4.2.3-4.2.4). Further the conservation of characterised maternal antisense, overlapping developmental genes, in both mouse and humans, in a syntenic manner, confirms the significance of these lncRNAs during development (Fig 3.2.14-3.2.17).

We propose a mechanism whereby the maternally deposited antisense acts as a means of regulation of overlapping protein-coding gene expression, post-transcriptionally, by RNA-RNA hybridisation (in the cytosol) before the zygotic genome activation. After the zygotic genome activation since these protein-coding genes are targeted by the PRC2 complex their expressions are regulated by histone modifying complexes. Therefore, implying that the maternally deposited antisense act as a backup mechanism for regulation of overlapping protein-coding genes because of the absence of repressive histone marks prior to MZT (Fig 3.3.1).

6.2 RNA dynamics with zygotic genome activation in zebrafish development

6.2.1 Non-coding RNA dynamic.

In ENCODE the addition of nuclear RNA sequencing data proved to be essential in identifying several ncRNA's which were only localised in the nucleus and not detected in the whole cell RNA sequencing data (Derrien et al., 2012). Our RNA-seq dataset detected several ncRNAs, with three-fold higher depth, that were not present in Pauli et al. dataset (Pauli et al., 2012) in the early stages of zebrafish development (Fig 4.2.15-4.2.17). We also observed majority (59%) of ncRNAs to be present in the cytosol before zygotic genome activation. However, with transcription activation ncRNAs become more and more nuclear (21-74%) (Fig 4.2.12). Both our nuclear RNA-sequencing and CAGE-sequencing data showed a bidirectional distribution of reads across the TSS of ncRNAs perhaps indicating absence of accurate annotations for these RNAs (Fig 4.2.14, 5.2.12). In general, we observed a stage specific expression of ncRNAs which has been previously seen in Pauli et al. RNA-seq (Pauli et al., 2012). However, the difference was that in our study we detect very different expression pattern for the different class of ncRNAs in the nuclear and cytosolic fractions (Fig 4.2.19). Thus, suggesting different stage specific function for the ncRNAs in the different subcellular compartments. The antisense were established to be the highly abundant class of ncRNAs in the cytosol whereas lincRNAs formed the majority in the nuclear fraction before zygotic genome activation (Fig 4.2.19). This further indicated towards the importance of overlapping or neighbouring protein-coding gene regulation prior to MZT. However, the abundance of snRNAs in both the nuclear and cytosolic fraction after zygotic genome activation suggested the shift in the requirement for splicing machinery with transcription of zygotic genes (also observed in human cell lines (Djebali et al., 2012)). The antisense ncRNAs were distinguished into a class, nuclear and cytosolic, based on their subcellular localisation. The nuclear antisense expression showed both maternal and zygotic contributions while the cytosolic antisense were mostly maternal with a majority not expressing in the shield stage (Fig 4.2.20-4.2.21). Out of all the categorised cytosolic antisense 56 antisense were specific to the 64-Cell stage and only 11 that were exclusive to the shield stage. In the category of nuclear antisense only 8 were specific to the 64-Cell stage while 92 antisense were explicitly expressed in the shield stage (Fig 4.2.23). Consequently, implying the importance of specific localisation of ncRNAs with embryonic development. The 56 antisense in the 64-Cell stage overlapped protein-coding genes more significantly enriched in TF activity and development (p-values greater than 10^{-6}) (Fig 4.2.24). However, the coding genes overlapping the nuclear antisense showed enrichment for insulin-like growth factor receptor binding and metabolism (p-values greater than 10^{-3}) (Fig 4.2.25). Thus, possibly suggesting presence of different mechanisms to regulate the expression of mRNA genes by antisense in the different subcellular compartments.

6.2.2 mRNA dynamic.

We observed rapid degradation (95-65%) of cytosolic mRNAs, perhaps maternal mRNAs, and a comparable increase in nuclear mRNA levels (5-35%) with maternal-to-zygotic transition (Fig 4.2.7). The read distribution, 4kb upstream and downstream, of mRNA TSS showed coverage at the TSS before zygotic genome activation. However, with zygotic transcription the reads occurred downstream suggesting mRNA synthesis. This was all in accordance with what has been previously observed about mRNA expression with zebrafish development (White et al., 2017).

6.3 Differential TSS usage and promoter organisation during zebrafish embryogenesis

Our CAGE-Seq study revealed differences in promoter as well the usage of TSS in the different subcellular compartment within the same stage proposing a spatial regulation of gene expression. On comparison with an existing CAGE-Seq dataset on zebrafish development (Nepal et al., 2013) our dataset identified more than three-fold higher number of TCs in stages after zygotic genome activation (Fig 5.2.2) providing a more comprehensive dataset for detection of transcription start sites. On observing the promoter width, we saw presence of broader promoters in the cytosolic fraction when compared to the nuclear during and after ZGA (1000-Cell, Dome and Prim-5). However, pre-MZT stages showed similar promoter width within the fractions (64-Cell and 256-Cell) (Fig 5.2.4-5.2.5). Expression profile revealed the presence of TSSs specific for maternal RNAs, zygotic RNAs, RNAs expressing exclusively during maternal-to-zygotic transition (Fig 5.2.6). This has been previously identified by Nepal et al. study (Nepal et al., 2013) on zebrafish development. However, our CAGE-Seq dataset also detected another class of TSSs that showed differences in expression between the nuclear and cytosolic fraction during MZT and post-MZT stages (Fig 5.2.6). Genes with multiple transcripts showed presence of different transcripts in the two subcellular compartments. Our nuclear CAGE tags showed enrichment at loci distal (2kb) to the H3K27me3 peaks suggesting presence of distal enhancer elements (Fig 5.2.16).

6.4 Future work and Conclusion

In our study on antisense ncRNAs (CHAPTER 3) we were successfully able to classify antisense ncRNAs into two categories based on expression, timing of expression, associated chromatin features and localisation during development. However, the features associated with maternal antisense and protein-coding gene pairs identified in this study needs further experimental validations to confirm the mechanism of regulation of mRNA expression by the antisense. It would be interesting to observe the results of chromatin associated with these lncRNAs using high-throughput techniques such as ChiRP-Seq (Chromatin isolation by RNA purification sequencing). Functional studies such as CRISPRi (Clustered Regularly Interspaced Short Palindromic Repeats interference) or addition of PolyA signals downstream of the start site of antisense would further help in understanding whether or not the absence of these antisense affects the expression of neighbouring mRNA genes. Analysis into the sequence and gene structure of some of the conserved antisense, identified overlapping developmental genes, in both humans and mouse would provide further insights into the significance of these antisense.

In summary, our nuclear RNA-seq and CAGE-seq generated as part of our study provides a vast resource for other researchers in the field of zebrafish development. It also revealed very specific (both stage and space) expression pattern for ncRNAs during zebrafish development which can be further validated *in vitro*. Both nuclear CAGE and RNA-seq could be used for annotating novel enhancer RNAs and ncRNAs essential in zebrafish development (an ongoing paper part of the DANIOCODE). Our nuclear data also provides opportunity for future functional and evolutionary studies such as differences in promoter sequence from nuclear to cytosolic during development.

APPENDIX I

Figure 1. Gapdh expression in RNAs isolated from zebrafish embryonic cells in different stages of development (as a control to check the quality of RNA extracted).

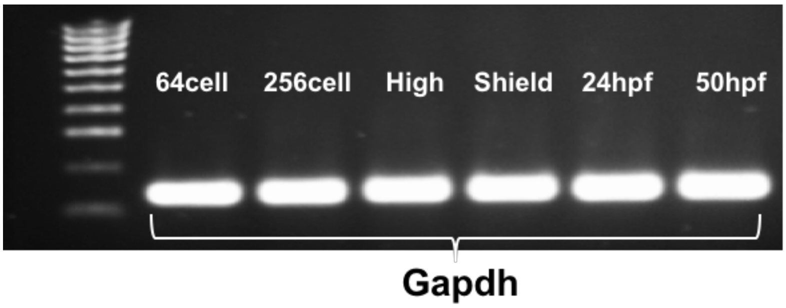


Figure 2. the heatmap below shows the expression of all ncRNAs (above 1 TPM) in the nuclear and cytosolic fractions. The X-axis represents the stages of development while the Y-axis represents ncRNAs. A dark colour indicates an expression from 80-100 TPM while lighter shade of blue is an indication of lower RNA levels (0-20 TPM).

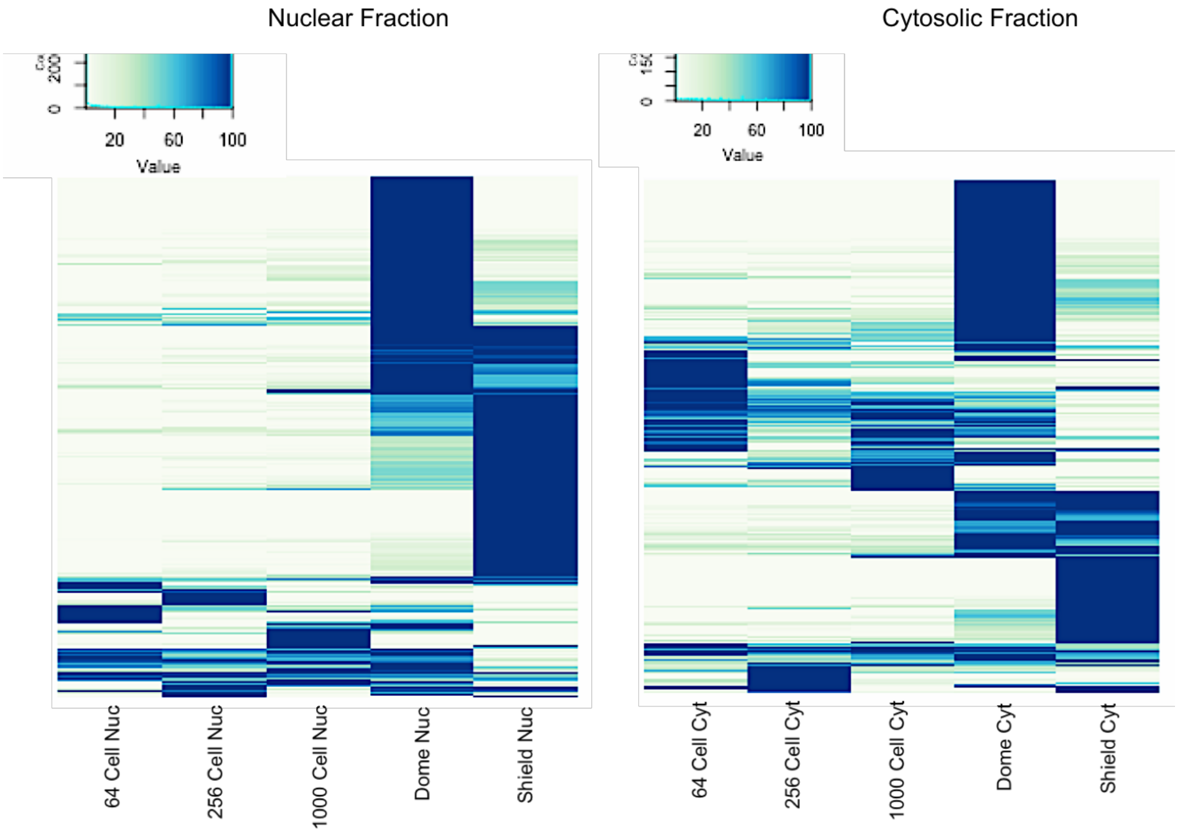


Figure 3. The violin plots below show the RNA levels of cytosolic and nuclear antisense with zebrafish development. the X-axis represents the stages in development while the Y-axis indicates RNA levels in TPM.

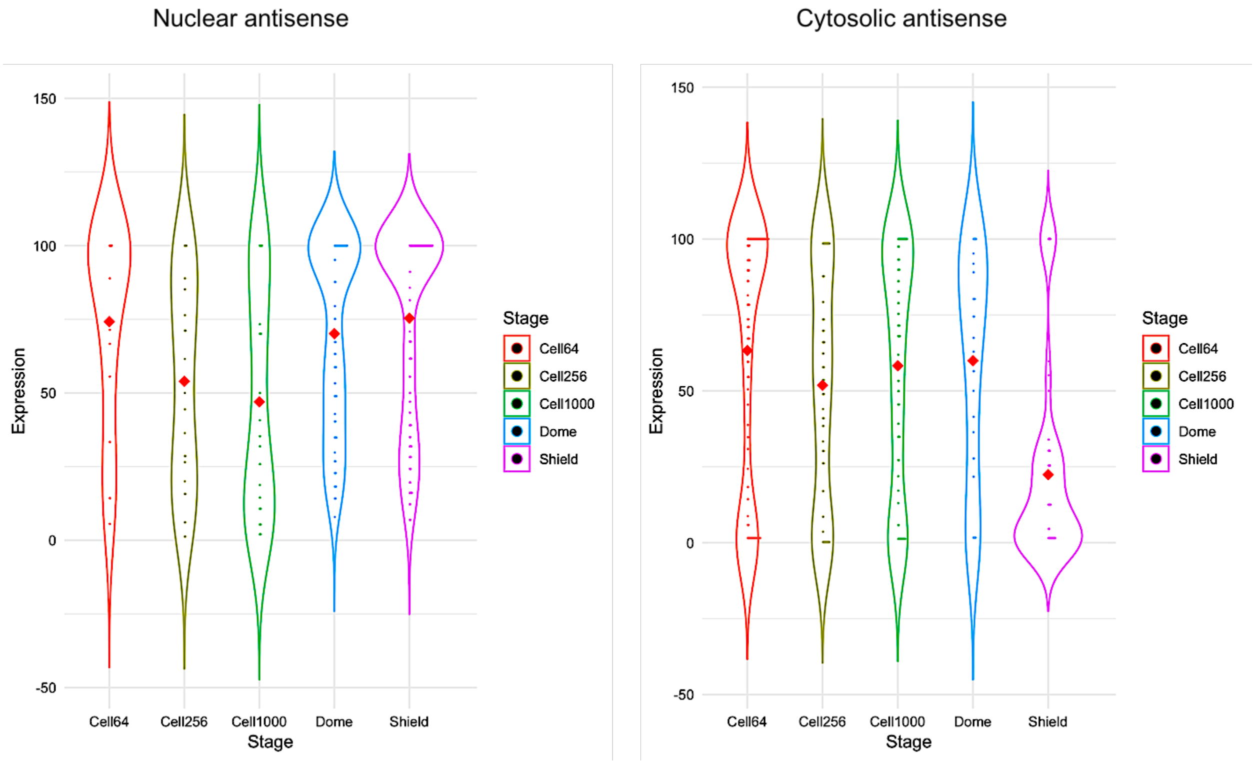
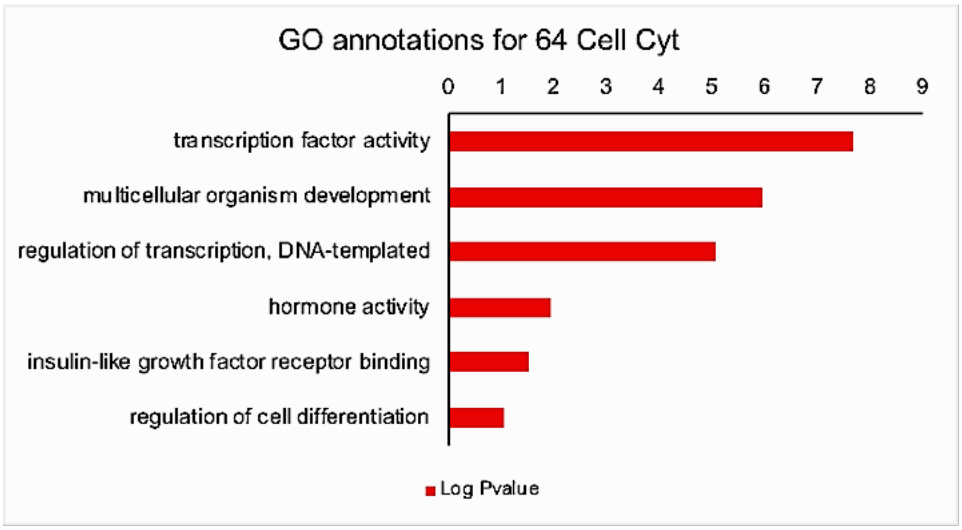


Figure 4. the graphs below show the gene ontology terms associated with mRNA genes overlapping cytosolic antisense (64-Cell stage) and the nuclear antisense (Shield stage). The analysis revealed that the unique cytosolic antisense observed in the 64-Cell stage overlap important developmental genes (p-value > 7.5) while the nuclear antisense mutually exclusive to the shield stage showed a higher enrichment for genes involved in signalling (p-value > 3).



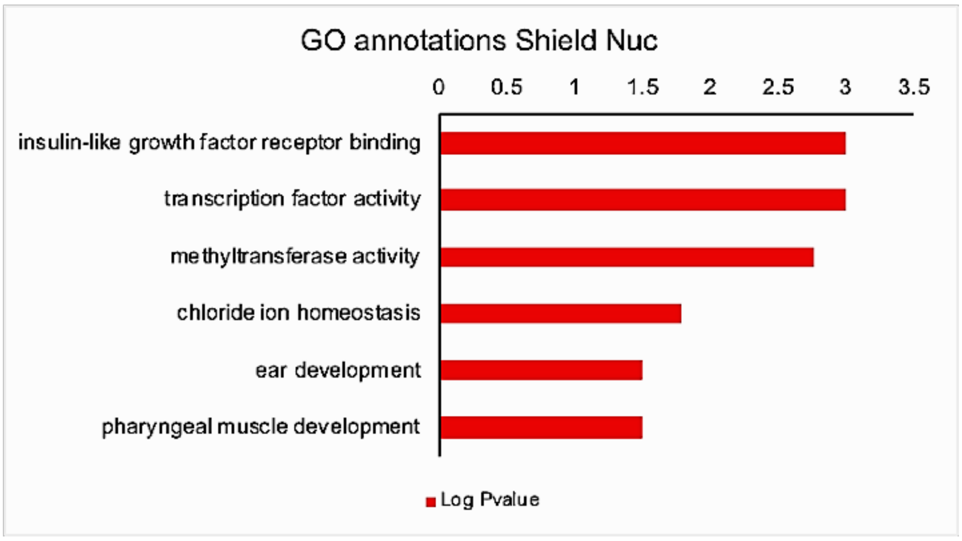
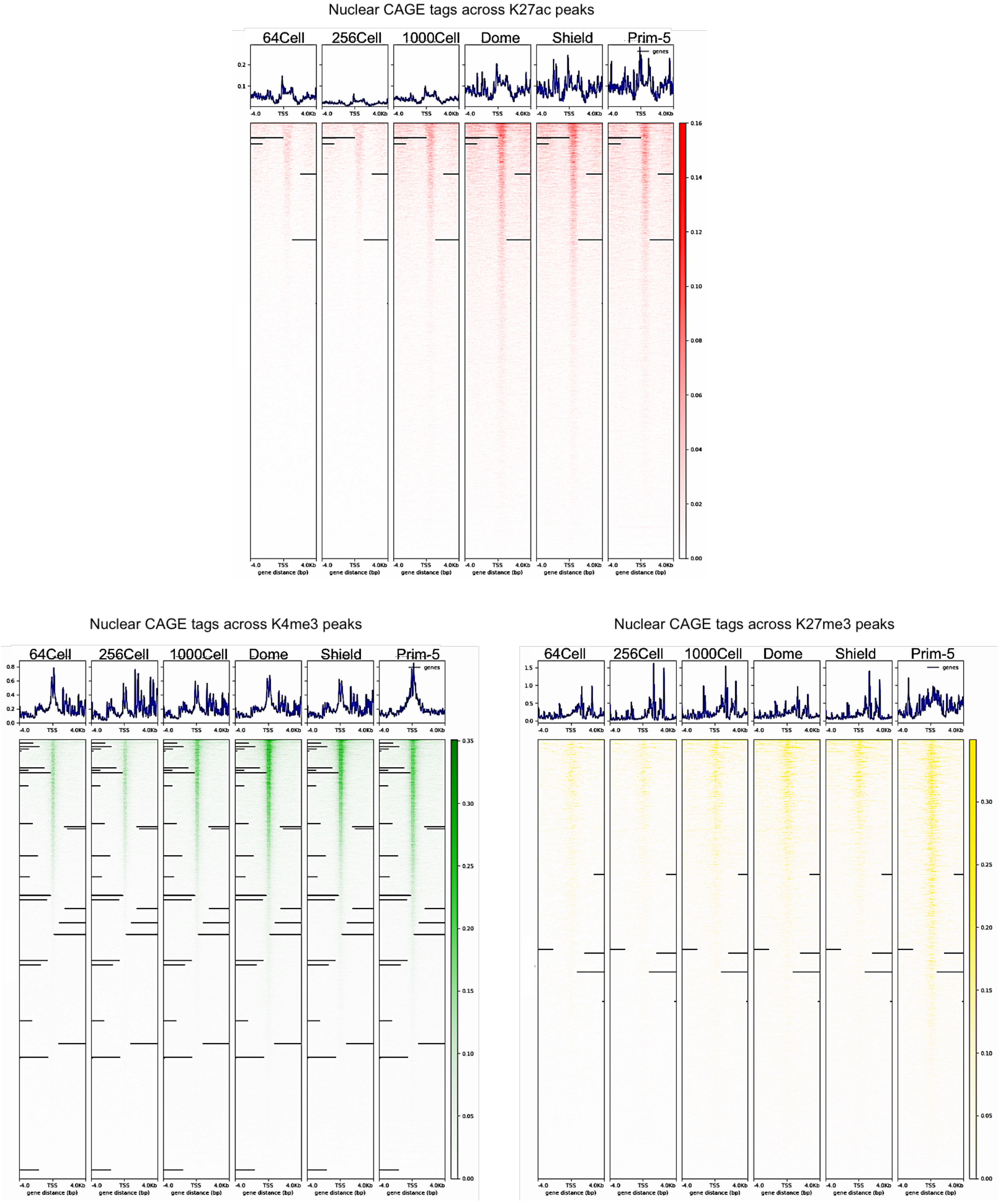


Figure 5. the heatmaps below represent the Nuclear CAGE-tags mapped to the H3K27ac (in red), H3K4me3 (in green) and H3K27me3 (in yellow) ChIP peaks. The Y-axis represents the ChIP loci while the X-axis signifies ±4kb of the ChIP peak center (TSS).



Nuclear CAGE tags across K27ac peaks

Nuclear CAGE tags across K4me3 peaks

Nuclear CAGE tags across K27me3 peaks

REFERENCES

AANES, H., WINATA, C. L., LIN, C. H., CHEN, J. P., SRINIVASAN, K. G., LEE, S. G., LIM, A. Y., HAJAN, H. S., COLLAS, P., BOURQUE, G., GONG, Z., KORZH, V., ALESTROM, P. & MATHAVAN, S. 2011. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res,* 21**,** 1328-38.

AASLAND, R., STEWART, A. F. & GIBSON, T. 1996. The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB. *Trends Biochem Sci,* 21**,** 87-8.

ABBOTT, A. L., ALVAREZ-SAAVEDRA, E., MISKA, E. A., LAU, N. C., BARTEL, D. P., HORVITZ, H. R. & AMBROS, V. 2005. The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans. *Dev Cell,* 9**,** 403-14.

ABRAMS, E. W. & MULLINS, M. C. 2009. Early zebrafish development: it's in the maternal genes. *Curr Opin Genet Dev,* 19**,** 396-403.

AFFYMETRIX, E. T. P. & COLD SPRING HARBOR LABORATORY, E. T. P. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature,* 457**,** 1028-32.

ALBERTS, B. & NATIONAL, A. 2002. From the National Academies. *Cell Biol Educ,* 1**,** 109-10.

ALOIA, L., DI STEFANO, B. & DI CROCE, L. 2013. Polycomb complexes in stem cells and embryonic development. *Development,* 140**,** 2525-34.

ANDERSSON, R., GEBHARD, C., MIGUEL-ESCALADA, I., HOOF, I., BORNHOLDT, J., BOYD, M., CHEN, Y., ZHAO, X., SCHMIDL, C., SUZUKI, T., NTINI, E., ARNER, E., VALEN, E., LI, K., SCHWARZFISCHER, L., GLATZ, D., RAITHEL, J., LILJE, B., RAPIN, N., BAGGER, F. O., JORGENSEN, M., ANDERSEN, P. R., BERTIN, N., RACKHAM, O., BURROUGHS, A. M., BAILLIE, J. K., ISHIZU, Y., SHIMIZU, Y., FURUHATA, E., MAEDA, S., NEGISHI, Y., MUNGALL, C. J., MEEHAN, T. F., LASSMANN, T., ITOH, M., KAWAJI, H., KONDO, N., KAWAI, J., LENNARTSSON, A., DAUB, C. O., HEUTINK, P., HUME, D. A., JENSEN, T. H., SUZUKI, H., HAYASHIZAKI, Y., MULLER, F., FORREST, A. R. R., CARNINCI, P., REHLI, M. & SANDELIN, A. 2014. An atlas of active enhancers across human cell types and tissues. *Nature,* 507**,** 455-461.

AUBLE, D. T., WANG, D., POST, K. W. & HAHN, S. 1997. Molecular analysis of the SNF2/SWI2 protein family member MOT1, an ATP-driven enzyme that dissociates TATA-binding protein from DNA. *Mol Cell Biol,* 17**,** 4842-51.

AYER, D. E. 1999. Histone deacetylases: transcriptional repression with SINers and NuRDs. *Trends Cell Biol,* 9**,** 193-8.

BAI, B. & LAIHO, M. 2016. Deep Sequencing Analysis of Nucleolar Small RNAs: Bioinformatics. *Methods Mol Biol,* 1455**,** 243-8.

BALWIERZ, P. J., CARNINCI, P., DAUB, C. O., KAWAI, J., HAYASHIZAKI, Y., VAN BELLE, W., BEISEL, C. & VAN NIMWEGEN, E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol,* 10**,** R79.

BANNISTER, A. J. & KOUZARIDES, T. 2011. Regulation of chromatin by histone modifications. *Cell Res,* 21**,** 381-95.

BELOSTOTSKY, D. 2009. Exosome complex and pervasive transcription in eukaryotic genomes. *Curr Opin Cell Biol,* 21**,** 352-8.

BELOTSERKOVSKAYA, R. & BERGER, S. L. 1999. Interplay between chromatin modifying and remodeling complexes in transcriptional regulation. *Crit Rev Eukaryot Gene Expr,* 9**,** 221-30.

BLACKWOOD, E. M. & KADONAGA, J. T. 1998. Going the distance: a current view of enhancer action. *Science,* 281**,** 60-3.

BOGDANOVIC, O., FERNANDEZ-MINAN, A., TENA, J. J., DE LA CALLE-MUSTIENES, E., HIDALGO, C., VAN KRUYSBERGEN, I., VAN HEERINGEN, S. J., VEENSTRA, G. J. & GOMEZ-SKARMETA, J. L. 2012. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res,* 22**,** 2043-53.

BOGU, G. K., VIZAN, P., STANTON, L. W., BEATO, M., DI CROCE, L. & MARTI-RENOM, M. A. 2015. Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. *Mol Cell Biol,* 36**,** 809-19.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.

BROCKDORFF, N. & TURNER, B. M. 2015. Dosage compensation in mammals. *Cold Spring Harb Perspect Biol,* 7**,** a019406.

BROWN, S. A., WEIRICH, C. S., NEWTON, E. M. & KINGSTON, R. E. 1998. Transcriptional activation domains stimulate initiation and elongation at different times and via different residues. *EMBO J,* 17**,** 3146-54.

BURATOWSKI, S. 1994. The basics of basal transcription by RNA polymerase II. *Cell,* 77**,** 1-3.

BUSHATI, N., STARK, A., BRENNECKE, J. & COHEN, S. M. 2008. Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila. *Curr Biol,* 18**,** 501-6.

CABILI, M. N., TRAPNELL, C., GOFF, L., KOZIOL, M., TAZON-VEGA, B., REGEV, A. & RINN, J. L. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev,* 25**,** 1915-27.

CAIRNS, B. R. 1998. Chromatin remodeling machines: similar motors, ulterior motives. *Trends Biochem Sci,* 23**,** 20-5.

CAO, R., WANG, L., WANG, H., XIA, L., ERDJUMENT-BROMAGE, H., TEMPST, P., JONES, R. S. & ZHANG, Y. 2002. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science,* 298**,** 1039-43.

CAREY, M. 1998. The enhanceosome and transcriptional synergy. *Cell,* 92**,** 5-8.

CARLSON, M. 1997. Genetics of transcriptional regulation in yeast: connections to the RNA polymerase II CTD. *Annu Rev Cell Dev Biol,* 13**,** 1-23.

CARNINCI, P. 2006. Tagging mammalian transcription complexity. *Trends Genet,* 22**,** 501-10.

CARNINCI, P., KVAM, C., KITAMURA, A., OHSUMI, T., OKAZAKI, Y., ITOH, M., KAMIYA, M., SHIBATA, K., SASAKI, N., IZAWA, M., MURAMATSU, M., HAYASHIZAKI, Y. & SCHNEIDER, C. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics,* 37**,** 327-36.

CARNINCI, P., SANDELIN, A., LENHARD, B., KATAYAMA, S., SHIMOKAWA, K., PONJAVIC, J., SEMPLE, C. A., TAYLOR, M. S., ENGSTROM, P. G., FRITH, M. C., FORREST, A. R., ALKEMA, W. B., TAN, S. L., PLESSY, C., KODZIUS, R., RAVASI, T., KASUKAWA, T., FUKUDA, S., KANAMORI-KATAYAMA, M., KITAZUME, Y., KAWAJI, H., KAI, C., NAKAMURA, M., KONNO, H., NAKANO, K., MOTTAGUI-TABAR, S., ARNER, P., CHESI, A., GUSTINCICH, S., PERSICHETTI, F., SUZUKI, H., GRIMMOND, S. M., WELLS, C. A., ORLANDO, V., WAHLESTEDT, C., LIU, E. T., HARBERS, M., KAWAI, J., BAJIC, V. B., HUME, D. A. & HAYASHIZAKI, Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet,* 38**,** 626-35.

CARTHEW, R. W. & SONTHEIMER, E. J. 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell,* 136**,** 642-55.

CONAWAY, R. C. & CONAWAY, J. W. 1993. General initiation factors for RNA polymerase II. *Annu Rev Biochem,* 62**,** 161-90.

CORDEN, J. L. 1990. Tails of RNA polymerase II. *Trends Biochem Sci,* 15**,** 383-7.

CORDEN, J. L. & PATTURAJAN, M. 1997. A CTD function linking transcription to splicing. *Trends Biochem Sci,* 22**,** 413-6.

CORMACK, B. P. & STRUHL, K. 1992. The TATA-binding protein is required for transcription by all three nuclear RNA polymerases in yeast cells. *Cell,* 69**,** 685-96.

CORSETTI, M. T., BRIATA, P., SANSEVERINO, L., DAGA, A., AIROLDI, I., SIMEONE, A., PALMISANO, G., ANGELINI, C., BONCINELLI, E. & CORTE, G. 1992. Differential DNA binding properties of three human homeodomain proteins. *Nucleic Acids Res,* 20**,** 4465-72.

DE RENZIS, S., ELEMENTO, O., TAVAZOIE, S. & WIESCHAUS, E. F. 2007. Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS Biol,* 5**,** e117.

DERRIEN, T., GUIGO, R. & JOHNSON, R. 2011. The Long Non-Coding RNAs: A New (P)layer in the "Dark Matter". *Front Genet,* 2**,** 107.

DERRIEN, T., JOHNSON, R., BUSSOTTI, G., TANZER, A., DJEBALI, S., TILGNER, H., GUERNEC, G., MARTIN, D., MERKEL, A., KNOWLES, D. G., LAGARDE, J., VEERAVALLI, L., RUAN, X., RUAN, Y., LASSMANN, T., CARNINCI, P., BROWN, J. B., LIPOVICH, L., GONZALEZ, J. M., THOMAS, M., DAVIS, C. A., SHIEKHATTAR, R., GINGERAS, T. R., HUBBARD, T. J., NOTREDAME, C., HARROW, J. & GUIGO, R. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res,* 22**,** 1775-89.

DESPIC, V., DEJUNG, M., GU, M., KRISHNAN, J., ZHANG, J., HERZEL, L., STRAUBE, K., GERSTEIN, M. B., BUTTER, F. & NEUGEBAUER, K. M. 2017. Dynamic RNA-protein interactions underlie the zebrafish maternal-to-zygotic transition. *Genome Res,* 27**,** 1184-1194.

DHALIWAL, N. K. & MITCHELL, J. A. 2016. Nuclear RNA Isolation and Sequencing. *Methods Mol Biol,* 1402**,** 63-71.

DHIMAN, H., KAPOOR, S., SIVADAS, A., SIVASUBBU, S. & SCARIA, V. 2015. zflncRNApedia: A Comprehensive Online Resource for Zebrafish Long Non-Coding RNAs. *PLoS One,* 10**,** e0129997.

DIERMEIER, S. D. & LANGST, G. 2014. Deep sequencing of small chromatin-associated RNA: bioinformatic analysis. *Methods Mol Biol,* 1094**,** 355-9.

DJEBALI, S., DAVIS, C. A., MERKEL, A., DOBIN, A., LASSMANN, T., MORTAZAVI, A., TANZER, A., LAGARDE, J., LIN, W., SCHLESINGER, F., XUE, C., MARINOV, G. K., KHATUN, J., WILLIAMS, B. A., ZALESKI, C., ROZOWSKY, J., RODER, M., KOKOCINSKI, F., ABDELHAMID, R. F., ALIOTO, T., ANTOSHECHKIN, I., BAER, M. T., BAR, N. S., BATUT, P., BELL, K., BELL, I., CHAKRABORTTY, S., CHEN, X., CHRAST, J., CURADO, J., DERRIEN, T., DRENKOW, J., DUMAIS, E., DUMAIS, J., DUTTAGUPTA, R., FALCONNET, E., FASTUCA, M., FEJES-TOTH, K., FERREIRA, P., FOISSAC, S., FULLWOOD, M. J., GAO, H., GONZALEZ, D., GORDON, A., GUNAWARDENA, H., HOWALD, C., JHA, S., JOHNSON, R., KAPRANOV, P., KING, B., KINGSWOOD, C., LUO, O. J., PARK, E., PERSAUD, K., PREALL, J. B., RIBECA, P., RISK, B., ROBYR, D., SAMMETH, M., SCHAFFER, L., SEE, L. H., SHAHAB, A., SKANCKE, J., SUZUKI, A. M., TAKAHASHI, H., TILGNER, H., TROUT, D., WALTERS, N., WANG, H., WROBEL, J., YU, Y., RUAN, X., HAYASHIZAKI, Y., HARROW, J., GERSTEIN, M., HUBBARD, T., REYMOND, A., ANTONARAKIS, S. E., HANNON, G., GIDDINGS, M. C., RUAN, Y., WOLD, B., CARNINCI, P., GUIGO, R. & GINGERAS, T. R. 2012. Landscape of transcription in human cells. *Nature,* 489**,** 101-8.

DOBIN, A. & GINGERAS, T. R. 2015. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics,* 51**,** 11 14 1-19.

DRIEVER, W., SOLNICA-KREZEL, L., SCHIER, A. F., NEUHAUSS, S. C., MALICKI, J., STEMPLE, D. L., STAINIER, D. Y., ZWARTKRUIS, F., ABDELILAH, S., RANGINI, Z., BELAK, J. & BOGGS, C. 1996. A genetic screen for mutations affecting embryogenesis in zebrafish. *Development,* 123**,** 37-46.

DUTTKE, S. H. C., LACADIE, S. A., IBRAHIM, M. M., GLASS, C. K., CORCORAN, D. L., BENNER, C., HEINZ, S., KADONAGA, J. T. & OHLER, U. 2015. Human promoters are intrinsically directional. *Mol Cell,* 57**,** 674-684.

DUVAL, C., BOUVET, P., OMILLI, F., ROGHI, C., DOREL, C., LEGUELLEC, R., PARIS, J. & OSBORNE, H. B. 1990. Stability of maternal mRNA in Xenopus embryos: role of transcription and translation. *Mol Cell Biol,* 10**,** 4123-9.

EBISUYA, M., YAMAMOTO, T., NAKAJIMA, M. & NISHIDA, E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol,* 10**,** 1106-13.

FAGHIHI, M. A. & WAHLESTEDT, C. 2009. Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol,* 10**,** 637-43.

FATICA, A. & BOZZONI, I. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet,* 15**,** 7-21.

FERNANDES, J. C. R., ACUNA, S. M., AOKI, J. I., FLOETER-WINTER, L. M. & MUXEL, S. M. 2019. Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease. *Noncoding RNA,* 5.

FIRE, A., XU, S., MONTGOMERY, M. K., KOSTAS, S. A., DRIVER, S. E. & MELLO, C. C. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature,* 391**,** 806-11.

GENG, Y. & JOHNSON, L. F. 1993. Lack of an initiator element is responsible for multiple transcriptional initiation sites of the TATA-less mouse thymidylate synthase promoter. *Mol Cell Biol,* 13**,** 4894-903.

GHILDIYAL, M. & ZAMORE, P. D. 2009. Small silencing RNAs: an expanding universe. *Nat Rev Genet,* 10**,** 94-108.

GIRALDEZ, A. J., MISHIMA, Y., RIHEL, J., GROCOCK, R. J., VAN DONGEN, S., INOUE, K., ENRIGHT, A. J. & SCHIER, A. F. 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science,* 312**,** 75-9.

GLOSS, B. S., SIGNAL, B., CHEETHAM, S. W., GRUHL, F., KACZOROWSKI, D. C., PERKINS, A. C. & DINGER, M. E. 2017. High resolution temporal transcriptomics of mouse embryoid body development reveals complex expression dynamics of coding and noncoding loci. *Sci Rep,* 7**,** 6731.

GREENBLATT, J. 1997. RNA polymerase II holoenzyme and transcriptional regulation. *Curr Opin Cell Biol,* 9**,** 310-9.

GUPTA, R. A., SHAH, N., WANG, K. C., KIM, J., HORLINGS, H. M., WONG, D. J., TSAI, M. C., HUNG, T., ARGANI, P., RINN, J. L., WANG, Y., BRZOSKA, P., KONG, B., LI, R., WEST, R. B., VAN DE VIJVER, M. J., SUKUMAR, S. & CHANG, H. Y. 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature,* 464**,** 1071-6.

GUTTMAN, M., DONAGHEY, J., CAREY, B. W., GARBER, M., GRENIER, J. K., MUNSON, G., YOUNG, G., LUCAS, A. B., ACH, R., BRUHN, L., YANG, X., AMIT, I., MEISSNER, A., REGEV, A., RINN, J. L., ROOT, D. E. & LANDER, E. S. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature,* 477**,** 295-300.

GUTTMAN, M. & RINN, J. L. 2012. Modular regulatory principles of large non-coding RNAs. *Nature,* 482**,** 339-46.

GUTTMAN, M., RUSSELL, P., INGOLIA, N. T., WEISSMAN, J. S. & LANDER, E. S. 2013. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell,* 154**,** 240-51.

HABERLE, V., FORREST, A. R., HAYASHIZAKI, Y., CARNINCI, P. & LENHARD, B. 2015. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res,* 43**,** e51.

HABERLE, V., LI, N., HADZHIEV, Y., PLESSY, C., PREVITI, C., NEPAL, C., GEHRIG, J., DONG, X., AKALIN, A., SUZUKI, A. M., VAN, I. W. F. J., ARMANT, O., FERG, M., STRAHLE, U., CARNINCI, P., MULLER, F. & LENHARD, B. 2014. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature,* 507**,** 381-385.

HAMMERSCHMIDT, M., PELEGRI, F., MULLINS, M. C., KANE, D. A., BRAND, M., VAN EEDEN, F. J., FURUTANI-SEIKI, M., GRANATO, M., HAFFTER, P., HEISENBERG, C. P., JIANG, Y. J., KELSH, R. N., ODENTHAL, J., WARGA, R. M. & NUSSLEIN-VOLHARD, C. 1996. Mutations affecting morphogenesis during gastrulation and tail formation in the zebrafish, Danio rerio. *Development,* 123**,** 143-51.

HAN, J. S., SZAK, S. T. & BOEKE, J. D. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature,* 429**,** 268-74.

HANNA-ROSE, W. & HANSEN, U. 1996. Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet,* 12**,** 229-34.

HANSEN, U. 1996. Mechanisms of eukaryotic transcription: surfaces, complexes, and contexts. (Cold Spring Harbor Cancer Cells Meeting, August 30-September 3, 1995). *Biochim Biophys Acta,* 1287**,** 59-62.

HAQUE, S., KAUSHIK, K., LEONARD, V. E., KAPOOR, S., SIVADAS, A., JOSHI, A., SCARIA, V. & SIVASUBBU, S. 2014. Short stories on zebrafish long noncoding RNAs. *Zebrafish,* 11**,** 499-508.

HARVEY, S. A., SEALY, I., KETTLEBOROUGH, R., FENYES, F., WHITE, R., STEMPLE, D. & SMITH, J. C. 2013. Identification of the zebrafish maternal and paternal transcriptomes. *Development,* 140**,** 2703-10.

HENGARTNER, C. J., MYER, V. E., LIAO, S. M., WILSON, C. J., KOH, S. S. & YOUNG, R. A. 1998. Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. *Mol Cell,* 2**,** 43-53.

HENIKOFF, S., AHMAD, K. & MALIK, H. S. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science,* 293**,** 1098-102.

HU, W., ALVAREZ-DOMINGUEZ, J. R. & LODISH, H. F. 2012. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep,* 13**,** 971-83.

HU, X., CHEN, W., LI, J., HUANG, S., XU, X., ZHANG, X., XIANG, S. & LIU, C. 2018. ZFLNC: a comprehensive and well-annotated database for zebrafish lncRNA. *Database (Oxford),* 2018.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

HUANG DA, W., SHERMAN, B. T., ZHENG, X., YANG, J., IMAMICHI, T., STEPHENS, R. & LEMPICKI, R. A. 2009b. Extracting biological meaning from large gene lists with DAVID. *Curr Protoc Bioinformatics,* Chapter 13**,** Unit 13 11.

HUNTZINGER, E. & IZAURRALDE, E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet,* 12**,** 99-110.

HUTVAGNER, G. & SIMARD, M. J. 2008. Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol,* 9**,** 22-32.

JADY, B. E. & KISS, T. 2001. A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. *EMBO J,* 20**,** 541-51.

JAYASENA, C. S. & BRONNER, M. E. 2012. Rbms3 functions in craniofacial development by posttranscriptionally modulating TGF-beta signaling. *J Cell Biol,* 199**,** 453-66.

JOHNSON, S. M., GROSSHANS, H., SHINGARA, J., BYROM, M., JARVIS, R., CHENG, A., LABOURIER, E., REINERT, K. L., BROWN, D. & SLACK, F. J. 2005. RAS is regulated by the let-7 microRNA family. *Cell,* 120**,** 635-47.

JOHNSTONE, O. & LASKO, P. 2001. Translational regulation and RNA localization in Drosophila oocytes and embryos. *Annu Rev Genet,* 35**,** 365-406.

JUVEN-GERSHON, T., HSU, J. Y., THEISEN, J. W. & KADONAGA, J. T. 2008. The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol,* 20**,** 253-9.

KADONAGA, J. T. 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell,* 116**,** 247-57.

KANE, D. A., HAMMERSCHMIDT, M., MULLINS, M. C., MAISCHEIN, H. M., BRAND, M., VAN EEDEN, F. J., FURUTANI-SEIKI, M., GRANATO, M., HAFFTER, P., HEISENBERG, C. P., JIANG, Y. J., KELSH, R. N., ODENTHAL, J., WARGA, R. M. & NUSSLEIN-VOLHARD, C. 1996a. The zebrafish epiboly mutants. *Development,* 123**,** 47-55.

KANE, D. A., MAISCHEIN, H. M., BRAND, M., VAN EEDEN, F. J., FURUTANI-SEIKI, M., GRANATO, M., HAFFTER, P., HAMMERSCHMIDT, M., HEISENBERG, C. P., JIANG, Y. J., KELSH, R. N., MULLINS, M. C., ODENTHAL, J., WARGA, R. M. & NUSSLEIN-VOLHARD, C. 1996b. The zebrafish early arrest mutants. *Development,* 123**,** 57-66.

KANHERE, A., VIIRI, K., ARAUJO, C. C., RASAIYAAH, J., BOUWMAN, R. D., WHYTE, W. A., PEREIRA, C. F., BROOKES, E., WALKER, K., BELL, G. W., POMBO, A., FISHER, A. G., YOUNG, R. A. & JENNER, R. G. 2010. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell,* 38**,** 675-88.

KATAYAMA, S., TOMARU, Y., KASUKAWA, T., WAKI, K., NAKANISHI, M., NAKAMURA, M., NISHIDA, H., YAP, C. C., SUZUKI, M., KAWAI, J., SUZUKI, H., CARNINCI, P., HAYASHIZAKI, Y., WELLS, C., FRITH, M., RAVASI, T., PANG, K. C., HALLINAN, J., MATTICK, J., HUME, D. A., LIPOVICH, L., BATALOV, S., ENGSTROM, P. G., MIZUNO, Y., FAGHIHI, M. A., SANDELIN, A., CHALK, A. M., MOTTAGUI-TABAR, S., LIANG, Z., LENHARD, B., WAHLESTEDT, C., GROUP, R. G. E. R., GENOME SCIENCE, G. & CONSORTIUM, F. 2005. Antisense transcription in the mammalian transcriptome. *Science,* 309**,** 1564-6.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

KERPPOLA, T. K. & KANE, C. M. 1991. RNA polymerase: regulation of transcript elongation and termination. *FASEB J,* 5**,** 2833-42.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol,* 14**,** R36.

KIMMEL, C. B., BALLARD, W. W., KIMMEL, S. R., ULLMANN, B. & SCHILLING, T. F. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn,* 203**,** 253-310.

KISS, T. 2002. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell,* 109**,** 145-8.

KIYOSAWA, H., MISE, N., IWASE, S., HAYASHIZAKI, Y. & ABE, K. 2005. Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res,* 15**,** 463-74.

KNIGHT, R. D., JAVIDAN, Y., ZHANG, T., NELSON, S. & SCHILLING, T. F. 2005. AP2-dependent signals from the ectoderm regulate craniofacial development in the zebrafish embryo. *Development,* 132**,** 3127-38.

KODZIUS, R., KOJIMA, M., NISHIYORI, H., NAKAMURA, M., FUKUDA, S., TAGAMI, M., SASAKI, D., IMAMURA, K., KAI, C., HARBERS, M., HAYASHIZAKI, Y. & CARNINCI, P. 2006. CAGE: cap analysis of gene expression. *Nat Methods,* 3**,** 211-22.

KOLESKE, A. J., CHAO, D. M. & YOUNG, R. A. 1996. Purification of yeast RNA polymerase II holoenzymes. *Methods Enzymol,* 273**,** 176-84.

KOPRUNNER, M., THISSE, C., THISSE, B. & RAZ, E. 2001. A zebrafish nanos-related gene is essential for the development of primordial germ cells. *Genes Dev,* 15**,** 2877-85.

KORNBERG, R. D. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science,* 184**,** 868-71.

KORNBERG, R. D. & LORCH, Y. 1999. Chromatin-modifying and -remodeling complexes. *Curr Opin Genet Dev,* 9**,** 148-51.

KURYAN, B. G., KIM, J., TRAN, N. N., LOMBARDO, S. R., VENKATESH, S., WORKMAN, J. L. & CAREY, M. 2012. Histone density is maintained during transcription mediated by the chromatin remodeler RSC and histone chaperone NAP1 in vitro. *Proc Natl Acad Sci U S A,* 109**,** 1931-6.

LALANCETTE, C., MILLER, D., LI, Y. & KRAWETZ, S. A. 2008. Paternal contributions: new functional insights for spermatozoal RNA. *J Cell Biochem,* 104**,** 1570-9.

LAVENDER, C. A., CANNADY, K. R., HOFFMAN, J. A., TROTTER, K. W., GILCHRIST, D. A., BENNETT, B. D., BURKHOLDER, A. B., BURD, C. J., FARGO, D. C. & ARCHER, T. K. 2016. Downstream Antisense Transcription Predicts Genomic Features That Define the Specific Chromatin Environment at Mammalian Promoters. *PLoS Genet,* 12**,** e1006224.

LAWRENCE, M., HUBER, W., PAGES, H., ABOYOUN, P., CARLSON, M., GENTLEMAN, R., MORGAN, M. T. & CAREY, V. J. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol,* 9**,** e1003118.

LECUYER, E., YOSHIDA, H., PARTHASARATHY, N., ALM, C., BABAK, T., CEROVINA, T., HUGHES, T. R., TOMANCAK, P. & KRAUSE, H. M. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell,* 131**,** 174-87.

LEE, M. T., BONNEAU, A. R. & GIRALDEZ, A. J. 2014. Zygotic genome activation during the maternal-to-zygotic transition. *Annu Rev Cell Dev Biol,* 30**,** 581-613.

LEE, M. T., BONNEAU, A. R., TAKACS, C. M., BAZZINI, A. A., DIVITO, K. R., FLEMING, E. S. & GIRALDEZ, A. J. 2013. Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature,* 503**,** 360-4.

LEE, T. I., WYRICK, J. J., KOH, S. S., JENNINGS, E. G., GADBOIS, E. L. & YOUNG, R. A. 1998. Interplay of positive and negative regulators in transcription initiation by RNA polymerase II holoenzyme. *Mol Cell Biol,* 18**,** 4455-62.

LEE, T. I. & YOUNG, R. A. 2000. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet,* 34**,** 77-137.

LEEB, M., STEFFEN, P. A. & WUTZ, A. 2009. X chromosome inactivation sparked by non-coding RNAs. *RNA Biol,* 6**,** 94-9.

LEWIS, B. A., KIM, T. K. & ORKIN, S. H. 2000. A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc Natl Acad Sci U S A,* 97**,** 7172-7.

LI, K., BLUM, Y., VERMA, A., LIU, Z., PRAMANIK, K., LEIGH, N. R., CHUN, C. Z., SAMANT, G. V., ZHAO, B., GARNAAS, M. K., HORSWILL, M. A., STANHOPE, S. A., NORTH, P. E., MIAO, R. Q., WILKINSON, G. A., AFFOLTER, M. & RAMCHANDRAN, R. 2010. A noncoding antisense RNA in tie-1 locus regulates tie-1 function in vivo. *Blood,* 115**,** 133-9.

LIM, C. Y., SANTOSO, B., BOULAY, T., DONG, E., OHLER, U. & KADONAGA, J. T. 2004. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev,* 18**,** 1606-17.

LONG, R. M., SINGER, R. H., MENG, X., GONZALEZ, I., NASMYTH, K. & JANSEN, R. P. 1997. Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. *Science,* 277**,** 383-7.

LUGER, K., RECHSTEINER, T. J., FLAUS, A. J., WAYE, M. M. & RICHMOND, T. J. 1997. Characterization of nucleosome core particles containing histone proteins made in bacteria. *J Mol Biol,* 272**,** 301-11.

LUGER, K. & RICHMOND, T. J. 1998. DNA binding within the nucleosome core. *Curr Opin Struct Biol,* 8**,** 33-40.

LUND, E., LIU, M., HARTLEY, R. S., SHEETS, M. D. & DAHLBERG, J. E. 2009. Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. *RNA,* 15**,** 2351-63.

LUO, S., LU, J. Y., LIU, L., YIN, Y., CHEN, C., HAN, X., WU, B., XU, R., LIU, W., YAN, P., SHAO, W., LU, Z., LI, H., NA, J., TANG, F., WANG, J., ZHANG, Y. E. & SHEN, X. 2016. Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell,* 18**,** 637-52.

MAAMAR, H., CABILI, M. N., RINN, J. & RAJ, A. 2013. linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis. *Genes Dev,* 27**,** 1260-71.

MALONE, C. D. & HANNON, G. J. 2009a. Molecular evolution of piRNA and transposon control pathways in Drosophila. *Cold Spring Harb Symp Quant Biol,* 74**,** 225-34.

MALONE, C. D. & HANNON, G. J. 2009b. Small RNAs as guardians of the genome. *Cell,* 136**,** 656-68.

MARLOW, F. L. & MULLINS, M. C. 2008. Bucky ball functions in Balbiani body assembly and animal-vegetal polarity in the oocyte and follicle cell layer in zebrafish. *Dev Biol,* 321**,** 40-50.

MARTIANOV, I., RAMADASS, A., SERRA BARROS, A., CHOW, N. & AKOULITCHEV, A. 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature,* 445**,** 666-70.

MATERA, A. G., TERNS, R. M. & TERNS, M. P. 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol,* 8**,** 209-20.

MATHAVAN, S., LEE, S. G., MAK, A., MILLER, L. D., MURTHY, K. R., GOVINDARAJAN, K. R., TONG, Y., WU, Y. L., LAM, S. H., YANG, H., RUAN, Y., KORZH, V., GONG, Z., LIU, E. T. & LUFKIN, T. 2005. Transcriptome analysis of zebrafish embryogenesis using microarrays. *PLoS Genet,* 1**,** 260-76.

MAYER, A., DI IULIO, J., MALERI, S., ESER, U., VIERSTRA, J., REYNOLDS, A., SANDSTROM, R., STAMATOYANNOPOULOS, J. A. & CHURCHMAN, L. S. 2015. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell,* 161**,** 541-554.

MEI, W., LEE, K. W., MARLOW, F. L., MILLER, A. L. & MULLINS, M. C. 2009. hnRNP I is required to generate the Ca2+ signal that causes egg activation in zebrafish. *Development,* 136**,** 3007-17.

MERCER, T. R., DINGER, M. E. & MATTICK, J. S. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet,* 10**,** 155-9.

MITCHELL, J. A., CLAY, I., UMLAUF, D., CHEN, C. Y., MOIR, C. A., ESKIW, C. H., SCHOENFELDER, S., CHAKALOVA, L., NAGANO, T. & FRASER, P. 2012. Nuclear RNA sequencing of the mouse erythroid cell transcriptome. *PLoS One,* 7**,** e49274.

MOHAMMAD, F., MONDAL, T. & KANDURI, C. 2009. Epigenetics of imprinted long non-coding RNAs. *Epigenetics,* 4**,** 277-286.

MONDAL, T., RASMUSSEN, M., PANDEY, G. K., ISAKSSON, A. & KANDURI, C. 2010. Characterization of the RNA content of chromatin. *Genome Res,* 20**,** 899-907.

NAGANO, T., MITCHELL, J. A., SANZ, L. A., PAULER, F. M., FERGUSON-SMITH, A. C., FEIL, R. & FRASER, P. 2008. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science,* 322**,** 1717-20.

NAGASHIMA, M., MAWATARI, K., TANAKA, M., HIGASHI, T., SAITO, H., MURAMOTO, K., MATSUKAWA, T., KORIYAMA, Y., SUGITANI, K. & KATO, S. 2009. Purpurin is a key molecule for cell differentiation during the early development of zebrafish retina. *Brain Res,* 1302**,** 54-63.

NEPAL, C., HADZHIEV, Y., PREVITI, C., HABERLE, V., LI, N., TAKAHASHI, H., SUZUKI, A. M., SHENG, Y., ABDELHAMID, R. F., ANAND, S., GEHRIG, J., AKALIN, A., KOCKX, C. E., VAN DER SLOOT, A. A., VAN IJCKEN, W. F., ARMANT, O., RASTEGAR, S., WATSON, C., STRAHLE, U., STUPKA, E., CARNINCI, P., LENHARD, B. & MULLER, F. 2013. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res,* 23**,** 1938-50.

NEWPORT, J. & KIRSCHNER, M. 1982. A major developmental transition in early Xenopus embryos: II. Control of the onset of transcription. *Cell,* 30**,** 687-96.

NIGUMANN, P., REDIK, K., MATLIK, K. & SPEEK, M. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics,* 79**,** 628-34.

OGBOURNE, S. & ANTALIS, T. M. 1998. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J,* 331 (Pt 1)**,** 1-14.

ORGEL, L. E. & CRICK, F. H. 1980. Selfish DNA: the ultimate parasite. *Nature,* 284**,** 604-7.

ORKIN, S. H. & HOCHEDLINGER, K. 2011. Chromatin connections to pluripotency and cellular reprogramming. *Cell,* 145**,** 835-50.

OROM, U. A., DERRIEN, T., BERINGER, M., GUMIREDDY, K., GARDINI, A., BUSSOTTI, G., LAI, F., ZYTNICKI, M., NOTREDAME, C., HUANG, Q., GUIGO, R. & SHIEKHATTAR, R. 2010a. Long noncoding RNAs with enhancer-like function in human cells. *Cell,* 143**,** 46-58.

OROM, U. A., DERRIEN, T., GUIGO, R. & SHIEKHATTAR, R. 2010b. Long noncoding RNAs as enhancers of gene expression. *Cold Spring Harb Symp Quant Biol,* 75**,** 325-31.

PAILLARD, L., OMILLI, F., LEGAGNEUX, V., BASSEZ, T., MANIEY, D. & OSBORNE, H. B. 1998. EDEN and EDEN-BP, a cis element and an associated factor that mediate sequence-specific mRNA deadenylation in Xenopus embryos. *EMBO J,* 17**,** 278-87.

PANDEY, R. R., MONDAL, T., MOHAMMAD, F., ENROTH, S., REDRUP, L., KOMOROWSKI, J., NAGANO, T., MANCINI-DINARDO, D. & KANDURI, C. 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell,* 32**,** 232-46.

PAULI, A., RINN, J. L. & SCHIER, A. F. 2011. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet,* 12**,** 136-49.

PAULI, A., VALEN, E., LIN, M. F., GARBER, M., VASTENHOUW, N. L., LEVIN, J. Z., FAN, L., SANDELIN, A., RINN, J. L., REGEV, A. & SCHIER, A. F. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res,* 22**,** 577-91.

PELISSON, A., MEJLUMIAN, L., ROBERT, V., TERZIAN, C. & BUCHETON, A. 2002. Drosophila germline invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. *Insect Biochem Mol Biol,* 32**,** 1249-56.

PENG, G. & WESTERFIELD, M. 2006. Lhx5 promotes forebrain development and activates transcription of secreted Wnt antagonists. *Development,* 133**,** 3191-200.

PENKOV, D., NI, R., ELSE, C., PINOL-ROMA, S., RAMIREZ, F. & TANAKA, S. 2000. Cloning of a human gene closely related to the genes coding for the c-myc single-strand binding proteins. *Gene,* 243**,** 27-36.

PERRY, R. B. & ULITSKY, I. 2016. The functions of long noncoding RNAs in development and stem cells. *Development,* 143**,** 3882-3894.

PERTEA, M., KIM, D., PERTEA, G. M., LEEK, J. T. & SALZBERG, S. L. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc,* 11**,** 1650-67.

PERTEA, M., PERTEA, G. M., ANTONESCU, C. M., CHANG, T. C., MENDELL, J. T. & SALZBERG, S. L. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol,* 33**,** 290-5.

PONJAVIC, J., OLIVER, P. L., LUNTER, G. & PONTING, C. P. 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet,* 5**,** e1000617.

PRIOLEAU, M. N., HUET, J., SENTENAC, A. & MECHALI, M. 1994. Competition between chromatin and transcription complex assembly regulates gene expression during early development. *Cell,* 77**,** 439-49.

PTASHNE, M. & GANN, A. 1997. Transcriptional activation by recruitment. *Nature,* 386**,** 569-77.

QI, L. S. & ARKIN, A. P. 2014. A versatile framework for microbial engineering using synthetic non-coding RNAs. *Nat Rev Microbiol,* 12**,** 341-54.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

RAHA, D., HONG, M. & SNYDER, M. 2010. ChIP-Seq: a method for global identification of regulatory elements in the genome. *Curr Protoc Mol Biol,* Chapter 21**,** Unit 21 19 1-14.

RAMIREZ, F., RYAN, D. P., GRUNING, B., BHARDWAJ, V., KILPERT, F., RICHTER, A. S., HEYNE, S., DUNDAR, F. & MANKE, T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res,* 44**,** W160-5.

REEVE, J. N. 2003. Archaeal chromatin and transcription. *Mol Microbiol,* 48**,** 587-98.

RICHARDSON, M. K., HANKEN, J., GOONERATNE, M. L., PIEAU, C., RAYNAUD, A., SELWOOD, L. & WRIGHT, G. M. 1997. There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat Embryol (Berl),* 196**,** 91-106.

RINN, J. L., KERTESZ, M., WANG, J. K., SQUAZZO, S. L., XU, X., BRUGMANN, S. A., GOODNOUGH, L. H., HELMS, J. A., FARNHAM, P. J., SEGAL, E. & CHANG, H. Y. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell,* 129**,** 1311-23.

ROY, K. & CHANFREAU, G. F. 2012. The Diverse Functions of Fungal RNase III Enzymes in RNA Metabolism. *Enzymes,* 31**,** 213-35.

RUZOV, A., DUNICAN, D. S., PROKHORTCHOUK, A., PENNINGS, S., STANCHEVA, I., PROKHORTCHOUK, E. & MEEHAN, R. R. 2004. Kaiso is a genome-wide repressor of transcription that is essential for amphibian development. *Development,* 131**,** 6185-94.

RYBAK, A., FUCHS, H., SMIRNOVA, L., BRANDT, C., POHL, E. E., NITSCH, R. & WULCZYN, F. G. 2008. A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. *Nat Cell Biol,* 10**,** 987-93.

SAN, B., CHRISPIJN, N. D., WITTKOPP, N., VAN HEERINGEN, S. J., LAGENDIJK, A. K., ABEN, M., BAKKERS, J., KETTING, R. F. & KAMMINGA, L. M. 2016. Normal formation of a vertebrate body plan and loss of tissue maintenance in the absence of ezh2. *Sci Rep,* 6**,** 24658.

SANA, J., FALTEJSKOVA, P., SVOBODA, M. & SLABY, O. 2012. Novel classes of non-coding RNAs and cancer. *J Transl Med,* 10**,** 103.

SCHARF, S. R. & GERHART, J. C. 1980. Determination of the dorsal-ventral axis in eggs of Xenopus laevis: complete rescue of uv-impaired eggs by oblique orientation before first cleavage. *Dev Biol,* 79**,** 181-98.

SCRUGGS, B. S., GILCHRIST, D. A., NECHAEV, S., MUSE, G. W., BURKHOLDER, A., FARGO, D. C. & ADELMAN, K. 2015. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol Cell,* 58**,** 1101-12.

SEYDOUX, G. & FIRE, A. 1994. Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. *Development,* 120**,** 2823-34.

SHEARWIN, K. E., CALLEN, B. P. & EGAN, J. B. 2005. Transcriptional interference--a crash course. *Trends Genet,* 21**,** 339-45.

SKELDING, Z., SARNOVSKY, R. & CRAIG, N. L. 2002. Formation of a nucleoprotein complex containing Tn7 and its target DNA regulates transposition initiation. *EMBO J,* 21**,** 3494-504.

SLACK, J. M., HOLLAND, P. W. & GRAHAM, C. F. 1993. The zootype and the phylotypic stage. *Nature,* 361**,** 490-2.

SMOLLE, M. & WORKMAN, J. L. 2013. Transcription-associated histone modifications and cryptic transcription. *Biochim Biophys Acta,* 1829**,** 84-97.

SOLNICA-KREZEL, L., STEMPLE, D. L., MOUNTCASTLE-SHAH, E., RANGINI, Z., NEUHAUSS, S. C., MALICKI, J., SCHIER, A. F., STAINIER, D. Y., ZWARTKRUIS, F., ABDELILAH, S. & DRIEVER, W. 1996. Mutations affecting cell fates and cellular rearrangements during gastrulation in zebrafish. *Development,* 123**,** 67-80.

SPENCER, V. A. & DAVIE, J. R. 1999. Role of covalent modifications of histones in regulating gene expression. *Gene,* 240**,** 1-12.

STEMPLE, D. L., SOLNICA-KREZEL, L., ZWARTKRUIS, F., NEUHAUSS, S. C., SCHIER, A. F., MALICKI, J., STAINIER, D. Y., ABDELILAH, S., RANGINI, Z., MOUNTCASTLE-SHAH, E. & DRIEVER, W. 1996. Mutations affecting development of the notochord in zebrafish. *Development,* 123**,** 117-28.

STOLL, S. J., BARTSCH, S. & KROLL, J. 2013. HOXC9 regulates formation of parachordal lymphangioplasts and the thoracic duct in zebrafish via stabilin 2. *PLoS One,* 8**,** e58311.

STRUBIN, M. & STRUHL, K. 1992. Yeast and human TFIID with altered DNA-binding specificity for TATA elements. *Cell,* 68**,** 721-30.

SUN, B. K., DEATON, A. M. & LEE, J. T. 2006. A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. *Mol Cell,* 21**,** 617-28.

SUZUKI, S., KATO, H., SUZUKI, Y., CHIKASHIGE, Y., HIRAOKA, Y., KIMURA, H., NAGAO, K., OBUSE, C., TAKAHATA, S. & MURAKAMI, Y. 2016. Histone H3K36 trimethylation is essential for multiple silencing mechanisms in fission yeast. *Nucleic Acids Res,* 44**,** 4147-62.

TADROS, W. & LIPSHITZ, H. D. 2009. The maternal-to-zygotic transition: a play in two acts. *Development,* 136**,** 3033-42.

TADROS, W., WESTWOOD, J. T. & LIPSHITZ, H. D. 2007. The mother-to-child transition. *Dev Cell,* 12**,** 847-9.

TAFT, R. J., GLAZOV, E. A., LASSMANN, T., HAYASHIZAKI, Y., CARNINCI, P. & MATTICK, J. S. 2009a. Small RNAs derived from snoRNAs. *RNA,* 15**,** 1233-40.

TAFT, R. J., KAPLAN, C. D., SIMONS, C. & MATTICK, J. S. 2009b. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle,* 8**,** 2332-8.

TAFT, R. J., PANG, K. C., MERCER, T. R., DINGER, M. & MATTICK, J. S. 2010. Non-coding RNAs: regulators of disease. *J Pathol,* 220**,** 126-39.

THISSE, B., HEYER, V., LUX, A., ALUNNI, V., DEGRAVE, A., SEILIEZ, I., KIRCHNER, J., PARKHILL, J. P. & THISSE, C. 2004. Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening. *Methods Cell Biol,* 77**,** 505-19.

TIAN, D., SUN, S. & LEE, J. T. 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell,* 143**,** 390-403.

TOKUSUMI, Y., MA, Y., SONG, X., JACOBSON, R. H. & TAKADA, S. 2007. The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Mol Cell Biol,* 27**,** 1844-58.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc,* 7**,** 562-78.

ULITSKY, I., SHKUMATAVA, A., JAN, C. H., SIVE, H. & BARTEL, D. P. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell,* 147**,** 1537-50.

VASTENHOUW, N. L., ZHANG, Y., WOODS, I. G., IMAM, F., REGEV, A., LIU, X. S., RINN, J. & SCHIER, A. F. 2010. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature,* 464**,** 922-6.

VENKATESH, S., SMOLLE, M., LI, H., GOGOL, M. M., SAINT, M., KUMAR, S., NATARAJAN, K. & WORKMAN, J. L. 2012. Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. *Nature,* 489**,** 452-5.

VENKATESH, S. & WORKMAN, J. L. 2015. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol,* 16**,** 178-89.

VILLEGAS, V. E. & ZAPHIROPOULOS, P. G. 2015. Neighboring gene regulation by antisense long non-coding RNAs. *Int J Mol Sci,* 16**,** 3251-66.

WANG, K. C. & CHANG, H. Y. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell,* 43**,** 904-14.

WANG, K. C., YANG, Y. W., LIU, B., SANYAL, A., CORCES-ZIMMERMAN, R., CHEN, Y., LAJOIE, B. R., PROTACIO, A., FLYNN, R. A., GUPTA, R. A., WYSOCKA, J., LEI, M., DEKKER, J., HELMS, J. A. & CHANG, H. Y. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature,* 472**,** 120-4.

WANG, Z., TOLLERVEY, J., BRIESE, M., TURNER, D. & ULE, J. 2009. CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo. *Methods,* 48**,** 287-93.

WATANABE, S. & PETERSON, C. L. 2010. The INO80 family of chromatin-remodeling enzymes: regulators of histone variant dynamics. *Cold Spring Harb Symp Quant Biol,* 75**,** 35-42.

WAXMAN, J. S., KEEGAN, B. R., ROBERTS, R. W., POSS, K. D. & YELON, D. 2008. Hoxb5b acts downstream of retinoic acid signaling in the forelimb field to restrict heart field potential in zebrafish. *Dev Cell,* 15**,** 923-34.

WEI, N., PANG, W., WANG, Y., XIONG, Y., XU, R., WU, W., ZHAO, C. & YANG, G. 2014. Knockdown of PU.1 mRNA and AS lncRNA regulates expression of immune-related genes in zebrafish Danio rerio. *Dev Comp Immunol,* 44**,** 315-9.

WHITE, R. J., COLLINS, J. E., SEALY, I. M., WALI, N., DOOLEY, C. M., DIGBY, Z., STEMPLE, D. L., MURPHY, D. N., BILLIS, K., HOURLIER, T., FULLGRABE, A., DAVIS, M. P., ENRIGHT, A. J. & BUSCH-NENTWICH, E. M. 2017. A high-resolution mRNA expression time course of embryonic development in zebrafish. *Elife,* 6.

WILUSZ, J. E., SUNWOO, H. & SPECTOR, D. L. 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev,* 23**,** 1494-504.

WINTER, J., JUNG, S., KELLER, S., GREGORY, R. I. & DIEDERICHS, S. 2009. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol,* 11**,** 228-34.

WOOLFE, A., GOODSON, M., GOODE, D. K., SNELL, P., MCEWEN, G. K., VAVOURI, T., SMITH, S. F., NORTH, P., CALLAWAY, H., KELLY, K., WALTER, K., ABNIZOVA, I., GILKS, W., EDWARDS, Y. J., COOKE, J. E. & ELGAR, G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol,* 3**,** e7.

XU, Z., WEI, W., GAGNEUR, J., CLAUDER-MUNSTER, S., SMOLIK, M., HUBER, W. & STEINMETZ, L. M. 2011. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol,* 7**,** 468.

XU, Z., WEI, W., GAGNEUR, J., PEROCCHI, F., CLAUDER-MUNSTER, S., CAMBLONG, J., GUFFANTI, E., STUTZ, F., HUBER, W. & STEINMETZ, L. M. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature,* 457**,** 1033-7.

YABE, T., GE, X., LINDEMAN, R., NAIR, S., RUNKE, G., MULLINS, M. C. & PELEGRI, F. 2009. The maternal-effect gene cellular island encodes aurora B kinase and is essential for furrow formation in the early zebrafish embryo. *PLoS Genet,* 5**,** e1000518.

ZERBINO, D. R., ACHUTHAN, P., AKANNI, W., AMODE, M. R., BARRELL, D., BHAI, J., BILLIS, K., CUMMINS, C., GALL, A., GIRON, C. G., GIL, L., GORDON, L., HAGGERTY, L., HASKELL, E., HOURLIER, T., IZUOGU, O. G., JANACEK, S. H., JUETTEMANN, T., TO, J. K., LAIRD, M. R., LAVIDAS, I., LIU, Z., LOVELAND, J. E., MAUREL, T., MCLAREN, W., MOORE, B., MUDGE, J., MURPHY, D. N., NEWMAN, V., NUHN, M., OGEH, D., ONG, C. K., PARKER, A., PATRICIO, M., RIAT, H. S., SCHUILENBURG, H., SHEPPARD, D., SPARROW, H., TAYLOR, K., THORMANN, A., VULLO, A., WALTS, B., ZADISSA, A., FRANKISH, A., HUNT, S. E., KOSTADIMA, M., LANGRIDGE, N., MARTIN, F. J., MUFFATO, M., PERRY, E., RUFFIER, M., STAINES, D. M., TREVANION, S. J., AKEN, B. L., CUNNINGHAM, F., YATES, A. & FLICEK, P. 2018. Ensembl 2018. *Nucleic Acids Res,* 46**,** D754-D761.

ZHANG, Y., VASTENHOUW, N. L., FENG, J., FU, K., WANG, C., GE, Y., PAULI, A., VAN HUMMELEN, P., SCHIER, A. F. & LIU, X. S. 2014. Canonical nucleosome organization at promoters forms during genome activation. *Genome Res,* 24**,** 260-6.

ZHAO, J., SUN, B. K., ERWIN, J. A., SONG, J. J. & LEE, J. T. 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science,* 322**,** 750-6.