Regulation of transposable elements during early mammalian development

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER FAKULTÄT FÜR BIOLOGIE
DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vorgelegt von
Clara Hermant

May 2025

Diese Dissertation wurde angefertigt

unter der Leitung von Prof. Dr. Maria-Elena Torres-Padilla

am Institut für Epigenetik und Stammzellen

des Helmholtz Zentrum München

Erstgutachter:	Prof. Dr. Maria-Elena Torres-Padilla
Zweitgutachter:	Prof. Dr. John Parsch
Tag der Abgabe:	15/05/2025
Tag der mündlichen Prüfung:	04/12/2025

ERKLÄRUNG

Ich versichere hiermit an Eides statt, dass meine Dissertation selbständig und ohne unerlaubte Hilfsmittel angefertigt worden ist.

Die vorliegende Dissertation wurde weder ganz, noch teilweise bei einer anderen Prüfungskommission vorgelegt.

Ich habe noch zu keinem früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

Wesentliche Teile dieser Arbeit sind in der folgenden Publikation veröffentlicht und zusammengefasst (Hermant et al., 2025):

Hermant C, Mourra-Díaz CM, Oomen ME, Altamirano-Pacheco L, Pal M, Nakatani T, Torres-Padilla ME. The transcription factor SRF regulates MERVL retrotransposons and gene expression during zygotic genome activation. Genes Dev. 2025 Feb 27. doi: 10.1101/gad.352270.124. Epub ahead of print. PMID: 40015990.

München, den 15.05.25 ……………………………………

Clara Hermant

TABLE OF CONTENTS

1 Summary 1

2 Introduction 2

2.1 Transposable elements 2
2.1.1 Classification and transposition mechanisms of TEs 2
2.1.2 Structure of major murine retrotransposons 4
2.1.2.1 Endogenous Retroviruses (ERVs) 4
2.1.2.2 Long Interspersed Nuclear Elements (LINEs) 5
2.1.2.3 Short interspersed elements 6
2.1.3 Impact of transposable elements on mammalian genomes 7

2.2 Mouse preimplantation development 9
2.2.1 Zygotic genome activation 9
2.2.2 Epigenetic reprogramming: a "window of opportunity"? 10

2.3 TE expression in early embryos: from parasites to co-opts 12
2.3.1 Evolutionary pressure for vertical transmission 12
2.3.2 TE cooption during mouse preimplantation development 13
2.3.2.1 A murine ERV as a regulator of ZGA 13
2.3.2.2 Other major retrotransposons families 15

3 Aims 17

4 Results 18

4.1 Identification of candidate TFs regulating TEs during preimplantation development 18
4.1.1 Selection of TE families of interest 18
4.1.2 A footprinting analysis identifies 54 candidate TFs potentially involved in TE regulation 21

4.2 Sequence heterogeneity and transcription factor binding motifs: insights from phylogenetic analysis 27
4.2.1 Phyloregulatory analysis of SINEB2 28
4.2.2 Phyloregulatory analysis of MaLRs 33
4.2.3 Phyloregulatory analysis of MT2_Mm 38
4.2.4 Uncovering the evolutionary history of MT2_Mm 41

4.3 Investigating TFs functions in regulating TEs in cell culture models 52
4.3.1 Targeted gain-of-function screen for TFs activating TE transcription in mESCs 52
4.3.2 SRF activates MT2_Mm and MT2C_Mm in reporter assays 55

4.4 A detailed expression profiling reveals that most candidate TFs are expressed in the embryo 62
4.4.1 Expression profiles at the mRNA level 62
4.4.2 Expression profiles at the protein level 66

4.5 Addressing the function of TFs in regulating genes and TEs in the early embryo: SRF, FOXJ3 and TBP 73
4.5.1 SRF regulates genes and MERVL retrotransposons during ZGA 73
4.5.1.1 A dominant negative strategy for SRF loss-of-function 73
4.5.1.2 SRF regulates genes and is essential for development 76
4.5.1.3 SRF controls TEs and host chimeric transcript expression 82
4.5.2 TBP regulates genes and MaLR during ZGA 89

4.5.2.1 Establishing Trim-Away for TBP and FOXJ3 in the embryo 89
4.5.2.2 Loss of TBP disrupts ZGA and delays development 90
4.5.2.3 TBP directly regulates MaLR in the mouse early embryo 97

5 Discussion 105

5.1 Overlap in TFBS across evolutionary distinct TEs 105

5.2 Contrasting TF impacts: *in-cellulo* vs *in-vivo* models 106

5.3 Challenging consensus sequences through phylogenetic analysis 107

5.4 A complex network of TFs to mediate TE expression in the early embryo 108

5.5 A remarkable cell-type specificity for TF and TE association 111

5.6 The early embryo: robustness and resilience 112

5.7 Concluding remarks 114

6 Material and Methods 115

6.1 Material 115

6.1.1 Chemicals and reagents 115
6.1.2 Kits and enzymes 116
6.1.3 Plasmids 117
6.1.3 Oligonucleotides 119
6.1.3.1 Cloning primers 119
6.1.3.2 qPCR primers 121
6.1.3.3 Smart-seq+5' oligonucleotides 122
6.1.3.4 Antisense oligonucleotides 122
6.1.4 Antibodies 122
6.1.4.1 Primary antibodies 122
6.1.4.2 Secondary antibodies 123
6.1.5 Buffers and media 124
6.1.6 Consumables 125
6.1.7 Equipment 126

6.2 Methods 127

6.2.1 Footprinting analysis 127
6.2.2 Phylogenetic analyses 127
6.2.3 Transcription factor binding site mapping 129
6.2.4 Divergence analysis 129
6.2.5 Consensus sequences and median-joining network analysis 129
6.2.6 Plasmid construction 130
6.2.6.1 TF overexpression plasmids 130
6.2.6.2 Reporter assays plasmids 131
6.2.6.3 Mutagenesis 131
6.2.6.4 *In-vitro* transcribed plasmids 131
6.2.7 Cell culture 132
6.2.7.1 mESCs 132
6.2.7.2 HEK293 132
6.2.8 Plasmid transfections in mESCs 132
6.2.9 RNA extraction, reverse transcription and qPCR 133
6.2.10 Luciferase reporter assay 133
6.2.11 Western blot analysis 134
6.2.12 Embryo collection 134

6.2.13 Immunostainings 134
6.2.13.1 Immunostaining in mESCs 135
6.2.13.2 Immunostaining in embryos 135
6.2.14 In-vitro transcription and antibody purification 135
6.2.15 Embryo manipulation and culture 136
6.2.15.1 Dominant-negative and ASO-mediated LOF 136
6.2.15.2 Reporter assay 136
6.2.15.3 Trim-Away LOF 136
6.2.16 RNAseq (Smart-seq+5') 137
6.2.17 RNA-seq analysis 138
6.2.17.1 ES cells RNA-seq analysis 138
6.2.17.2 Smart-seq2 analysis 138
6.2.17.3 Mapping and processing of Smart-seq+5' 138
6.2.17.4 Differential expression analysis 139
6.2.18 Identification of TE-initiated gene transcripts 141
6.2.19 CUT&Tag 141
6.2.20 CUT&Tag analysis 142
6.2.21 Data deposition 143

7 References 144

8 List of Abbreviations 160

9 Appendices 161

10 Acknowledgements 169

1 Summary

Transposable elements (TEs) are mobile genetic elements that make up a significant portion of mammalian genomes. They have evolved *cis*-regulatory elements to hijack host-encoded transcription machinery for their own expression. While typically silenced in somatic cells, TEs undergo a dramatic and stage-specific surge in transcriptional activity during early mammalian development, particularly around the time of genome activation. Once considered a byproduct of epigenetic reprogramming, TE expression is now emerging as a tightly regulated process with critical biological implications. Understanding the molecular mechanisms mediating TE expression, particularly during critical stages such as preimplantation development, is of utmost importance. We hypothesized that, like genes, TEs are regulated by sequence-specific transcription factors (TFs) and are under evolutionary pressure to preserve TF binding sites (TFBS) for factors expressed during these developmental stages. In this study, we aimed to identify novel TFs involved in regulating six highly expressed TE families in the mouse embryo. By analyzing chromatin accessibility data, we identified 54 candidate TFs potentially involved in TE regulation. Phylogenetic analysis, combined with TFBS profiling over individual insertions, revealed correlations between TE subfamilies, TFBS and expression patterns during development. We reconstructed the evolutionary history of MERVL LTRs, identifying SRF as a regulator of MT2, and pinpointed the acquisition of SRF and DUX binding sites during MT2 evolution. A targeted gain-of-function screen in mESCs showed that 10 TFs induced TE transcription, and luciferase reporter assays confirmed the role of SRF as an activator of MT2. Three TFs, SRF, FOXJ3 and TBP, were further investigated for their functional roles in TE regulation *in-vivo*. Loss of function (LOF) experiments confirmed the role of SRF as a regulator of MERVL at the 2-cell stage, while TBP was found to regulate MaLR ORR. Together, this work sheds light on TE regulation during early mammalian development, and identifed SRF and TBP as novel regulators of TEs, expanding our understanding of the dynamic interplay between TFs and TEs in early embryogenesis.

2 Introduction

To contextualize the questions addressed during my PhD, I will first introduce transposable elements (TEs), with an emphasis on the structure of cis-regulatory regions of major mouse TE families. To situate the relevant biological framework, I will then outline preimplantation development, focusing on key events following fertilization such as zygotic genome activation and epigenetic reprogramming. Finally, I will present how TE expression during preimplantation development is not merely a spurious wave of transcription, but rather a regulated mechanism, likely governed by transcription factors (TFs). These TFs are recruited to TE cis-regulatory regions, and the resulting TE transcription comes along, in some cases, with the activation of genes specifically active during the time of zygotic genome activation. Hence, through the binding of TFs, from parasites, TEs become co-opts, whose transcriptional activation can be beneficial to the host. During my PhD I set out to understand how TEs are regulated, looking for new TFs involved in TE regulation.

2.1 Transposable elements

Transposable elements (TEs) are DNA sequences that have, in principle, the ability to move within the genome. Roughly 46% and 37% of the human and mouse genomes, respectively, have been identified as remnants of TE insertions, meaning that a substantial fraction of mammalian genomes are composed of TE fossils (Lander et al. 2001; Waterston et al. 2002). The true fraction of mammalian genomes derived from TEs may be underestimated, given significant decay that older insertions have undergone (de Koning et al. 2011). Regardless, TEs have proven to be remarkably successful at colonizing mammalian genomes.

2.1.1 Classification and transposition mechanisms of TEs

Based on the transposition mechanism and intermediate used, TEs are divided into two major classes: Class I TEs, or Retrotransposons, and Class II TEs, or DNA transposons (Finnegan 1989; Wicker et al. 2007). Class I TEs spread throughout the genome via a copy-and-paste mechanism that relies upon the reverse transcription of a full-length RNA intermediate (**Figure 1A**), Class II TEs directly cut-and-paste themselves in and out the genome (**Figure 1B**) (Wicker et al. 2007).

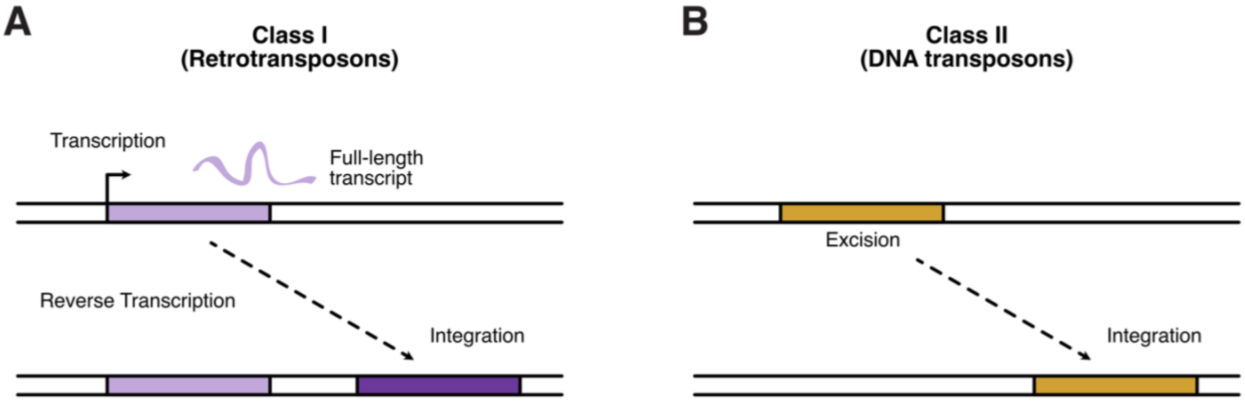


Figure 1. Schematic representation of the two major classes of TEs and their transposition mechanism. *(A)* Transposition mechanism of Class I transposons (Retrotransposons), which require transcription of a full-length transcript, followed by reverse transcription. *(B)* Transposition mechanism of Class II transposons (DNA transposons), involving direct excision of the DNA fragment and integration elsewhere in the genome. Figure adapted from (Jansz 2019), licensed under CC BY 4.0.

Whether they belong to Class I or Class II, TEs have the ability to transpose independently of host genome DNA replication. Autonomous TEs encode their own machinery for replication. Others, known as non-autonomous TEs, rely on the transposition machinery of other TEs to mobilize.

Perhaps because the copy-and-paste replication mechanism facilitates waves of copy number increase (Gifford et al. 2013), retrotransposons largely dominate the mammalian TE repertoires (Lander et al. 2001; Waterston et al. 2002; Rodriguez-Terrones and Torres-Padilla 2018; Osmanski et al. 2023). Based on structural components, retrotransposons are further divided into two major groups: the long terminal repeats (LTR)-containing elements and the elements which do not contain an LTR. LTRs are sequences that range from 100bp to over 5kb in size and flank the element on each end (Mager and Stoye 2015). While not exclusively, LTR-containing elements are primarily composed of Endogenous Retroviruses (ERVs). Non-LTR retrotransposons include two major subclasses: Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs) (Goodier and Kazazian 2008). Within these subclasses, TEs are categorized in superfamilies, themselves further divided into families.

TEs must be transcribed to mobilize. Retrotransposon transposition relies upon a full-length RNA intermediate. While DNA transposons do not need this RNA intermediate, they do need to transcribe and translate a functioning transposase, necessary for transposition. While replication itself is self-encoded and independent from that of host DNA, TEs depend on the recruitment of the host transcription machinery. To do so, TEs have evolved *cis*-regulatory sequences, of different forms and structures that effectively mimic host promoters by recruiting host-encoded factors such as RNA polymerases or transcription factors (TFs).

2.1.2 Structure of major murine retrotransposons

2.1.2.1 Endogenous Retroviruses (ERVs)

As their name indicates, ERVs are remnants of ancient retroviral infection which, when integrated into the germline, were subsequently vertically transmitted, and eventually endogenized (Eickbush and Malik 2007). ERVs exist in a wide variety of different structural configurations within the mouse genome, most of which are incomplete or truncated, and constitute about 11.5% of the mouse genome (estimates from mm10 and RepeatMasker). A complete ERV is structurally similar to a provirus, as it contains an internal coding sequence for the viral proteins Gag, Pol and occasionally Env, surrounded by two initially identical LTRs (**Figure 2**). A significant fraction are also present as solo LTRs, which result from homologous recombination between two LTRs of a complete element. ERVs have been classified according to phylogenetic analysis of the reverse transcriptase (RT) encoded by the Pol gene. Three major superfamilies result from this classification system: ERVI, ERVII and ERVIII which are related to gamma/epsilon, alpha/beta and spuma viruses, respectively (Rowe and Trono 2011; Mager and Stoye 2015). This work mainly focuses on ERVIII retrotransposons, which have lost the env gene (**Figure 2**). Mammalian apparent LTR Retrotransposons (MaLRs), which is a large category of elements (~4.8% of the mouse genome according to the initial mouse genome sequencing consortium (Waterston et al. 2002)), are internally depleted: their internal sequence is short and non-coding (**Figure 2**). Based on slight homology with the Gag gene, they are commonly classified as belonging to ERVIII. MaLRs are non-autonomous TEs and rely upon the replication machinery of ERVIII to propagate.

LTRs emerge from the reverse transcription process and contain the *cis*-regulatory elements required for activating and regulating transcription. An LTR can be divided into three regions: U3, R and U5 (**Figure 2**). U3 is the fragment at the 5' end of the LTR and corresponds to the sequence unique to the 3' end of the RNA. Conversely, U5 is the 3' most region of the LTR and is exclusively found at the 5' end of the RNA (**Figure 2**). U3 and U5 are duplicated upon reverse transcription. Finally, the R region, located in between U3 and U5 within the LTR, is repeated at both sides of the RNA (Mager and Stoye 2015; Vogt 1997) (**Figure 2**). The transcription of a full-length RNA is therefore expected to begin at the U3/R boundary and end at the R/U5 boundary (Vogt 1997; Boeke and Corces 1989) (**Figure 2**). Most of the control elements of a provirus, such as the RNA polymerase II (RNAPII) promoter and several enhancer sequences responding host TFs are found within U3 regions (Vogt 1997; Mager and Stoye 2015; Chuong et al. 2017).

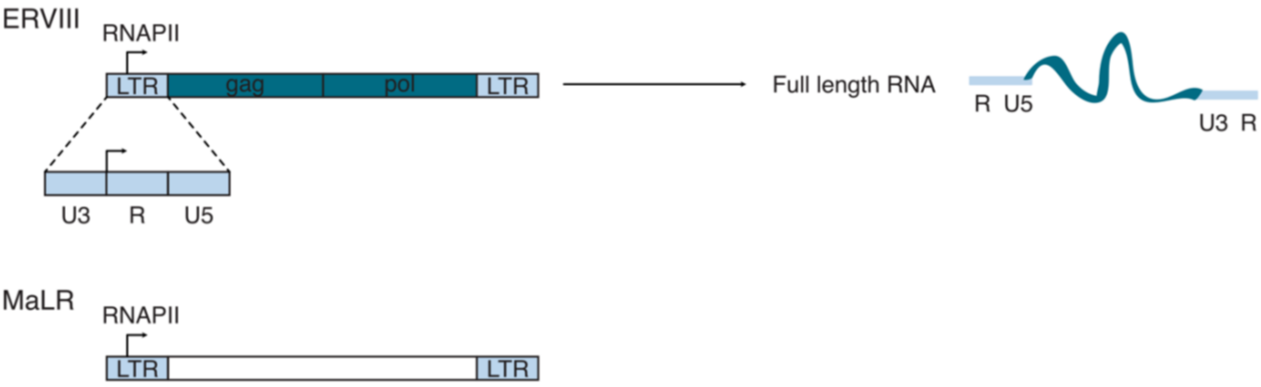


Figure 2. Schematic representation of ERVIII, MaLR and their regulatory elements. Boxes represent the different elements as indicated. Arrows represent the transcription start sites (TSSs). The full-length RNA structure is displayed, with highlighted the positions of the R, U3 and U5 regions of the LTR. (LTR) long terminal repeat, (ERV) endogenous retrovirus, (MaLR) mammalian apparent LTR retrotransposon. Figure adapted from (Hermant and Torres-Padilla 2021), licensed under CC BY-NC 4.0.

2.1.2.2 Long Interspersed Nuclear Elements (LINEs)

LINEs can be divided into many different clades (Malik et al. 1999; Bao et al. 2015; Rodriguez-Terrones and Torres-Padilla 2018). Three of these are widely distributed across mammals and not restricted to certain taxonomic groups: LINE1 (L1), LINE2 (L2) and CR1 (Rodriguez-Terrones and Torres-Padilla 2018). CR1 and L2 are older, and expanded earlier during mammalian evolution (Chalopin et al. 2015), but all three are found in all mammals indicating that their entry within the genome predates mammalian radiation which happened around 100 million years ago (MYA). In nearly all mammals, L1 consistently outnumbers L2 genomic content (Rodriguez-Terrones and Torres-Padilla 2018). In fact, L1 is the most successful retrotransposon in many mammals, and accounts for nearly 20% of the mouse genome (~19.9% according to mm10 and RepeatMasker annotations). Full-length L1 elements are composed of two open reading frames, Orf1 and Orf2, which encode proteins involved in transposition (**Figure 3**). A 5'UTR and a 3'UTR surround the coding regions (**Figure 3**) and many, if not most, copies in the mouse genome are truncated at the 5' end (Voliva et al. 1983).

In contrast to ERVs, L1s do not harbor an LTR sequence. A typical RNAPII promoter initiates transcription downstream of its promoter region. LTRs facilitate ERV transcription and reverse transcription without loss of promoter elements. Murine L1s have evolved a different, bipartite promoter system with a series of tandem repeats, called the monomers, which are associated to the coding regions through a structure termed the tether (Padgett et al. 1988; Naas et al. 1998) (**Figure 3**). Monomers are around 200bp in length, and repeated on average three times in young L1 families (Zhou and Smith 2019). Each monomer of the tandem repeat is sufficient to drive the expression of a reporter (Padgett et al. 1988; Naas et al. 1998; Furano 2000). Therefore, in the case of monomer loss, transcription can, in principle, still be initiated. This monomeric system provides the mouse L1 with a platform for recruitment of transcription regulators and explains the persistent 5' truncations observed in the genome (Voliva et al. 1983; Loeb et al. 1986). However, some monomers may be recovered through unequal homologous recombination between misaligned monomers during meiosis, leading to non-reciprocal exchange events that duplicate deleted monomers from the homologous template (Loeb et al. 1986).

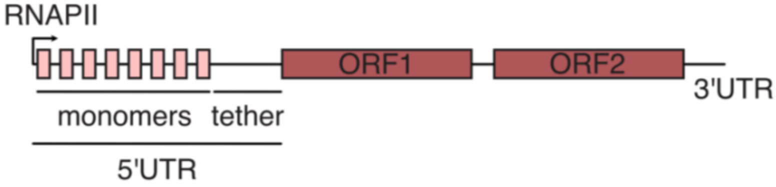


Figure 3. Schematic representation of murine L1 and its regulatory elements. Boxes represent the different elements as indicated. Arrow represents the TSS. (ORF) open reading frame, (UTR) untranslated region. Figure adapted from (Hermant and Torres-Padilla 2021), licensed under CC BY-NC 4.0.

2.1.2.3 Short interspersed elements

SINEs are non-autonomous retrotransposons, which do not encode for any protein and rely on the transposition machinery of LINEs to mobilize (Dewannieux et al. 2003; Dewannieux and Heidmann 2005). Full-length SINEs range between 100 to 600bp in size (Kramerov and Vassetzky 2011), and display a composite structure including a head on the 5' end, a body and a tail on the 3' end (**Figure 4**). The origin of the SINE head is used to classify SINE families. SINE1 heads are derived from 7SL RNAs, while SINE2 heads come from tRNAs. Even though SINE1 (SINEB1 in rodents) are present in the mouse genome, the most prevalent SINE family is SINE2 (SINEB2) in mice. These rodent-specific elements, with oldest members found in both rat and mouse genomes, started expanding around 50 MYA (Schmidt et al. 2012) and make up about 2.4% of the mouse genome.

SINEB2 elements, like the tRNA sequences their 5' end originates from, are primarily transcribed by RNA Polymerase III (RNAPIII). Unlike RNAPII, RNAPIII can initiate transcription upstream of its promoter. An RNAPIII-driven transcription mechanism enables SINEB2 to be transcribed from the +1 nucleotide, ensuring that promoter elements are preserved throughout successive transposition cycles. SINEB2 elements are derived from type II RNAPIII promoters, which contain two internal motifs: the A and B boxes. These regulatory boxes are each about 11bp-long and are involved in recruiting RNAPIII to the promoter (Schramm and Hernandez 2002) (**Figure 4**).

SINEB2 elements also contain an internal RNAPII promoter that is located outside of the tRNA-derived regions and drives transcription in the opposite direction (Ferrigno et al. 2001; Allen et al.

2004) (**Figure 4**). Additionally, many SINEs are found embedded within the 3'UTR of genes and pervasively transcribed by RNAPII (Chen et al. 2009; Roy-Engel et al. 2005).

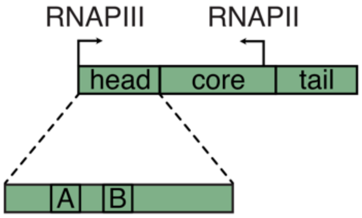


Figure 4. Schematic representation of SINEB2 regulatory elements. Boxes represent the different elements as indicated. Arrows represent the TSSs. A and B refer to the A and B boxes. Figure adapted from (Hermant and Torres-Padilla 2021), licensed under CC BY-NC 4.0.

2.1.3 Impact of transposable elements on mammalian genomes

While most TEs in the genome are no longer mobile, many still retain transcriptional potential. Over waves of transposition and genomic expansion, these *cis*-regulatory sequences have been dispersed throughout the genome, contributing to the pool of transcriptional regulators and causing regulatory innovations. First put forward as an insightful intuition by Barbara McClintock in the 1950s, following her discovery of TEs in maize (McClintock 1950; Mcclintock 1956), the impact of TEs in genome evolution is now undeniable. TEs have acted as powerful genetic elements, profoundly shaping the structure of genomes, whether influencing the proteins the genome encodes, or the regulatory mechanisms controlling the expression of these proteins (Feschotte 2008; Chuong et al. 2017; van de Lagemaat et al. 2003; Wang et al. 2007; Cohen et al. 2009; Bourque et al. 2008; Chuong et al. 2016; Modzelewski et al. 2021).

Indeed, this is now supported by a wealth of evidence, made possible by advances in genomics techniques, such as genome-wide mapping of promoters or TF binding locations. Using different methods, consecutive studies showed that at least one fourth of human promoters overlap with TE-derived sequences, one third of TSSs were located within TEs, and nearly 44% of open chromatin regions were located within TEs in different cell types and species (Jordan et al. 2003; Faulkner et al. 2009; Jacques et al. 2013). Further, many TFs were found to bind to TEs. Numerous TFs have been found to preferentially associate with TEs in a species- and cell type-specific manner across cancer cell lines, but also pluripotent cells such as stem cells. This suggests significant evolutionary pressures, on both TEs and the host genome, to maintain transcription factor binding sites (TFBS) over time (Wang et al. 2007; Bourque et al. 2008; Kunarso et al. 2010; Sundaram et al. 2014; Ito et al. 2017; Sun et al. 2018; Jiang and Upton 2019; Du et al. 2024). The selective binding of TFs to specific TEs further reinforces the idea that TEs originally carry the TFBS, scattering them across the genome during periods of genomic expansion (Bourque et al. 2008; Chuong et al. 2017; Schmidt et al. 2012; Hermant and Torres-Padilla 2021).

While the benefit of TEs in host evolution is indisputable, their uncontrolled spread could rapidly result in lethality. Indeed, given their mobile nature, TEs can compromise genome integrity. For example, a new insertion landing in an essential gene may be fatal for the cell. In fact, TE activity is associated with multiple diseases, particularly cancers (Hancks and Kazazian 2016; Kazazian and Moran 2017; Burns 2017). To avoid TE-induced genome damages, the cell has evolved a variety of mechanisms involved in restricting TE mobilization at various stages of the transposition cycle (reviewed in (Goodier 2016)). The host’s first line of defense against transposition is blocking transcription. In most somatic cells TEs are transcriptionally inert, with a few notable exceptions, such as neurons or cells in the placenta. TE transcriptional repression is mainly controlled by the two main components of constitutive heterochromatin: DNA methylation and H3K9me3. DNA methylation at CpG dinucleotides is initially established by de novo methyltransferases (DNMT3a, DNMT3b and DNMT3L) and subsequently maintained by DNMT1 (reviewed in (Greenberg and Bourc’his 2019)). H3K9me3 is catalyzed by a group of histone methyltransferases (HMTs), including SUV39H1, SUV39H2, SETDB1, SETDB2, G9a and EHMT1 (reviewed in (Padeken et al. 2022)). H3K9me3 is directed to TEs through the action of KRAB-zinc finger proteins (KZFPs), which possess a DNA-binding domain involved in sequence-specific recognition of TEs. KZFPs recruit the effector protein KAP/TRIM28, which in turn mobilizes the heterochromatin inducing machinery to the genomic location (Rosspopoff and Trono 2023).

The prevalence of TEs in mammalian genomes reflects a fine-tuned balance over evolutionary time between the disruptive potential of TE mobilization and the evolutionary benefits they may confer by enhancing host fitness and genetic diversity (Oomen and Torres-Padilla 2024).

2.2 Mouse preimplantation development

In mammals, embryonic development begins with the fertilization of the oocyte by the sperm (**Figure 5**), both of which are transcriptionally inert. During spermatogenesis, the sperm chromatin becomes highly condensed as histones are replaced by protamines. Meanwhile, growing mouse oocytes accumulate transcripts and proteins within the cytoplasm and upon maturation, oocytes remain arrested in metaphase II (MII), with condensed meiotic chromosomes. Fertilization merges these highly specialized cells, leading to the formation of a totipotent zygote (**Figure 5**). Totipotency refers to the ability of a single cell to give rise to a complete new organism by itself. Following fertilization, the zygote carries the parental genomes which remain physically segregated in pronuclei until they unite in a process called syngamy. This occurs during metaphase of the first mitosis and results in the formation of the first diploid nuclei in the 2-cell stage embryo. During the first few cell divisions, the embryo gradually transits from totipotency to pluripotency, ultimately forming the blastocyst, which implants into the female uterus (**Figure 5**). The formation of the blastocyst marks the allocation of the first lineages, separating extraembryonic from embryonic tissues. The inner cell mass (ICM), which forms the embryonic tissues, has lost the ability to generate supportive extraembryonic tissues, but can differentiate into all three germ layers of the embryo, including the germline, and is therefore referred to as pluripotent.

2.2.1 Zygotic genome activation

It is during preimplantation development that the new organism transcribes its genome for the first time in a process called zygotic genome activation (ZGA). Simultaneously, the pool of maternally loaded mRNAs is degraded, altogether enabling the maternal-to-zygotic transition (MZT) to occur (**Figure 5**). In the mouse, ZGA occurs in two waves: the minor wave starts at the end of the zygote stage (Mintz 1964), while the major wave happens at the late 2-cell stage and consists of the transcriptional activation of about 4000 genes simultaneously (Aoki et al. 1997; Hamatani et al. 2004; Jukam et al. 2017) (**Figure 5**). Concomitant with activation of gene expression, murine preimplantation development is characterized by a robust activation of TEs (Peaston et al. 2004; Fadloun et al. 2013), a phenomenon that is also observed in other mammalian species (Rodriguez-Terrones and Torres-Padilla 2018; Modzelewski et al. 2021; Oomen et al. 2025). TEs are transcriptionally active as early as the zygote stage for certain elements and sometimes until the blastocyst stage. For example, ERVLs (ERVIII) are robustly transcribed from the onset of minor ZGA and peak at the late 2-cell stage (Kigami et al. 2003; Peaston et al. 2004). Young L1s also get transcriptionally activated with a peak at the late 2-cell stage (Fadloun et al. 2013;

Jachowicz et al. 2017). SINEB2, on the other hand, become transcriptionally activated by the time of ZGA and peaks later in development around the morula to blastocyst stages (**Figure 5**) (Bachvarova 1988; Fadloun et al. 2013). Therefore, TE expression is dynamic, abundant, stage specific, and involves many, but not all, TE families (Fadloun et al. 2013; Oomen and Torres-Padilla 2024).

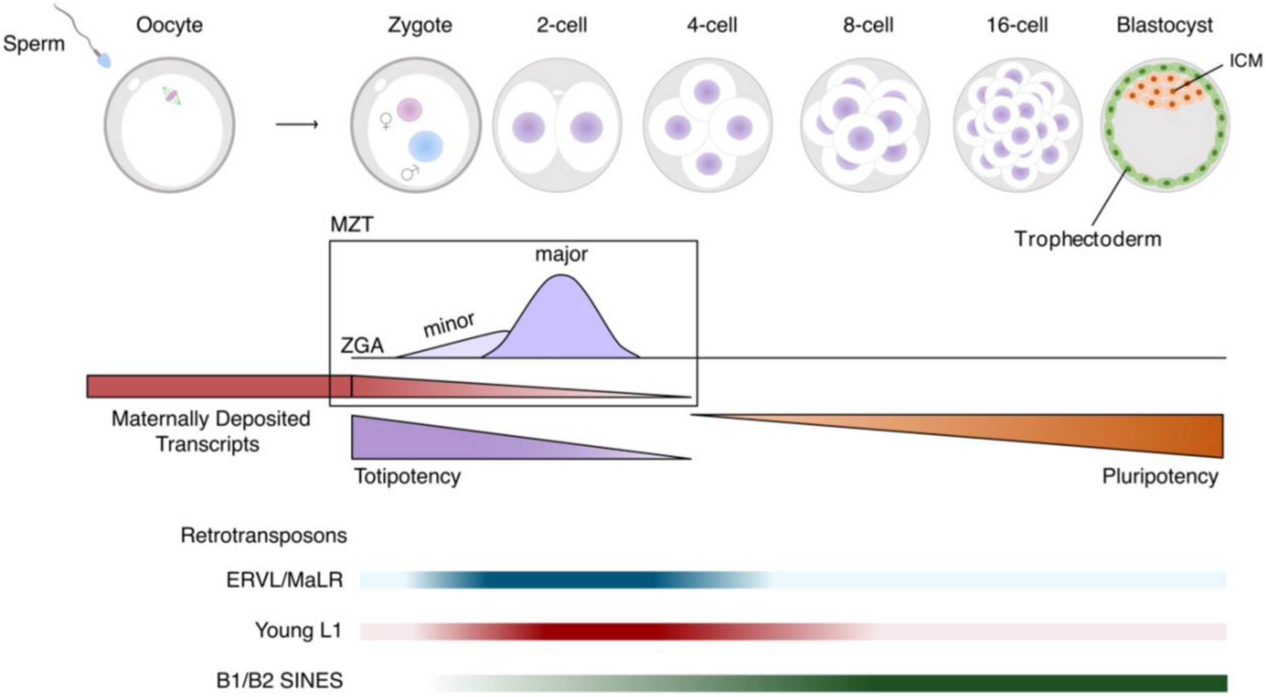


Figure 5. Murine preimplantation development overview. Sperm, oocyte and cleavage stage embryos up to blastocyst stage are depicted. The major biological processes occurring during this stage of development are highlighted as well as expression patterns of main retrotransposons. (ICM) inner cell mass, (MZT) maternal-to-zygotic transition, (ZGA) zygotic genome activation. Figure adapted from (Hermant and Torres-Padilla 2021), licensed under CC BY-NC 4.0.

This peculiar embryonic transcription characterized by expression of many repeated elements tightly silenced in most cell types raises the central question of my PhD project:

How is TE expression during preimplantation development regulated?

2.2.2 Epigenetic reprogramming: a "window of opportunity"?

TE de-repression is not unique to blastomeres during preimplantation development, it has also been observed in cancer cells, as well as in the germline. These cells have a critical feature in common: they all undergo extensive epigenetic reprogramming. Indeed, these disease or developmental contexts are characterized by globally open chromatin structure and high cellular plasticity, which may promote the robust transcriptional activation of TEs.

During preimplantation development and following fertilization, large-scale epigenetic changes and chromatin organization remodeling occur, which are believed to underlie the reprogramming of highly specialized gametes into totipotent cells and allow zygotic transcription to start (Burton and Torres-Padilla 2010, 2014; Eckersley-Maslin et al. 2018). At these developmental stages, the global chromatin structure is largely relaxed (Bošković et al. 2014; Wu et al. 2016). While the changes in the chromatin composition occur at different levels including DNA methylation, histone post-translational modifications and histone variants (Burton and Torres-Padilla 2014), I will focus on the two classical marks of heterochromatin involved in TE regulation; DNA methylation and H3K9me3.

First, the DNA becomes largely demethylated upon fertilization up to the blastocyst stage (Smith et al. 2014; Eckersley-Maslin et al. 2018). Interestingly, the dynamics of DNA demethylation differ between the two parental genomes (Smith et al. 2014, 2012; Wang et al. 2014b; Guo et al. 2014). While paternal DNA demethylation is active and rapid, maternal DNA demethylation is largely passive. Similarly, H3K9me3 is intensely remodeled. Levels of H3K9me3 fall drastically right after fertilization occurs, also asymmetrically between the parental genomes (Liu et al. 2004; Puschendorf et al. 2008; Wang et al. 2018). These atypical chromatin features at the early stages of development could constitute a “window of opportunity” for pervasive TE transcription, as all the classical chromatin modifications involved in TE repression in somatic cells are largely absent. However, the chromatin remodeling described above cannot, alone, explain the pattern of TE transcriptional activation. Despite an increase in H3K9me3, and stable levels of DNA methylation deposited over MERVL elements between the zygote and 2-cell stages, these elements are heavily upregulated upon ZGA at the 2-cell stage (Fadloun et al. 2013; Wang et al. 2018). Whereas LINEs, ERVL and MaLR are hypomethylated (DNA) after fertilization and highly expressed at the 2-cell stage, SINEs are hypomethylated upon fertilization, yet their highest transcriptional activity is observed at the morula and blastocyst stages (Bachvarova 1988; Fadloun et al. 2013; Smith et al. 2014; Wang et al. 2018). Furthermore, increased H3K9me3 from Suv39h1 overexpression and removal of heterochromatic marks prior to ZGA do not affect TE expression (Burton et al. 2020).

This suggests that the pattern of transcriptional activation of TEs during preimplantation development does not fully trace the chromatin dynamics at these elements, implying that their transcription is not simply due to the loss of classic heterochromatin signature, but instead modulated by the action of sequence-specific transcription factors.

2.3 TE expression in early embryos: from parasites to co-opts

Rather than a spurious wave of transcription, TE expression in the early embryo emerges as a tightly regulated process of key biological relevance to the host. The stage-specific TE transcription across classes, families (Rodriguez-Terrones and Torres-Padilla 2018; Oomen et al. 2025) and occasionally even subfamilies (Carter et al. 2022), led us to hypothesize the involvement of sequence-specific transcription factor-mediated activation and repression. Not all TE families are expressed at the same level; rather a specific subset is known to be transcribed. In addition, TE transcription at these stages is not merely a result of read-through transcription of TEs embedded within genes but is driven by the *cis*-regulatory elements of the TEs themselves (Evsikov et al. 2004; Peaston et al. 2004; Fadloun et al. 2013; Oomen et al. 2025). Understanding how TEs are regulated at this stage of development becomes a central question, at the intersection of their evolutionary pressures as parasitic elements, reliant upon vertical transmission, and the biological significance of their transcription for the host.

2.3.1 Evolutionary pressure for vertical transmission

When considering TEs as purely selfish elements, whose fitness is dependent on the ability of new insertions to be transmitted to the next generation, then a strong selection likely occurs for their mobilization prior to, or during germline development. In the mouse, primordial germ cells (PGCs) arise soon after implantation, around 7.5 days after fertilization. It is therefore an ideal timeframe for TEs to transpose, as cells during preimplantation development will later contribute to the germline formation. A new insertion within blastomeres of a preimplantation developing embryo can be transmitted to the germline and therefore, transmitted to the next generation. TEs likely evolved under the selective pressure to be transcribed during preimplantation development. This can be fulfilled by recruiting TFs that are specifically expressed at this stage, driving their expression in a timely manner.

Some examples, mostly regarding human ERVs started to emerge supporting this hypothesis. For example, it was reported that LTR5Hs, a human-specific ERVK (ERVII), is transcriptionally activated in human embryos from the 8-cell stage, the time at which human embryonic genome activation happens. Conversely, LTR5a and LTR5b, the evolutionary predecessors of LTR5Hs found in non-human primates, are not expressed during preimplantation development. Despite overall 90% identity between the three consensus sequences, only LTR5Hs harbored a motif for the TF OCT4 (Grow et al. 2015). In pluripotent cells, OCT4 was found to bind exclusively to LTR5Hs, and siRNA-mediated knock-down of OCT4 reduced HERVK transcription (Grow et al.

2015). This suggested that OCT4 regulates HERVK, and is responsible, through the evolution of a binding site within the youngest LTR, of its expression in pluripotent cells.

Similarly, HERVH (ERVI) is known to be highly expressed in pluripotent cells, specifically in cells of the ICM, in naïve human embryonic stem cells (hESCs) as well as during reprogramming of somatic cells to induced pluripotent stem cells (iPSCs). Many different studies have reported a variety of TFs to bind and activate HERVH, including OCT4, NANOG, SP1 and SOX2 (Kunarso et al. 2010; Santoni et al. 2012; Kelley and Rinn 2012; Wang et al. 2014a; Ohnuki et al. 2014; Fort et al. 2014; Göke et al. 2015; Ito et al. 2017; Zhang et al. 2019b; Pontis et al. 2019; Carter et al. 2022). The HERVH LTRs are subdivided into four families: LTR7, LTR7b, LTR7c and LTR7y. While only some LTR7 insertions were found to be expressed in pluripotent stem cells, 7c and 7y families were characteristically expressed earlier in development, by the time of human genome activation (Göke et al. 2015; Carter et al. 2022). Using phylogenetic analysis of all LTR7 insertions, Carter and colleagues were able to partly explain these differences in temporal expression patterns. Indeed, their phylogenetic analysis of all HERVL LTRs led to the identification of eight previously unknown subfamilies based on sequence similarity. By deriving consensus sequences for the subfamilies, they identified an 8bp insertion in the most recent ones, specifically expressed in pluripotent cells. They could associate this insertion with the introduction of a SOX2/3 binding motif, which conferred specific expression in pluripotent cells to these insertions (Carter et al. 2022). This suggests that the evolution of LTR7 was influenced by the acquisition of TFBS, enabling the colonization of distinct developmental environments for expression and genome amplification.

2.3.2 TE cooption during mouse preimplantation development

2.3.2.1 A murine ERV as a regulator of ZGA

The mouse-specific ERV with leucine tRNA primer binding site called MERVL (ERVIII) is transiently transcribed at the two-cell stage (Kigami et al. 2003; Peaston et al. 2004; Evsikov et al. 2004; Svoboda et al. 2004). ERVLs are found in all placental mammals, suggesting the existence of a common ancestor more than 70 MYA (Bénit et al. 1999), but underwent species-specific expansions. It was suggested, based on studies on a few insertions, that MERVL experienced two waves of expansion within the mouse genome. A first one soon after the divergence between *Mus* and *Rattus* (~14 MYA), and a second around 2 MYA (Bénit et al. 1999; Costas 2003). All internal regions are annotated using the consensus MERVL-int, and the most recent LTR, the only family still associated with complete or fragmented MERVL-int, is MT2_Mm. The ancestors of MT2_Mm are MT2C_Mm, MT2B and MT2A, among which MT2B and MT2A are older, as they are also found in the rat genome (Franke et al. 2017).

At the onset of the 21st century, MERVL was found to have the potential to initiate host gene transcription in mouse embryos (Peaston et al. 2004). Several years later, mouse embryonic stem cells (mESCs) were found to spontaneously fluctuate to a “2-cell like cell” state (2CLC) (Macfarlan et al. 2012). This was initially identified based on the transcriptional activation of MERVL. Several features have been described to be similar between 2CLCs and the 2-cell stage embryo (reviewed in (Genet and Torres-Padilla 2020)). At the transcriptome level, MERVL expression in 2CLCs comes along with the activation of “2-cell specific” genes, largely corresponding to ZGA, suggesting a direct role for MERVL in changing embryonic stem cells fate. (Macfarlan et al. 2012). The discovery of 2CLCs provided the community with a tool to explore the molecular features of totipotency (Rodriguez-Terrones et al. 2018), and also positioned an ERV as a putative key driver in the change of fate and reprogramming to totipotency (Torres-Padilla 2020). Recently, by a combination of knock-downs of MERVL using antisense oligonucleotides (ASO) and CRISPRi-based repression in the embryo, it was shown that MERVL transcription is indeed important for development to proceed (Sakashita et al. 2023). Although embryos were not blocked at the 2-cell stage upon ASO-mediated MERVL knock-down, they failed to develop to blastocyst (Sakashita et al. 2023). About 1000 genes were differentially expressed at the 2-cell stage, during major ZGA, suggesting a direct role for MERVL transcription in regulating these genes (Sakashita et al. 2023). Similarly, using CRISPRi-mediated MT2_Mm repression in zygotes, about 1500 genes were differentially regulated at the 2-cell stage, most of which were downregulated (Yang et al. 2024). Transcription of MT2_Mm and MERVL is therefore functionally important during preimplantation development and for proper progression through MZT to occur.

Some TFs regulating MT2_Mm and MERVL are known. For example, GATA2 was found to regulate MERVL in mESCs (Choi et al. 2017). Transcriptional activation by GATA2 was specific to MT2_Mm, as the evolutionary older MT2A and MT2B1/2 were not affected by GATA2 increase in expression (Choi et al. 2017). In addition, MT2_Mm includes a binding site for the TF DUX (*Duxf3*), which is a double homeodomain TF. Human DUX4, the DUX human ortholog, functions as a pioneer TF, recruiting histone acetyltransferases (HATs) p300/CBP mediating H3K27Ac deposition and allowing for transcription activation (Choi et al. 2016). The pioneering activity is conserved in the mouse DUX but the DNA binding homeodomains display low (33%) sequence homology between mouse DUX and human DUX4 (Eidahl et al. 2016). DUX4 was initially identified as aberrantly expressed in facioscapulohumeral muscular dystrophy (FSHD) (Geng et al. 2012; Young et al. 2013). Overexpression of either DUX in mESCs or DUX4 in human iPSCs induces robust expression of their respective species-specific ERVL: MERVL and HERVL (De Iaco et al. 2017; Hendrickson et al. 2017; Whiddon et al. 2017). The regulatory sequences of

MERVL and HERVL have diverged significantly through species-specific amplification, suggesting that DUX and DUX4 have evolved independently within their species to perform analogous functions. In mESCs, transcriptional activation of MERVL upon ectopic DUX expression is concurrent with the activation of a substantial part of the two-cell transcriptional program (De Iaco et al. 2017). However, the overlap of genes regulated by DUX in mESCs and the two-cell genes activated during ZGA is incomplete. In fact, *Dux*-/- mice are born, albeit at submendelian ratios, indicating that DUX is not essential for development (Guo et al. 2019; Chen and Zhang 2019; De Iaco et al. 2020). Although DUX is a key regulator of ZGA, other TFs must exist and play redundant roles in development. When, and how, the DUX binding site appeared within MT2_Mm, and how this affected MERVL genomic expansion is unknown. The oocyte-specific homeobox genes (*Obox*), which constitute a cluster of more than 60 genes all located on chromosome 7, were also implicated in regulating ZGA and MERVL (Royall et al. 2018; Ji et al. 2023; Guo et al. 2024). The OBOX gene cluster is specific to rodents but occurs in the syntenic region where human TPRX1, TPRX2 and TPRXL are located (Wilming et al. 2015), which have also been involved in regulating human ZGA (Royall et al. 2018; Zou et al. 2022). To overcome potential redundancies between *Obox* genes when studying knock-out models, 1.2 Mb within the cluster were deleted including the *Obox1/2/3/4/5/7* (which contain both maternal and zygotic *Obox* genes). *Obox1/2/3/4/5/7*-null embryos arrested at the 2-cell stage. ZGA was impaired and MERVL/MT2_Mm expression downregulated but not entirely abolished, again suggesting the existence of additional players. While our understanding of the molecular players involved in MERVL regulation is expanding, it is clear that other factors are also involved in its transcriptional regulation. Further, the evolutionary path that led to its co-option as a key regulator of ZGA remains largely unknown.

2.3.2.2 Other major retrotransposons families

Expression of TEs during preimplantation development includes other elements in addition to MERVL, encompassing transcription from various other TE families. This includes other ERVs, such as the MaLRs. MaLRs can be divided into different families, dominated in rodents by the ORR1 and MT families. MaLR MT can be further divided into six main families, among which MTA_Mm is the youngest and is exclusively found in mice (Franke et al. 2017). MT MaLRs are major components of the oocyte transcriptome, and have been coopted as oocyte-specific promoters in oocyte reprogramming (Peaston et al. 2004; Veselovska et al. 2015; Franke et al. 2017; Modzelewski et al. 2021). The ORR1 MaLRs are also classified into several families and include the youngest member, ORR1A. Even though ORR1A is the youngest ORR1 MaLR, ORR1A is not restricted to *Mus* genomes and therefore is older than both MT2_Mm and

MTA_Mm. ORR1 MaLRs are highly expressed at the 2-cell stage, in a way that resembles MERVL patterns more than their oocyte-specific MT MaLR counterparts (Peaston et al. 2004). The expression of ORR1 during preimplantation development was described more than two decades ago, and their role as alternative promoters has been documented in several studies for both MT and ORR1 (Peaston et al. 2004; Veselovska et al. 2015; Franke et al. 2017; Modzelewski et al. 2021; Yang et al. 2024; Honda et al. 2024; Oomen et al. 2025). However, the regulatory mechanisms underlying their expression are largely unknown. The distinctive expression patterns of MT MaLRs and ORR1 MaLRs in different developmental contexts suggests the involvement of sequence-specific TFs.

Beyond ERVs, LINEs and SINEs are also transcribed during development. Young L1 families (Af, Gf and Tf) are transcribed as early as the zygote stage in mouse (Fadloun et al. 2013; Jachowicz et al. 2017), and appear to regulate global chromatin accessibility during development (Jachowicz et al. 2017). Using a sequence-specific targeting approach, timely activation of L1 was shown to be necessary for development (Jachowicz et al. 2017). The regulators of L1 transcription remain mostly unidentified. The ubiquitous TF YY1 contains a binding site within the monomers of two young L1 families, Tf and Gf (DeBerardinis and Kazazian 1999). Very recently, using siRNA-mediated depletion in embryos, YY1 was shown to be required for L1MdT transcription in morulae (Sakamoto and Ishiuchi 2024). SINE transcripts are present in oocytes, inherited in zygotes but their transcription increases upon ZGA and peaks by the morula stage (Taylor and Pikó 1987; Bachvarova 1988; Fadloun et al. 2013). SINEB1s are bound and regulated by NR5A2 during early embryo development, potentially regulating mouse ZGA (Gassler et al. 2022). Whether NR5A2 is indeed involved in regulating ZGA is still debated (Lai et al. 2023; Festuccia et al. 2024). A role for SINEs, particularly SINEB2, in the regulation of chromatin organization has been documented (Schmidt et al. 2012). CTCF was associated with SINEB2 elements in mESCs, accounting for 33.8% of the CTCF binding regions (Bourque et al. 2008; Kunarso et al. 2010; Schmidt et al. 2012). CTCF, also known as 11-zinc finger protein or CCCTC-binding factor, is a TF that acts as an architectural protein that promotes loop extrusion by acting as a boundary, leading to the formation of topologically associating domains (TADs) (Dixon et al. 2012; Nora et al. 2012; Sexton et al. 2012). While no role has been ascribed to SINEB2 transcription during preimplantation development, these data collectively suggest that SINEB2 may play a key role in higher order chromatin organization, and that expansion of these elements has contributed to species-specific expansion of CTCF binding sites across the genome. The mechanisms governing SINEB2 transcription during preimplantation development nonetheless remain entirely unexplored.

3 Aims

We hypothesized that transcription of TEs during preimplantation development is modulated by the action of sequence-specific TFs. The primary goal of my PhD project was to identify novel TFs involved in modulating TE expression during development. This objective was pursued through six specific aims:

In-silico:

- Conduct a footprinting analysis to identify TFs potentially activating specific TE families.

- Perform a phylogenetic analysis of the different TE families to gain deeper insight into the sequence determinants of their regulation.

In-cellulo:

- Carry out a targeted gain-of-function screen to assess the role of TFs in TE activation in mESCs.

- Use a heterologous system based on luciferase reporter to evaluate the ability of TFs to activate TEs.

In-vivo:

- Describe the expression patterns of selected TFs at both mRNA and protein levels in embryos.

- Investigate the role of TFs *in-vivo*, determining their function in TE regulation in the embryo. If possible, define the genomic locations of TF binding within the embryo.

4 Results

4.1 Identification of candidate TFs regulating TEs during preimplantation development

4.1.1 Selection of TE families of interest

Aiming to extend the repertoire of TFs that regulate TEs during preimplantation development, I chose to focus on major families of young, rodent-specific TEs. I first included the well-characterized MERVL LTR, MT2_Mm. While several TFs have already been identified as regulators of MERVL both *in-vitro* and *in-vivo* (Hendrickson et al. 2017; De Iaco et al. 2017; Whiddon et al. 2017; Choi et al. 2017; Zhang et al. 2019a; Ji et al. 2023; Guo et al. 2024), it was suggested that there may be additional factors involved in MERVL regulation. Using publicly available high resolution single embryo RNA-seq dataset across the different stages of preimplantation development, I could confirm the precise expression pattern of MT2_Mm (**Figure 6**). Importantly, this dataset was generated using a method called Smart-Seq+5', which is a variation of Smart-Seq2, enabling the capture of the 5' end of the transcript, representing true TE-driven transcription (Oomen et al. 2025). As described previously (Peaston et al. 2004; Svoboda et al. 2004), MT2_Mm was transcriptionally activated at the same time as the major wave of ZGA, with an induction at the early 2-cell stage, a peak at the late 2-cell stage and silencing by the 8-cell stage (**Figure 6**).

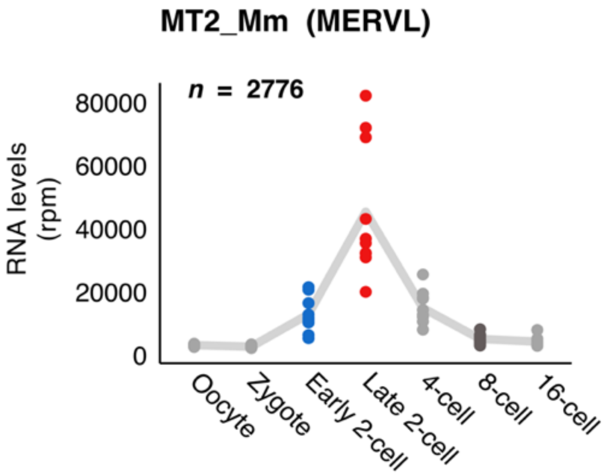


Figure 6. Expression pattern of MT2_Mm across preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot represents the sum rpm of all insertions (*n*) belonging to MT2_Mm family per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages.

Then, I proceeded to explore additional LTR-containing elements, with a particular focus on MaLRs. Using the same dataset as described above, I looked at the expression dynamics of different MaLR families across preimplantation development. The expression of MTA_Mm and

MTB_Mm was characteristic of oocyte transcription, with high transcript levels found in the oocyte, decreasing after fertilization (**Figure 7A, B**), as previously reported (Peaston et al. 2004; Franke et al. 2017; Modzelewski et al. 2021). MTA_Mm expression peaked at the early 2-cell stage (**Figure 7A**), corresponding to the timing of minor ZGA, but decreased substantially by the late 2-cell stage. ORR MaLRs though, such as ORR1A0 and ORR1A1, displayed transcription kinetics reminiscent to that of MT2_Mm (**Figure 7C, D**), characterized by high expression at the late 2-cell stage. In fact, ORR1A0 and ORR1A1 are among the highest transcribed LTR-containing TEs at the late 2-cell stage, after MT2_Mm. Therefore, I chose to focus on these latter two families while excluding the oocyte-specific families.

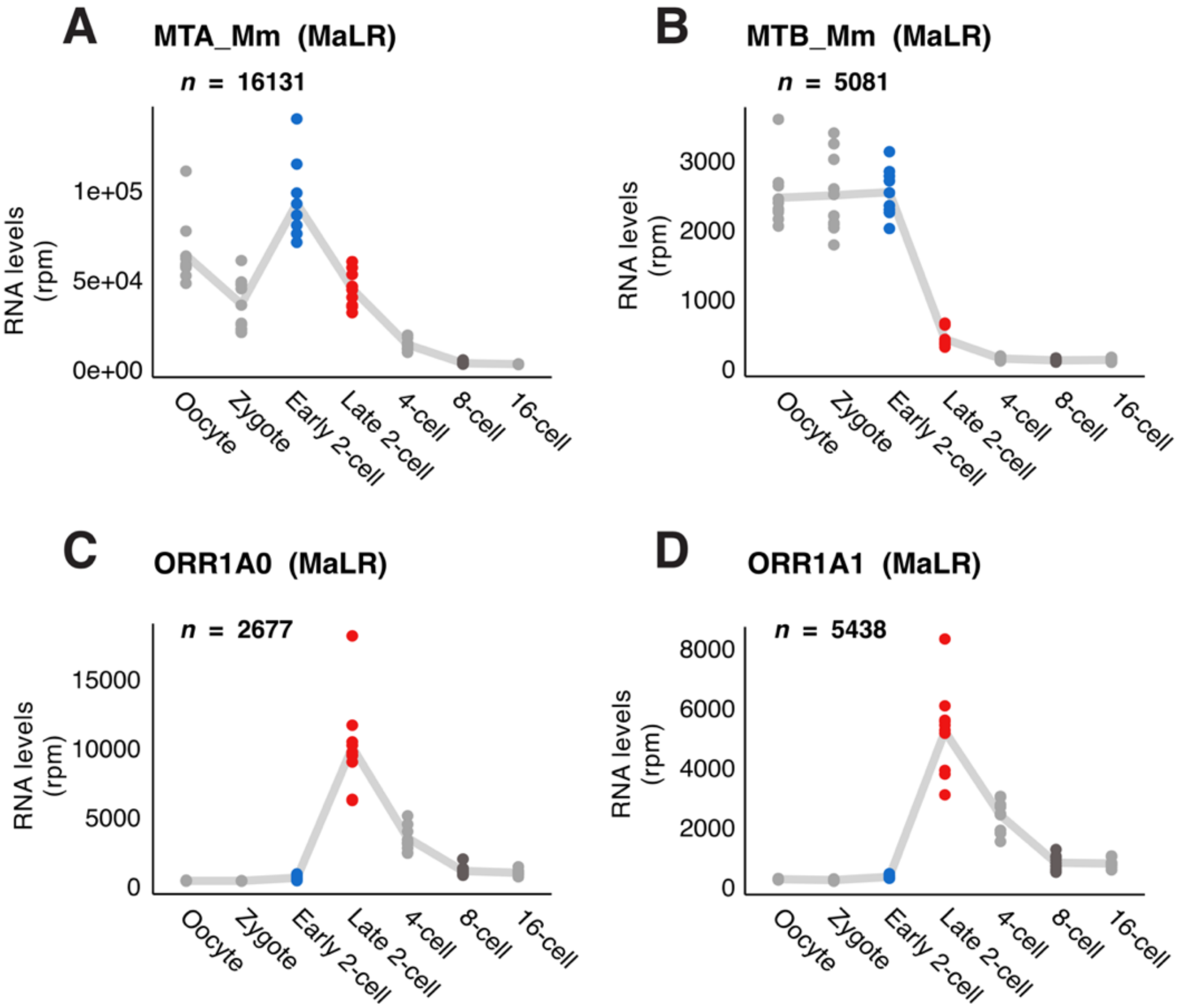


Figure 7. Expression patterns of major, recent families of MaLRs: MT MaLRs, MTA_Mm (A) and MTB_Mm (B) and ORR MaLRs, ORR1A0 (C) and ORR1A1 (D) across preimplantation development. (A-D) Data was reanalyzed from (Oomen et al. 2025). Each dot represents the sum rpm of all insertions (*n*) belonging to the indicated TE family per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages.

Additionally, I included key families of non-LTR TEs. Notably, L1, and more specifically L1MdTf, one of the youngest mouse L1 element, stands out for its expression during preimplantation development (Fadloun et al. 2013; Jachowicz et al. 2017). Because the L1 regulatory elements are contained within its monomers, I examined the expression of the monomers themselves, which I re-annotated using the L1MdTf_II monomer RepBase consensus sequence. As previously described for complete L1MdTf_II elements (Fadloun et al. 2013; Jachowicz et al. 2017), L1MdTf_II monomers transcription was initiated upon major ZGA at the 2-cell stage, and RNA levels increased through development, with a peak at the 8- to 16-cell stages (**Figure 8**).

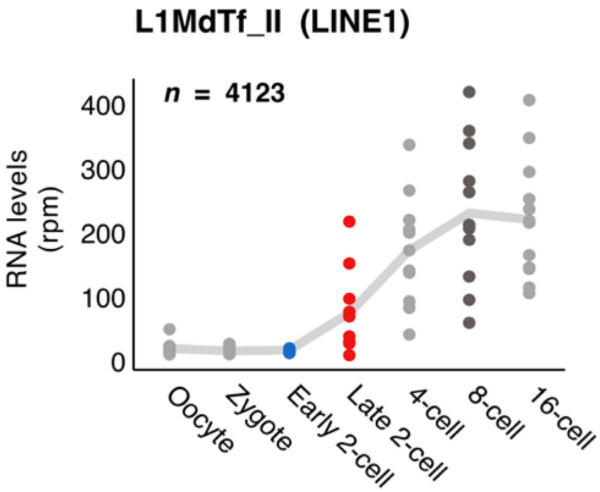


Figure 8. Expression pattern of L1MdTf_II monomers across preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot represents the sum rpm of all insertions (*n*) belonging to L1MdTf_II family (only monomers) per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages.

Finally, I analyzed the expression profiles of SINEB2 across development. The main families of SINEB2 are the following: B3A, B3, B2_Mm2, B2_Mm1t and B2_Mm1a.

B3A and B3 are phylogenetically closely related, and upon examining their expression using the same dataset (Oomen et al. 2025), I observed that they exhibit highly similar expression profiles (**Figure 9A, B**). B3A and B3 correspond to the oldest SINEB2 families, with their expansion estimated to have begun around 50 MYA, although B3A is slightly older than B3 (Schmidt et al. 2012). B2_Mm1a and B2_Mm1t are the youngest, *Mus*-specific SINEB2 families (Schmidt et al. 2012). Similarly to B3A and B3, B2_Mm1a and B2_Mm1t are phylogenetically close to each other and their expression profiles are similar (**Figure 9C, D**) (Schmidt et al. 2012). I chose to focus on the youngest SINEB2, B2_Mm1a, as well as the oldest, B3A, since both exhibit high expression levels with transcriptional induction coinciding with major ZGA (**Figure 9B, D**).

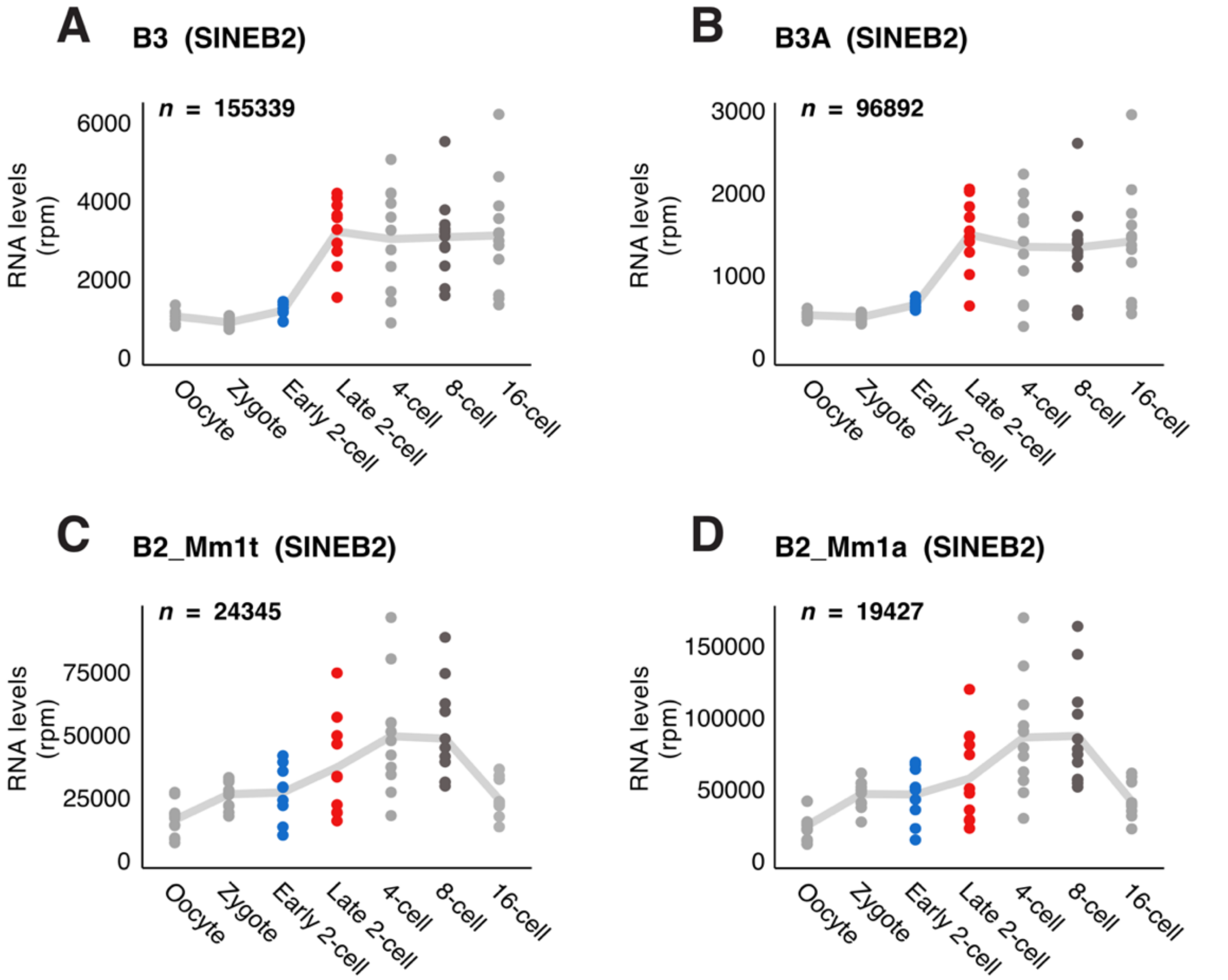


Figure 9. Expression patterns of B3 (A), B3A (B), B2_Mm1t (C) and B2_Mm1a (D) families across preimplantation development. (A-D) Data was reanalyzed from (Oomen et al. 2025). Each dot represents the sum rpm of all insertions (*n*) belonging to the indicated TE family per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages.

In conclusion, I decided to focus on six TE families, all expressed around the time of major ZGA, for further analysis: **MT2_Mm, ORR1A0, ORR1A1, L1MdTf_II, B2_Mm1a and B3A.**

4.1.2 A footprinting analysis identifies 54 candidate TFs potentially involved in TE regulation

This part of my PhD was performed in collaboration with Dr. Natasha Jansz.

To identify potential candidate transcriptional regulators of these six families of TEs, we performed a transcription factor footprinting analysis using publicly available ATAC-seq data that have been generated in the mouse preimplantation embryo (Wu et al. 2016). ATAC-seq utilizes a Tn5 transposase to infer chromatin accessibility; the greater the signal from the Tn5 insertion, the greater the chromatin accessibility. The binding of proteins can protect the DNA from Tn5 insertion, allowing TFBS to be inferred from signatures in the ATAC-seq reads when mapping the insertion sites, providing single nucleotide resolution (**Figure 10**). For each family, we mapped the Tn5 insertion sites over their promoter: the LTR for MT2_Mm and ORR1A0/A1, the Tf monomers for L1MdTf_II and the complete elements for B2_Mm1a and B3A.

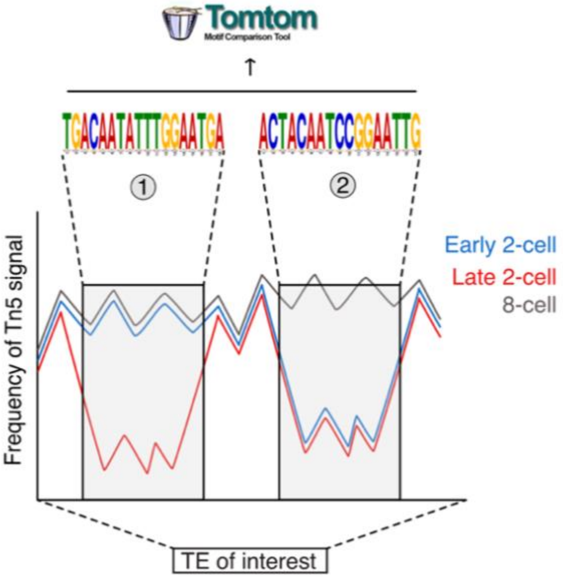


Figure 10. Schematic representation of the transcription factor footprinting analysis. The colored lines represent the aggregated Tn5 insertion signal at the different stages. A footprint corresponded to a local depletion compared to flanking regions and compared to the other stages. Footprints were identified by visual inspection. The sequence corresponding to the footprint position within the consensus sequence was extracted and subjected to motif search using Tomtom (MEME suite).

By manual inspection of the mapped Tn5 insertion frequencies, we identified signatures in the reads corresponding to local depletion in signal relative to the other stages and to the flanking regions. As we looked for transcriptional activators of TEs during ZGA, we focused on footprints that were specifically found at the early 2-cell stage, the late 2-cell stage, or both (**Figure 10**).

As DUX is a known regulator of MT2_Mm and its binding site within the LTR has been identified (Hendrickson et al. 2017), we decided to use MT2_Mm and DUX as proof-of-principle for this analysis. Indeed, we identified two footprints within MT2_Mm present at both early and late 2-cell stages, as well as two footprints present only at the early 2-cell stage. By examining the sequence corresponding to the footprint position and comparing it with the known binding site of DUX (Hendrickson et al. 2017), we could determine that the second early 2-cell specific footprint (highlighted by the inset in Figure 11) corresponded to DUX binding site (**Figure 11**).

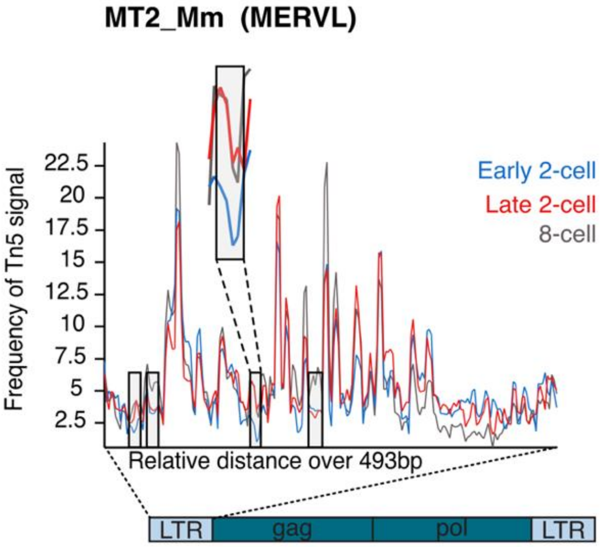


Figure 11. Tn5 insertion frequency plot over MT2_Mm. Early 2-cell insertion frequency is displayed in blue, late 2-cell in red and 8-cell in brown. The size of the MT2_Mm consensus sequence is indicated. The inset is a zoom on a specific footprint (DUX footprint).

Despite ~96% sequence similarity of ORR1A0 and ORR1A1 consensus sequences (Repbase), we identified different footprints within the Tn5 insertion frequency plots. We found three late 2-cell specific footprints within ORR1A0 LTRs (**Figure 12A**), while in ORR1A1 we identified three early 2-cell footprints and one late 2-cell footprint (**Figure 12B**).

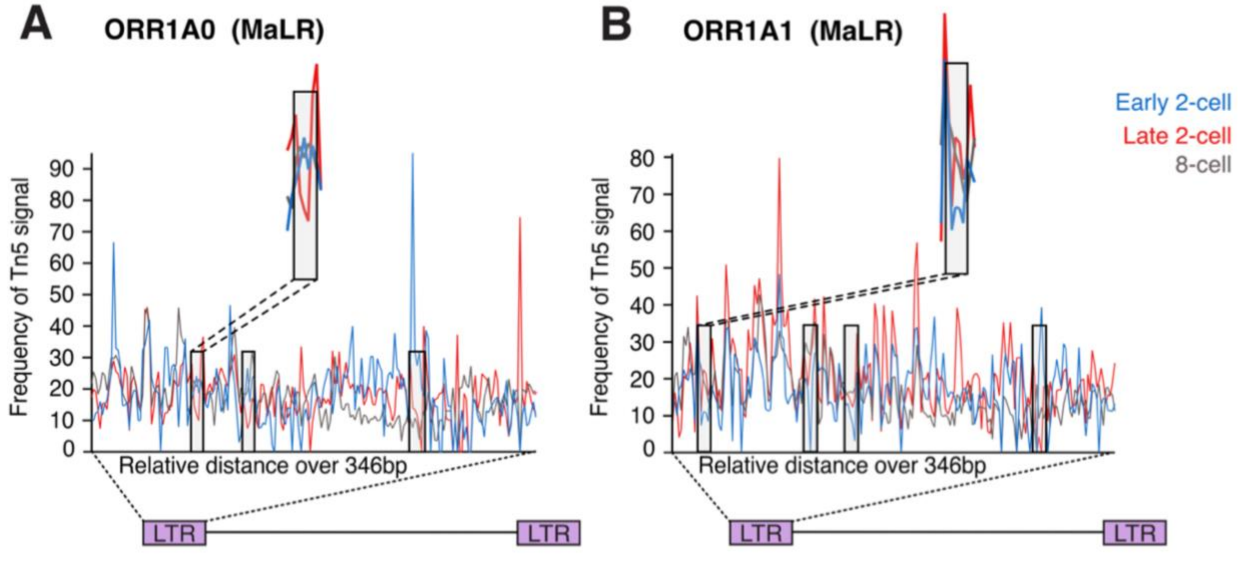


Figure 12. Tn5 insertion frequency plots over ORR1A0 (A) and ORR1A1 (B). (A, B) Early 2-cell insertion frequency is displayed in blue, late 2-cell in red and 8-cell in brown. The sizes of ORR1A0 and ORR1A1 consensus sequences are indicated. The insets are zooms on specific footprints.

In L1MdTf_II monomers, the Tn5 insertion frequency signal was clearly higher at the late 2-cell stage and the 8-cell stage compared to the early 2-cell stage (**Figure 13**), which aligns with the transcriptional activation timing of L1MdTf_II shown in Figure 8. Nonetheless, we identified one early 2-cell specific footprint as well as three late 2-cell specific footprints within the monomers.

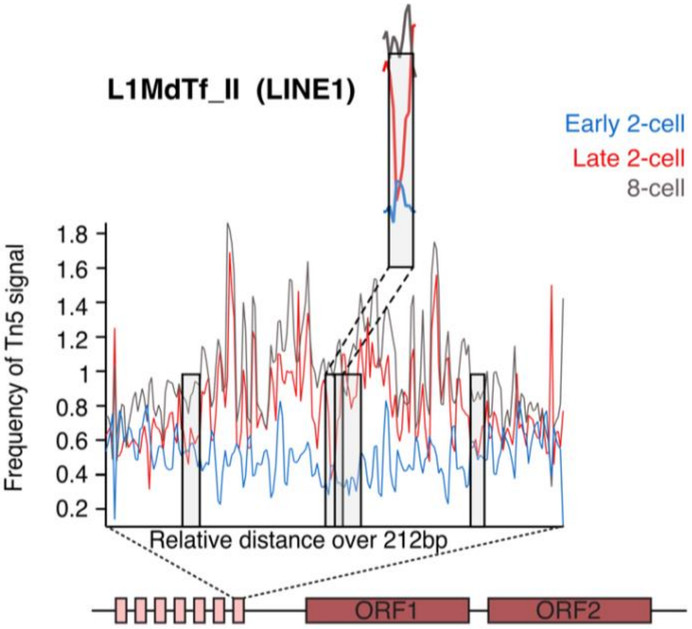


Figure 13. Tn5 insertion frequency plot over L1MdTf_II monomers. Early 2-cell insertion frequency is displayed in blue, late 2-cell in red and 8-cell in brown. The size of the L1MdTf_II monomer consensus sequence is indicated. The inset is a zoom on a specific footprint.

Finally, in both SINEB2 families analyzed, we found two late 2-cell footprints (**Figure 14A, B**). Interestingly, for both B2_Mm1a and B3A, one of these footprints was located at the 3' end of the element, possibly overlapping with the reverse RNAPII promoter region (**Figure 14A, B**).

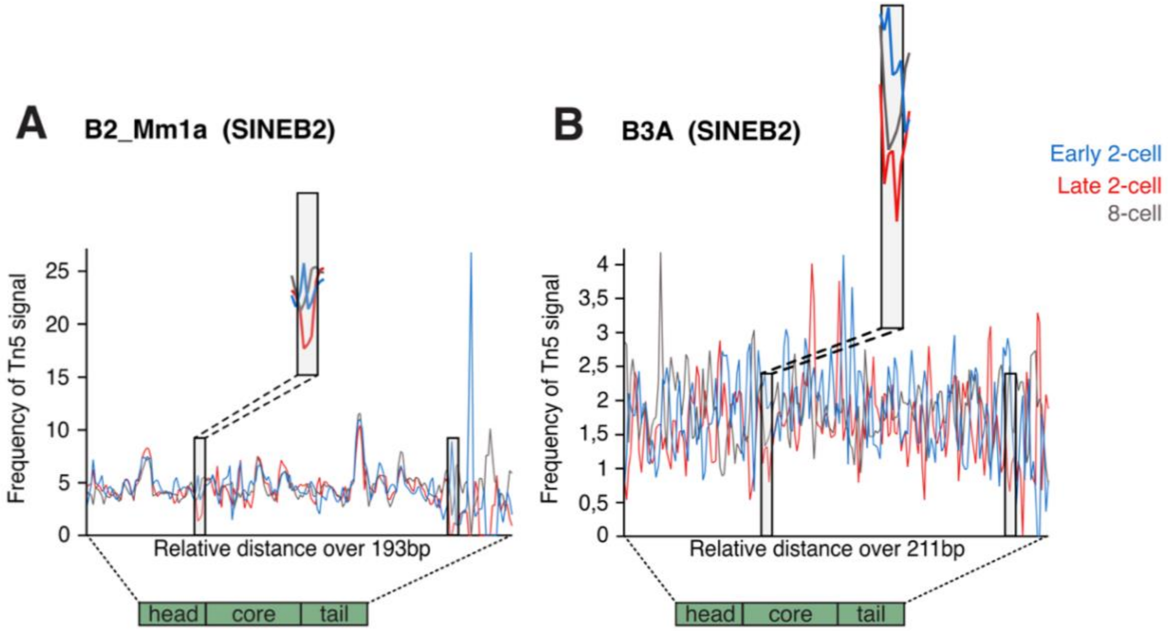


Figure 14. Tn5 insertion frequency plots over B2_Mm1a (A) and B3A (B) complete elements. (A, B) Early 2-cell insertion frequency is displayed in blue, late 2-cell in red and 8-cell in brown. The sizes of B2_Mm1a and B3A consensus sequences are indicated. The insets are zooms on specific footprints.

At the corresponding position of the identified footprints within the consensus sequence of each TE of interest, we extracted the DNA sequence. Using the Tomtom tool from the MEME suite (Bailey et al. 2015), which compares input DNA sequence against the UniPROBE mouse dataset (Newburger and Bulyk 2009; Hume et al. 2015), we identified putative TFBS (**Figure 10**). This analysis resulted in an extensive list of potential TFs, which I then refined by filtering based on their expression at the 2-cell stage, using publicly available RNA-seq datasets from preimplantation development (Deng et al. 2014), retaining only those TFs expressed at the 2-cell stage. Each TF had two binding motifs potentially, a primary and a secondary, which could exist in either forward or reverse orientation and could be located within different footprints, either within the same element, or between different elements (**Appendix 1**). Altogether, this process resulted in a list of 54 potential TFs, with considerable overlap across the six families, despite large evolutionary distance separating them (**Figure 15**). Using the same RNA-seq dataset as for the filtering step (Deng et al. 2014), I looked at the expression profiles of these TFs, and found that several TFs had transcripts in zygotes, pointing to maternal inheritance, while others displayed expression pattern concomitant with ZGA (**Figure 16**).

To summarize, through a footprinting analysis over six TE families using accessibility data spanning stages prior, during and after the major wave of ZGA, we identified **54 potential new TE regulators**, which are expressed during development. This work therefore expands the pool of putative TFs involved in regulating TEs at the onset of development.

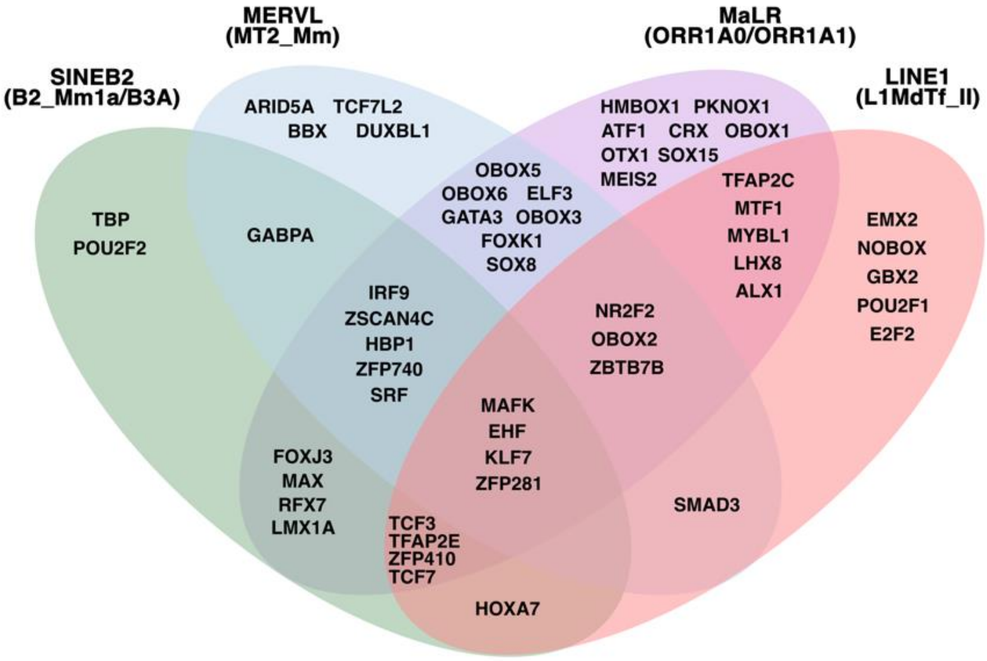


Figure 15. Venn diagram showing the 54 TFs obtained from the footprinting analysis within each TE family.

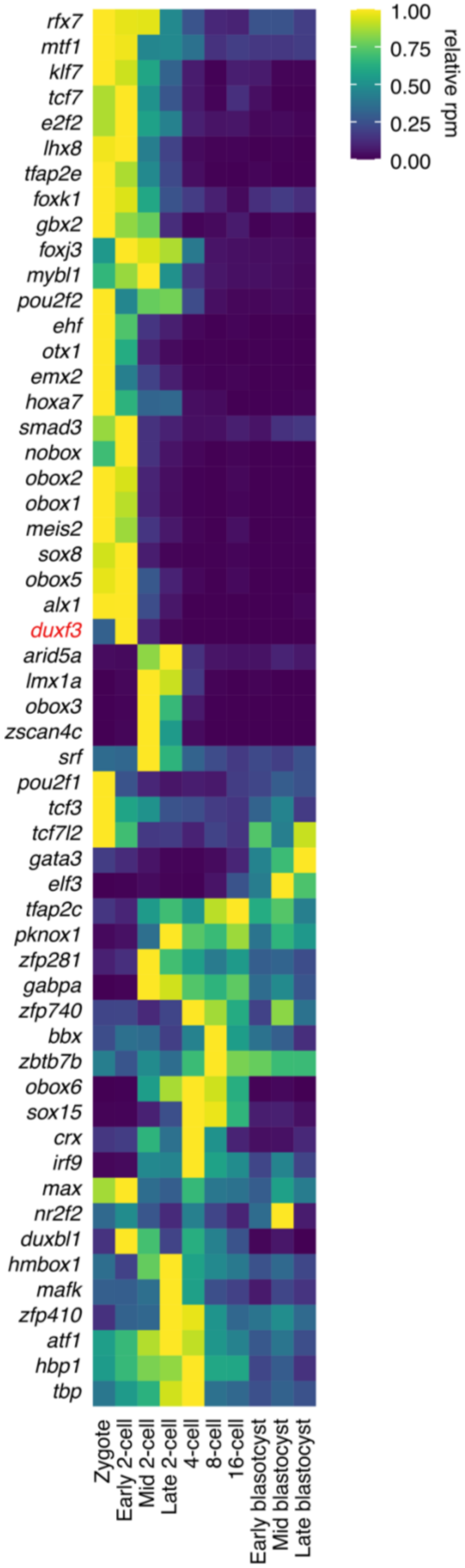


Figure 16. Expression patterns of the candidate TFs during preimplantation development. Heatmap showing the expression of the 54 candidate TFs (with *Duxf3* added) across all the stages of early embryo development, reanalyzed from (Deng et al. 2014). Values are normalized counts centered around each row mean. TFs were ordered by unsupervised hierarchical clustering.

4.2 Sequence heterogeneity and transcription factor binding motifs: insights from phylogenetic analysis

Setting up the phylogenetic analysis pipeline was performed in collaboration with Luis Altamirano-Pacheco.

To investigate further the regulation of these TE families, I decided to take an evolutionary approach. While the footprinting enabled me to generate a list of potential new candidate TFs involved in regulating TEs during development, this analysis was based on the TE families consensus sequences. By definition, a consensus sequence is an average over a group of sequences. As a result, this could mask potential heterogeneity among the individual TE insertions within each family across the genome (Carter et al. 2022). Therefore, I reasoned that examining the phylogenetic relationships between individual insertions could uncover sequence heterogeneity, which may reflect underlying regulatory differences. For each TE family, we took all insertions within the genome and performed a size distribution analysis (**Figure 17**). Based on the obtained size distribution, we selected length corresponding to the consensus of intact elements and therefore excluded fragmented elements from further analysis. We then performed a multiple sequence alignment (MSA) using MUSCLE (Sievers et al. 2011) which we used as input for phylogenetic tree reconstruction using IQ-TREE (Minh et al. 2020) (**Figure 17**).

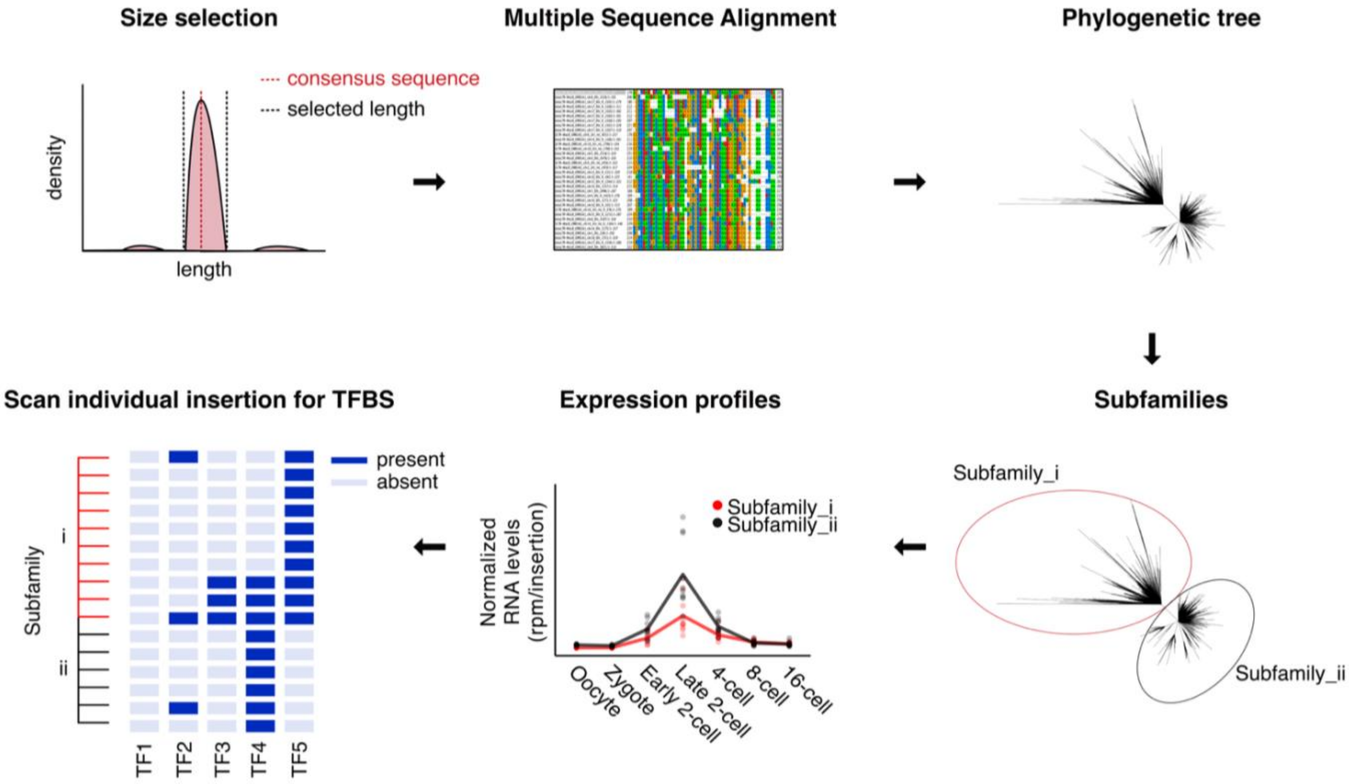


Figure 17. Schematic representation of the phyloregulatory analysis pipeline.

Using semi-arbitrary criteria following a method (a minimum of 10 sequences, on a high confidence node (>95%) that separates from the rest of the tree with a branch longer than 0.015) previously used for similar TE analyses of LTR families in the human genome (Carter et al. 2022), we clustered individual insertions into new subfamilies for each TE family (**Figure 17**). However, the criteria that I used were adjusted for each TE family based on the tree structure and in order to obtain clear clades, that would not lead to a high number of subfamilies containing too few insertions as these would likely be uninformative. To investigate the biological significance of these newly defined subfamilies, I first plotted their expression across development. This allowed me to assess whether the expression patterns or expression intensities between subfamilies differed (**Figure 17**). Finally, using the list of TFs identified within the footprints of each family consensus sequence from **section 4.1.2**, I scanned individual insertions for the presence of these TFBS. This approach, which we refer to as phyloregulatory analysis, aimed to identify potential sequence determinants that might account for observed differences in expression (**Figure 17**).
I performed this analysis on L1MdTf_II monomers using the annotation that we generated using the consensus sequence (see methods section) and used for the expression (**section 4.1.1**) and footprinting (**section 4.1.2**) analyses. The monomers were very homogenous in sequence and the phylogenetic analysis led to four subfamilies, containing one main large clade and three very small clades. I therefore decided to remove their description from this section. The results can be found in **Appendix 2**.

4.2.1 Phyloregulatory analysis of SINEB2

SINEB2 elements represent large families with a high number of insertions (**Figure 9,** *n* numbers), making phylogenetic analysis particularly challenging. A size distribution analysis of all insertions annotated as B2_Mm1a revealed that most elements cluster around the consensus length of 193bp (**Figure 18A**). I restricted the analysis to focus on elements within 189 – 195bp, encompassing approximately 6700 sequences, while excluding the density distribution tail that contained numerous slightly shorter elements (**Figure 18A**).

For B3A, the oldest SINEB2 family, the size distribution was markedly broad, with only a few elements matching consensus length (**Figure 18B**). This is consistent with the expected TE erosion over evolutionary time, eventually resulting in a large proportion of fragmented elements. To align with the footprinting analysis in which I used the consensus sequence length, I applied a narrow size distribution around 211bp (**Figure 18B**), capturing around 1000 sequences.

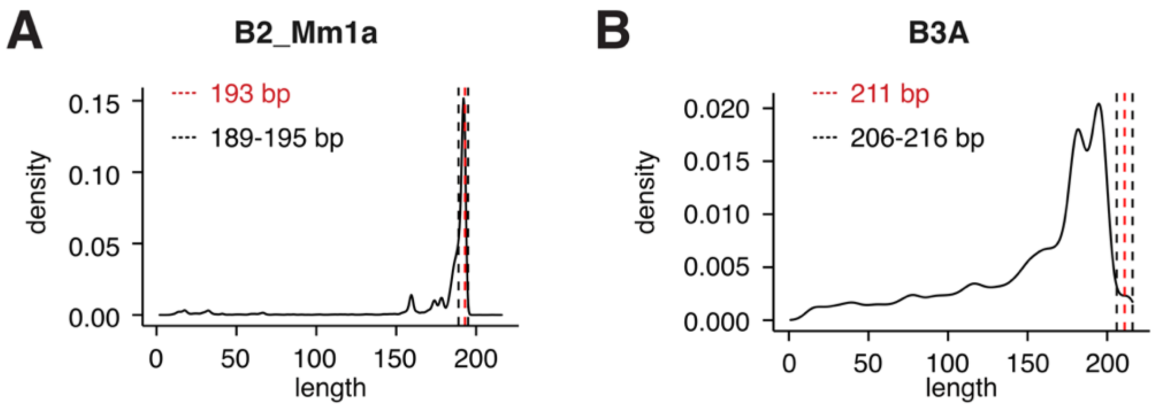


Figure 18. Size density distribution plots of B2_Mm1a (A) and B3A (B) full elements. (A, B) For each family the consensus length is displayed in a red dashed line. The selected size range is indicated in black dashed lines.

The B2_Mm1a phylogenetic tree exhibited considerable heterogeneity. While the tree structure featured a central major node with numerous attached branches, the branch lengths were relatively long (**Figure 19A**), indicating large genetic distance between insertions.

By increasing the minimum number of leaves required to define a subfamily to 75, and applying the same other two criteria, I identified 7 distinct subfamilies. These subfamilies varied greatly in size, ranging from 75 insertions in the smallest to 5971 insertions in the largest, corresponding to the main node (**Figure 19A**). Examining the expression of these subfamilies, I observed that while the patterns appeared similar to each other, small variations were evident (**Figure 19B**). For example, subfamilies B2_Mm1a_i, ii, iii and vii all exhibited an expression profile like that of the entire B2_Mm1a group, with a peak at the 8-cell stage (**Figures 19B and 9D**). In contrast, B2_Mm1a_iv and v peaked at the 4-cell stage and declined at the 8-cell stage (**Figure 19B**). B2_Mm1a_vi, however, displayed a unique pattern, peaking at the late 2-cell stage and slightly decreasing by the 4-cell stage (**Figure 19B**). Analysis of the expression of individual insertions at the late 2-cell stage revealed that all subfamilies contained insertions that were expressed, with B2_Mm1a_i containing most the of highest expressed individual insertions (**Figure 19C**).

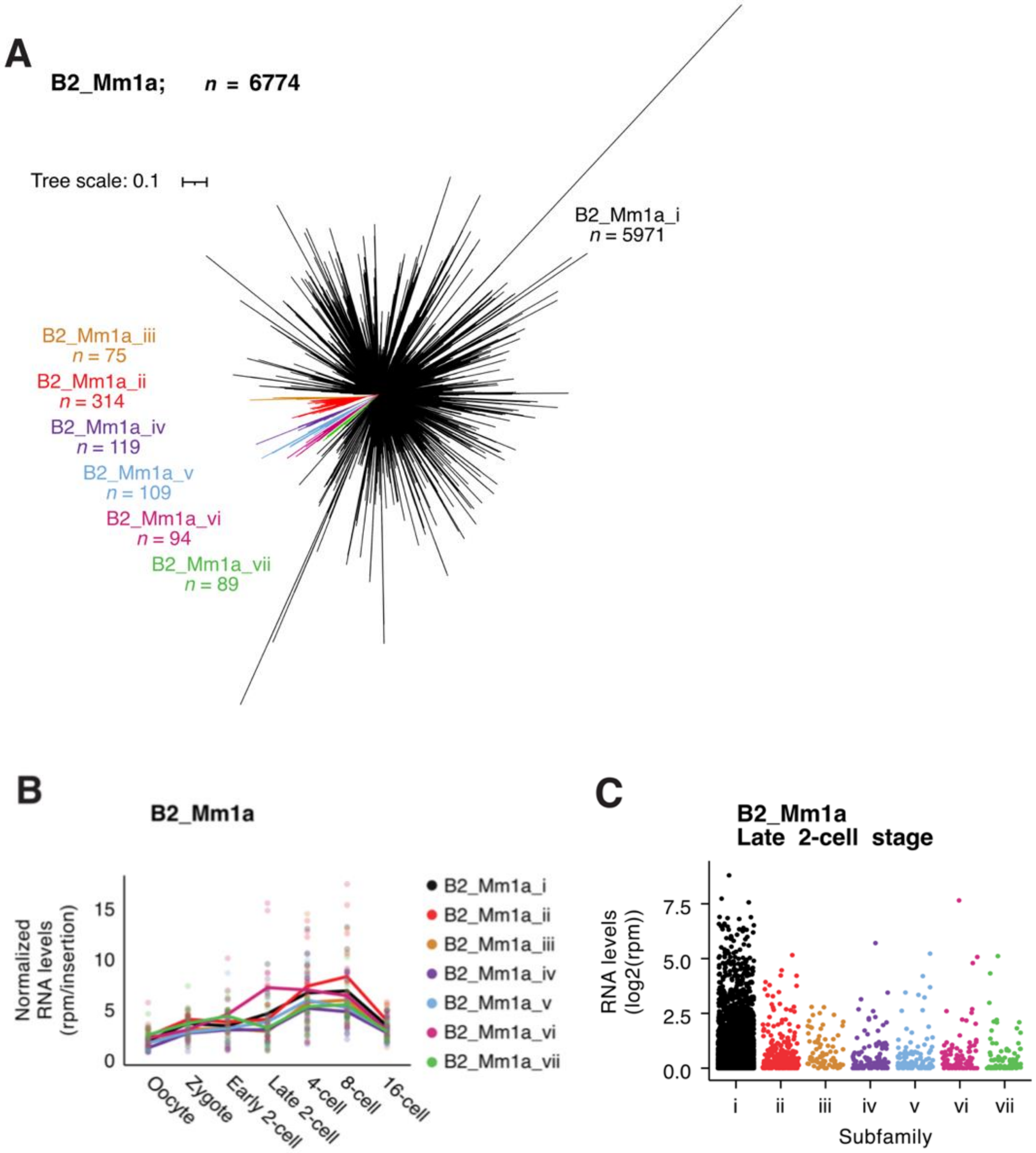


Figure 19. Phylogenetic analysis of B2_Mm1a. (A) Unrooted phylogenetic tree of B2_Mm1a insertions ranging between 189 and 195bp in length. *n* is the number of sequences per group indicated. The sequences in gray correspond to outgroups which did not qualify as subfamily (3 insertions). The tree scale is indicated. (B) Expression of all B2_Mm1a subfamilies across the different stages of preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot is the mean rpm of all insertions belonging to the indicated subfamily (rpm/insertion number) per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages. (C) Expression of single B2_Mm1a insertions at the late 2-cell stage. Data was reanalyzed from (Oomen et al. 2025). Each point is the log2 transformed mean rpm value of a single insertion in all late 2-cell stage embryos.

The expression pattern differences led me to investigate the variability in TFBS among individual insertions and across subfamilies. Scanning for the presence of TFBS identified in **section 4.1.2** revealed a strong prevalence of FOXJ3 and TBP across nearly all B2_Mm1a insertions (**Figure 20**). B2_Mm1a_vii, which shared expression profile with B2_Mm1a_i, ii and iii, lacked both FOXJ3 and TBP binding sites (**Figure 20**), suggesting that these binding sites may not be sufficient for expression at these stages. B2_Mm1a_iv, which peaked at the 4-cell stage, was characterized by the presence of a LMX1A binding sites and the absence of FOXJ3 binding sites (**Figure 20**). Finally, B2_Mm1a_iii displayed all three binding sites: LMX1A, FOXJ3 and TBP (**Figure 20**).

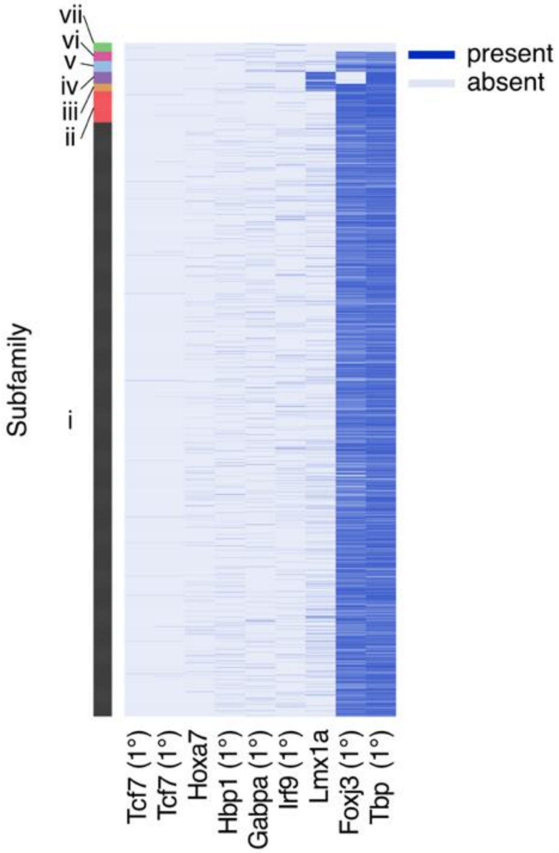


Figure 20. Heatmap showing the presence or absence of TFBS in B2_Mm1a insertions. Insertions are ordered by subfamily indicated by the bar legend on the left, and TFs are clustered based on the number of insertions that contain the binding site. 1° and 2° refer to “primary” and “secondary” binding sites.

While phylogenetic analysis and transcription factor binding profiling appeared to be related, subfamily expression did not correlate with the presence or absence of specific TFBS in the case of B2_Mm1a. Further, despite heterogeneity among individual B2_Mm1a insertions, their phyloregulatory analysis revealed near-ubiquitous presence of TFBS for FOXJ3 and TBP.

I further analyzed SINEB2 phylogenies by focusing on the B3A family. It is important to note that only a very small fraction of B3A elements were selected and therefore the phylogeny only represents a small subset of B3A elements. The phylogenetic analysis revealed very distinct clades, which I classified from B3A_i to B3A_vi, each with a relatively consistent number of insertions, from 96 in the smallest subfamily to 329 in the largest (**Figure 21A**).

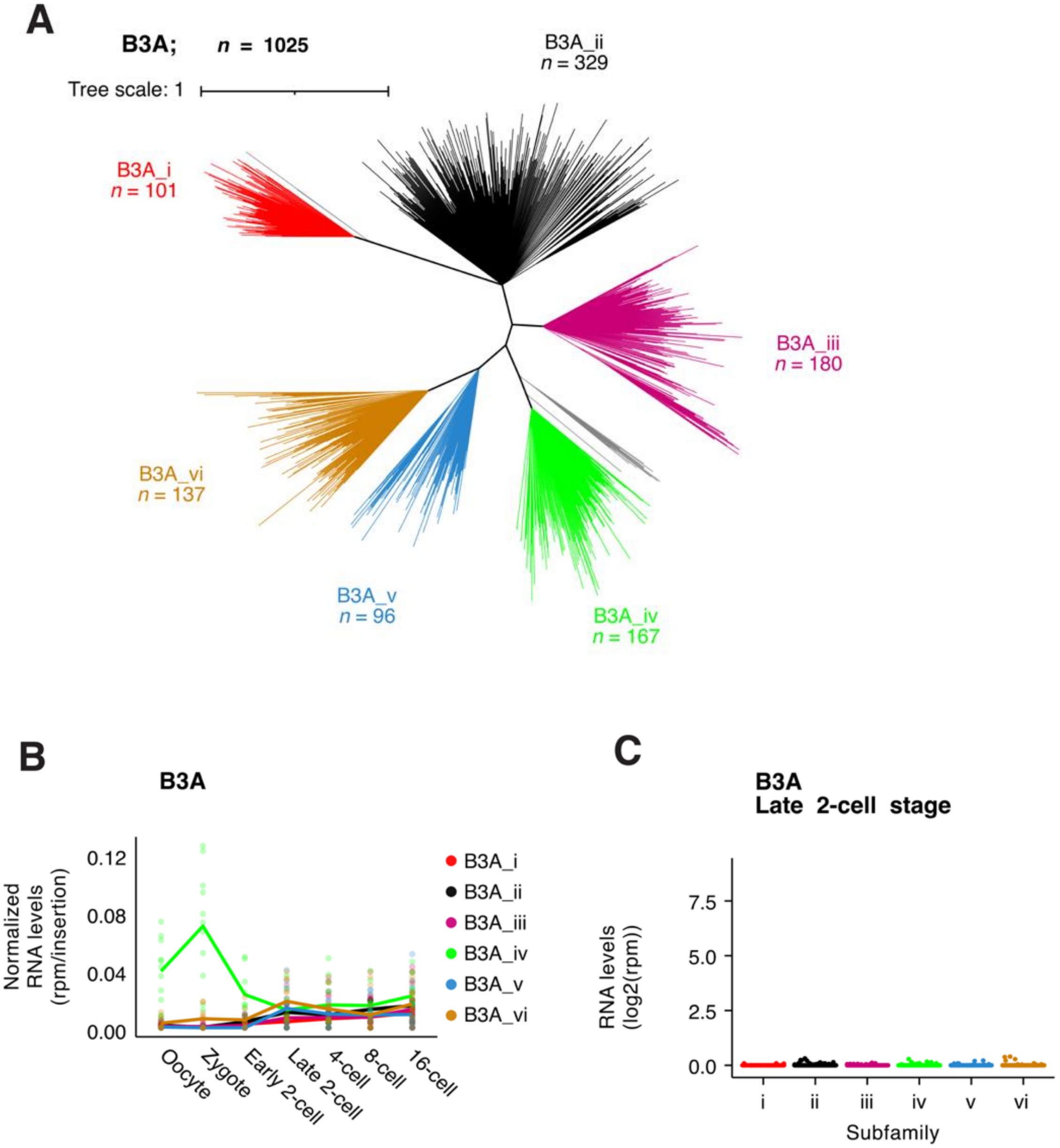


Figure 21. Phylogenetic analysis of B3A. (A) Unrooted phylogenetic tree of B3A insertions ranging between 206 and 216bp in length. *n* is the number of sequences per group indicated. The sequences in gray correspond to outgroups which did not qualify as subfamily (15 insertions, in different groups). The tree scale is indicated. (B) Expression of all B3A subfamilies across the different stages of preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot is the mean rpm of all insertions belonging to the indicated subfamily (rpm/insertion number) per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages. (C) Expression of single B3A insertions at the late 2-cell stage. Data was reanalyzed from (Oomen et al. 2025). Each point is the log2 transformed mean rpm value of a single insertion in all late 2-cell stage embryos.

Analysis of expression per insertion across these subfamilies revealed strikingly low expression levels, with highest being B3A_iv in zygotes, barely reaching 1 rpm/insertion (**Figure 21B**). The expression of B3A_iv was also in sharp contrast with the expression pattern of the entire B3A family (**Figure 9B**). Examining individual insertion expression showed that all subfamilies exhibited very low, if any, expression of individual insertions at the late 2-cell stage (**Figure 21C**).

POU2F2 and ZSCAN4C were the main two TFs present in many individual B3A insertions (**Figure 22**). However, I did not observe any pattern linking the differential expression of B3A_iv to the presence of specific TFBS.

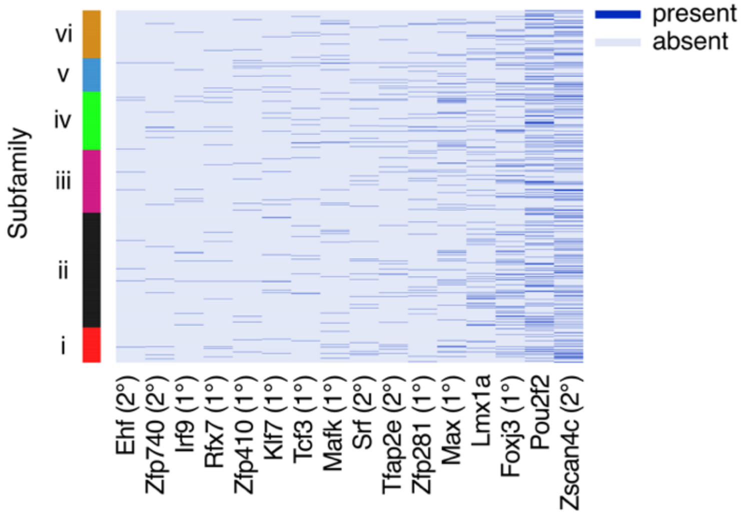


Figure 22. Heatmap showing the presence or absence of TFBS in B3A insertions. Insertions are ordered by subfamily indicated by the bar legend on the left, and TFs are clustered based on the number of insertions that contain the binding site. 1° and 2° refer to "primary" and "secondary" binding sites.

4.2.2 Phyloregulatory analysis of MaLRs

MaLR LTRs exist in three forms within the genome: the 5', 3' and solo LTRs. Size distribution analysis of ORR1A0 revealed that most insertions correspond to intact LTRs (**Figure 23A**), with the majority displaying lengths similar to the consensus sequence (346bp). Insertions ranging from 282 to 410bp were selected (**Figure 23A**), covering about 1800 sequences. However, while most ORR1A1 5' and 3' LTRs were approximately 346bp long, matching the consensus length, a subset of solo LTRs was shorter, ranging from 250bp to 300bp (**Figure 23B**). Manual inspection revealed that these shorter insertions, around 300 sequences in total, were primarily truncated solo LTRs located on the Y chromosome (**data not shown**). I decided to focus on intact LTRs, using the same length range as for ORR1A0 (**Figure 23**), excluding these shorter fragments from the size selection, and resulting in about 3000 insertions.

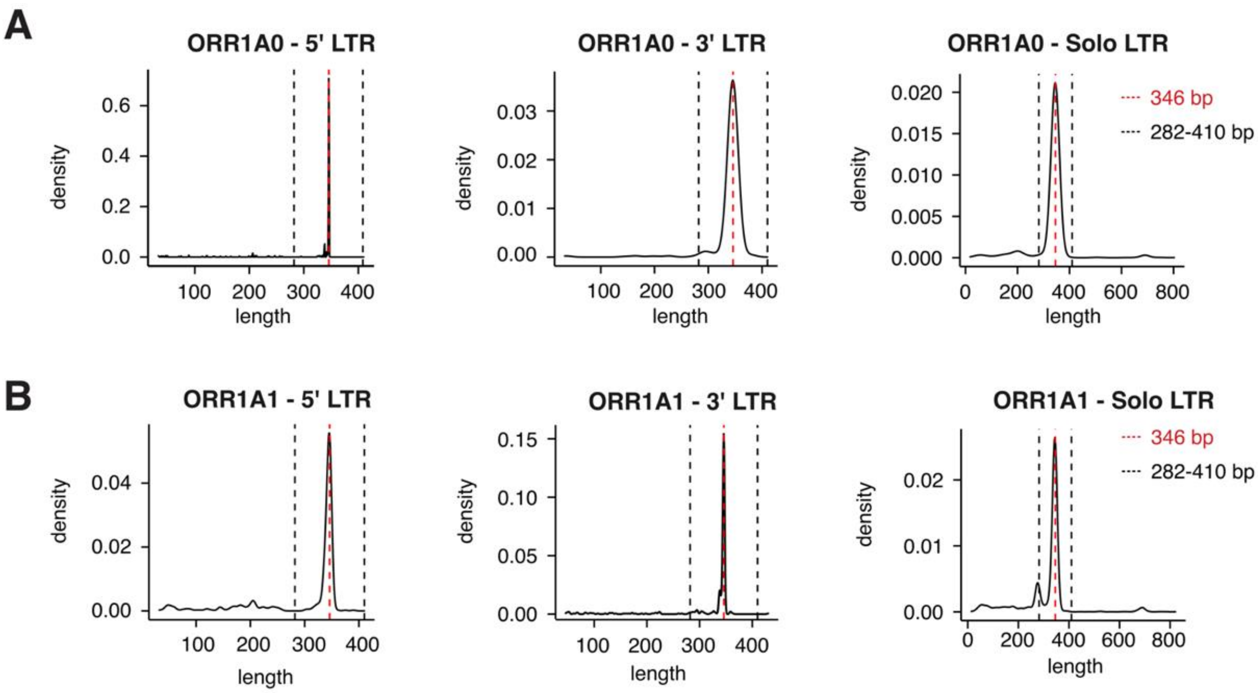


Figure 23. Size density distribution plots of 5', 3' and solo ORR1A0 (A) and ORR1A1 (B) LTRs. (A, B) For each family the consensus length is displayed in red dashed line. The selected size range is indicated in black dashed lines.

The ORR1A0 phylogenetic tree resembled that of B2_Mm1a, with a central node and radiating branches (**Figure 24A**). While subfamilies were identifiable due to heterogeneity, sequence variation was less than in B2_Mm1a (**Figures 19A and 24A**). By defining subfamilies with a minimum of 20 sequences, I identified seven subfamilies, with insertions counts ranging between 21 to 1666 (**Figure 24A**). The smallest subfamily, ORR1A0_vii, was characterized by very long branches, probably constituting a cluster of very divergent insertions (**Figure 24A**).

While most ORR1A0 subfamilies shared a global expression pattern with to the entire ORR1A0 family (**Figures 24B and 7C**), there were significant differences in expression intensities across subfamilies (**Figure 24B**). Specifically, subfamilies ORR1A0_i and iv were generally more expressed than the other five subfamilies (**Figure 24B**). These subfamilies also contained the highest expressed individual insertions, even though all subfamilies included insertions expressed at the late 2-cell stage (**Figure 24C**). TFBS analysis revealed no consistent TFBS pattern correlating with expression (**Figure 25**). In fact, most insertions lacked the examined binding sites, except perhaps KLF7 which was found in many sequences across subfamilies (**Figure 25**).

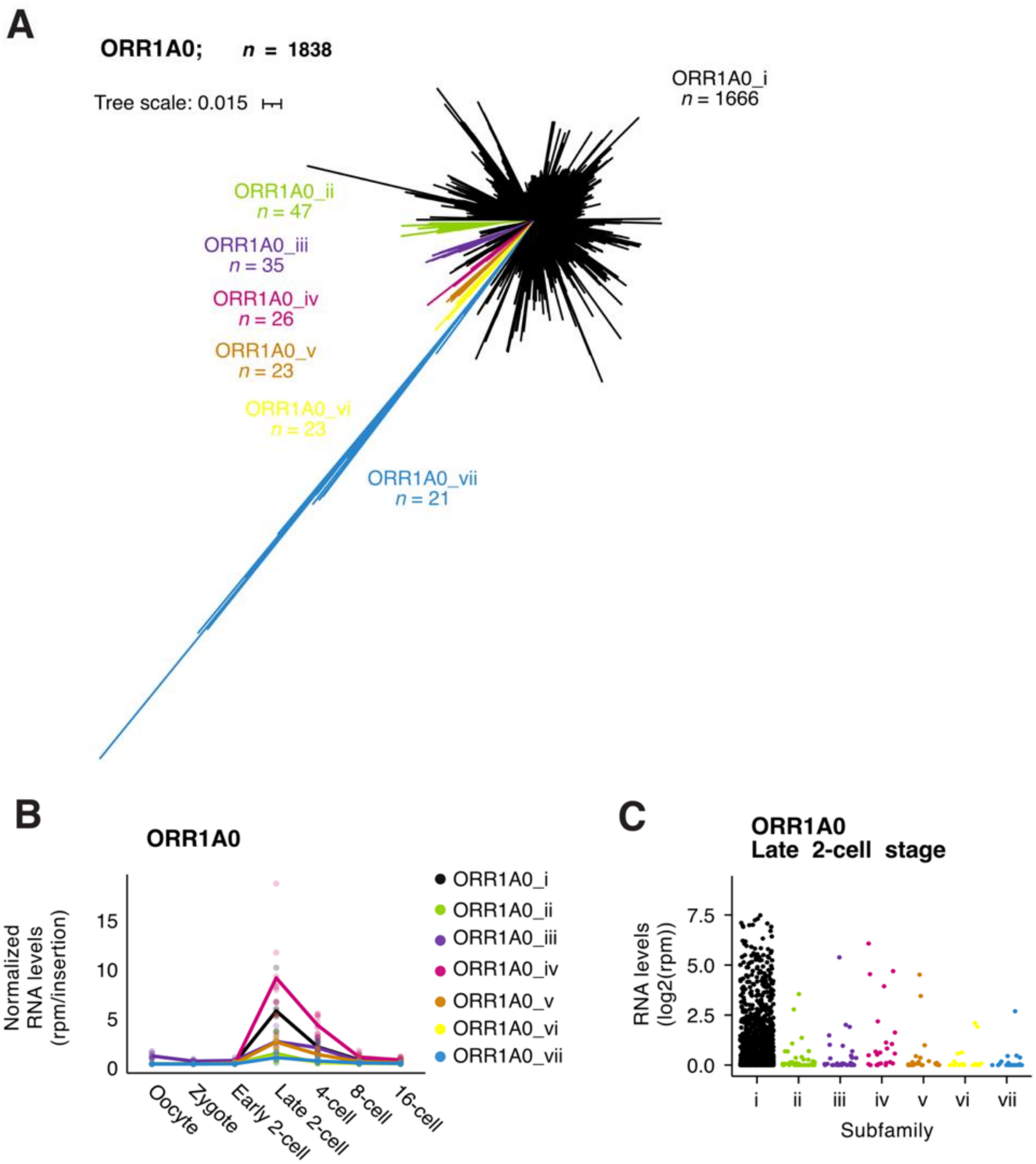


Figure 24. Phylogenetic analysis of ORR1A0. (A) Unrooted phylogenetic tree of ORR1A0 insertions ranging between 282 and 410bp in length. *n* is the number of sequences per group indicated. The tree was pruned to remove 3 sequences corresponding to an outgroup. The tree scale is indicated. (B) Expression of all ORR1A0 subfamilies across the different stages of preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot is the mean rpm of all insertions belonging to the indicated subfamily (rpm/insertion number) per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages. (C) Expression of single ORR1A0 insertions at the late 2-cell stage. Data was reanalyzed from (Oomen et al. 2025). Each point is the log2 transformed mean rpm value of a single insertion in all late 2-cell stage embryos.

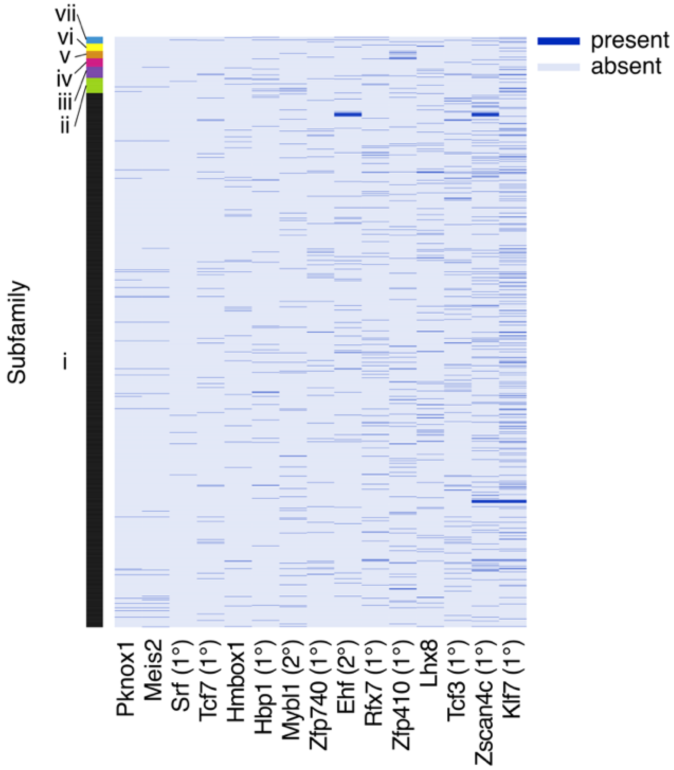


Figure 25. Heatmap showing the presence or absence of TFBS in ORR1A0 insertions. Insertions are ordered by subfamily indicated by the bar legend on the left, and TFs are clustered based on the number of insertions that contain the binding site. 1° and 2° refer to "primary" and "secondary" binding sites.

The ORR1A1 phylogenetic tree displayed a more complex structure than that of ORR1A0, with longer branch lengths, suggesting greater heterogeneity (**Figure 26A**). When aligning the consensus sequences of ORR1A0, ORR1A1 and their evolutionary predecessors ORR1B1 and B2 (Franke et al. 2017), I noticed that ORR1A1 is older than ORR1A0, as it is more closely related to ORR1B elements (**data not shown**). Therefore, it is perhaps not surprising that ORR1A1 elements show greater divergence from each other. I set the minimum number of insertions per subfamilies to 50 and increased the minimum branch length required, which resulted in the identification of 8 clear new subfamilies (**Figure 26A**). The three was characterized by two major clades, separating ORR1A1_i to v from ORR1A1_vi to viii (**Figure 26A**). Subfamilies ORR1A1_i to v contained most of the insertions and were more similar to each other, as indicated by shorter branch lengths, whereas subfamilies ORR1A1_vi to viii consisted primarily of long branches (**Figure 26A**).

Strikingly, mainly subfamilies ORR1A1_i to v were expressed during preimplantation development, following the same expression pattern as described for the complete ORR1A1 family, that is, a global increase in expression levels at the late 2-cell stage (**Figures 26B and 7D**). Indeed, subfamilies vi, vii and viii were almost not expressed at all, and this was also clear when looking at individual insertion expression (**Figure 26B, C**). Remarkably, expression during preimplantation development was correlated with the presence of binding sites for the TFs SRF, FOXK1 and FOXJ3 (**Figure 27**). Indeed, the binding sites for these TFs were mainly found in subfamilies ORR1A1_i to v, and were either sparser, or in the case of SRF, almost absent from subfamilies ORR1A1_vi to viii (**Figure 27**), potentially outlining the observed differences in expression in the embryo.

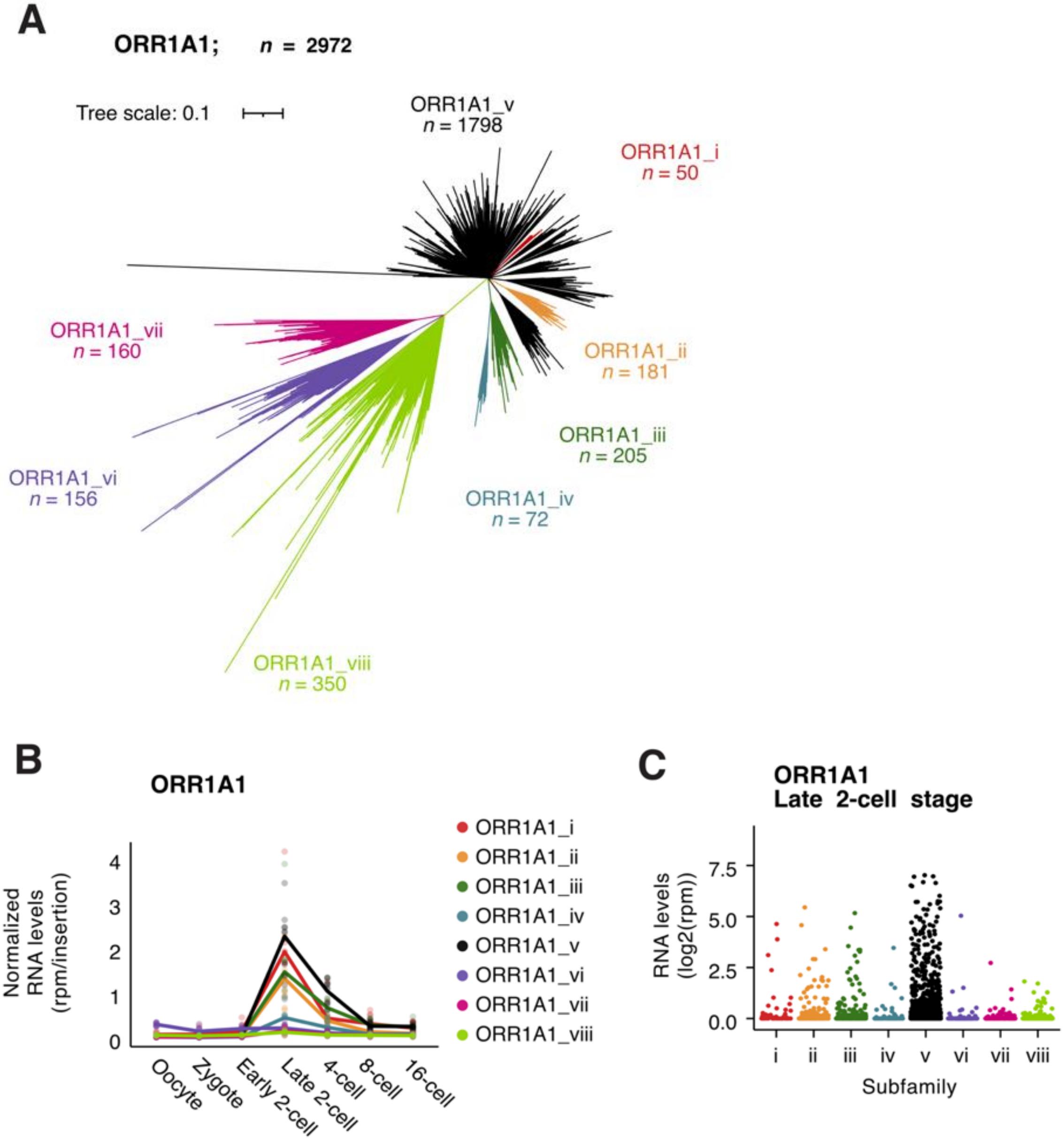


Figure 26. Phylogenetic analysis of ORR1A1. (A) Unrooted phylogenetic tree of ORR1A1 insertions ranging between 282 and 410bp in length. *n* is the number of sequences per group indicated. The tree scale is indicated. (B) Expression of all ORR1A1 subfamilies across the different stages of preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot is the mean rpm of all insertions belonging to the indicated subfamily (rpm/insertion number) per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages. (C) Expression of single ORR1A1 insertions at the late 2-cell stage. Data was reanalyzed from (Oomen et al. 2025). Each point is the log2 transformed mean rpm value of a single insertion in all late 2-cell stage embryos.

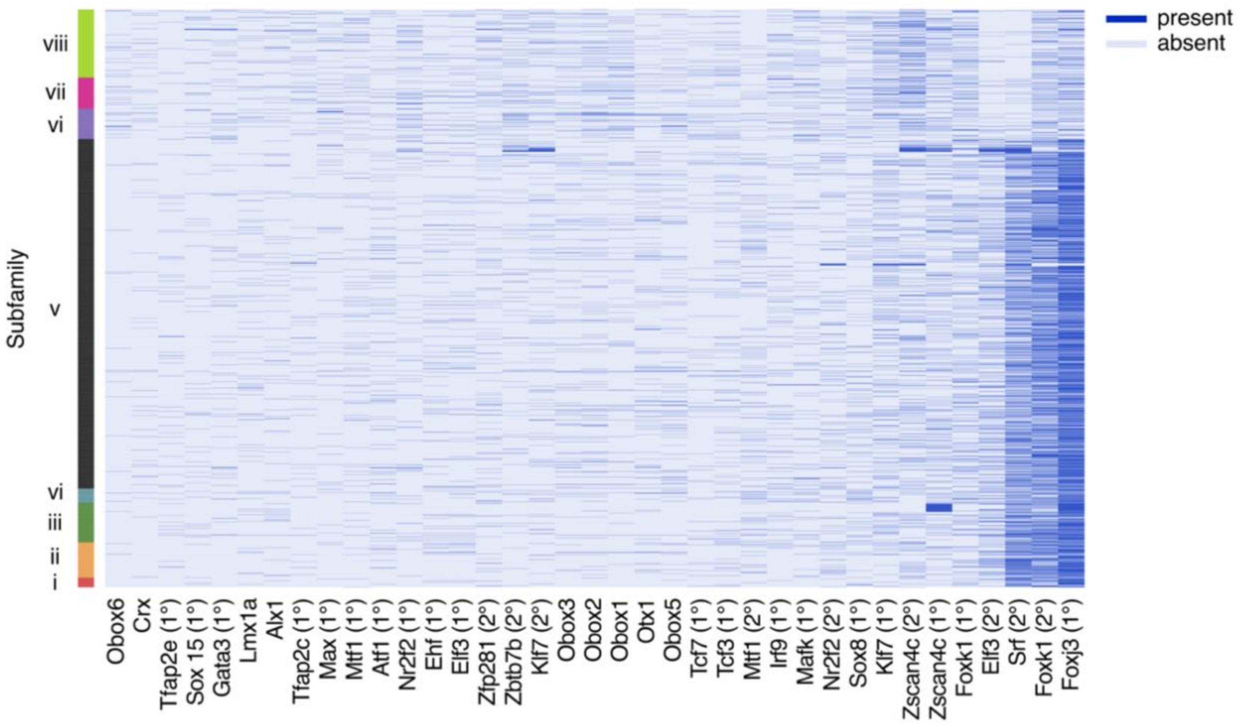


Figure 27. Heatmap showing the presence or absence of TFBS in ORR1A1 insertions. Insertions are ordered by subfamily indicated by the bar legend on the left, and TFs are clustered based on the number of insertions that contain the binding site. 1° and 2° refer to "primary" and "secondary" binding sites.

4.2.3 Phyloregulatory analysis of MT2_Mm

This part of my PhD was performed in collaboration with Carlos Michel Mourra-Díaz.

As with the previous four TE families, we began analyzing the evolution of the MERVL LTR, MT2_Mm by conducting a size distribution analysis. Much like the other LTR families, this analysis showed that most MT2_Mm insertions, including 5', 3' and solo LTRs, are intact, with most insertions matching the length of the known consensus sequence (**Figure 28**). We selected insertions considered full length, ranging from 400 to 586bp in length, encompassing a total of 2307 LTRs.

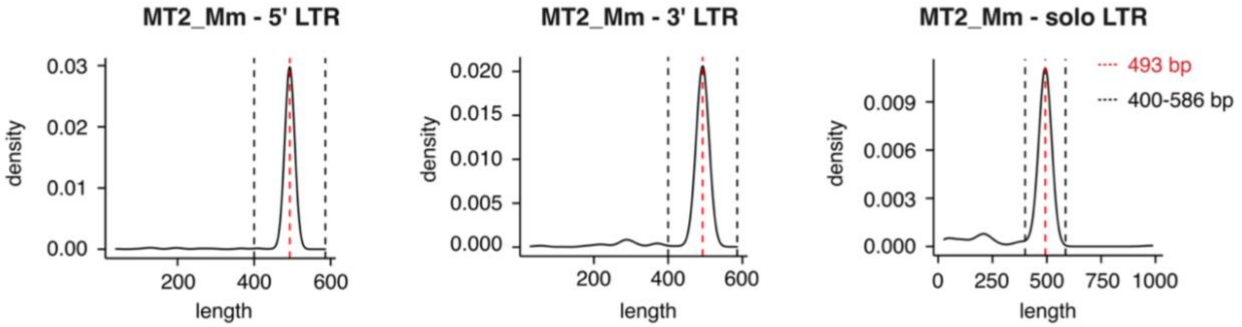


Figure 28. Size density distribution plots of 5', 3' and solo MT2_Mm LTRs. The consensus length is displayed in red dashed line. The selected size range is indicated in black dashed lines.

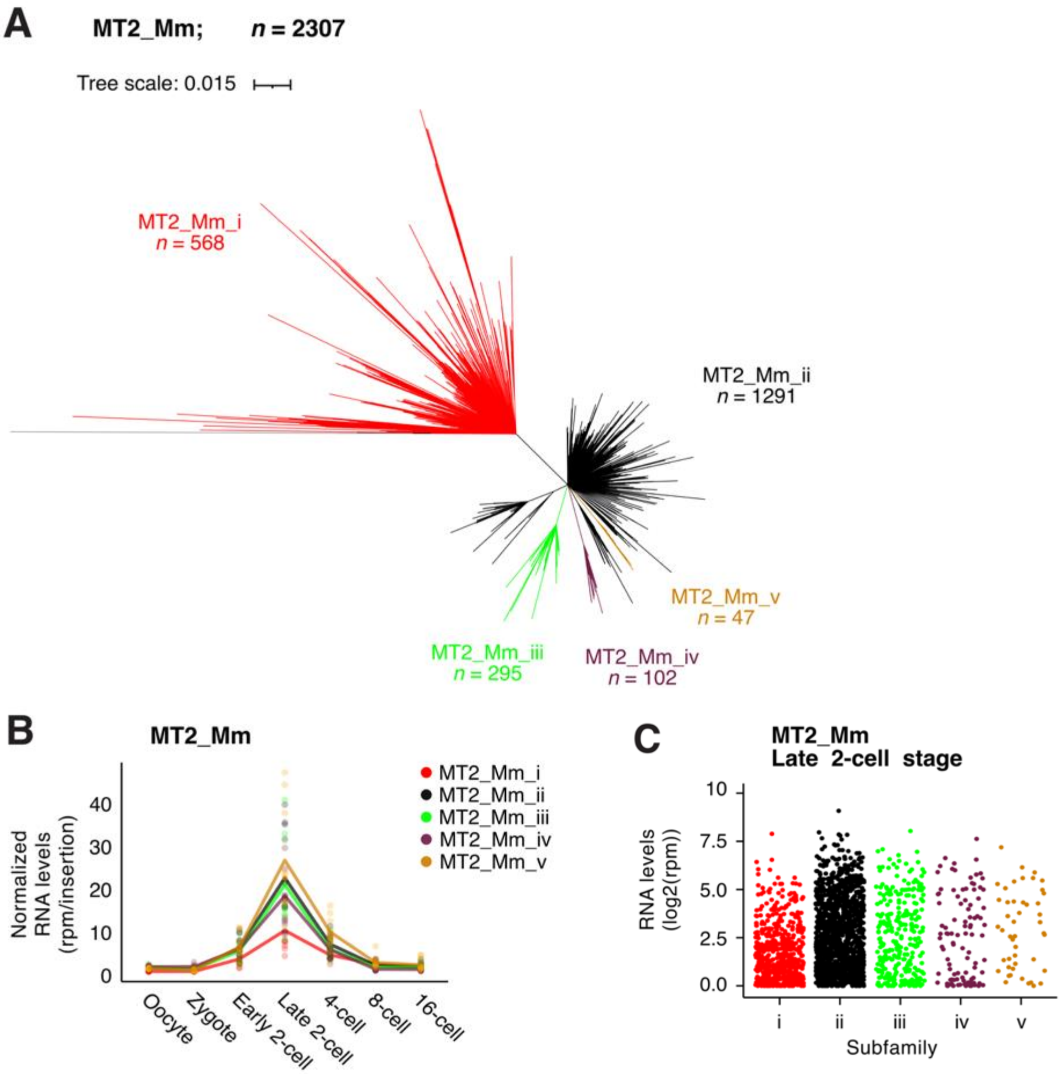


Figure 29. Phylogenetic analysis of MT2_Mm. (A) Unrooted phylogenetic tree of MT2_Mm insertions ranging between 400 and 586bp in length. *n* is the number of sequences per group indicated. The sequences in gray correspond to outgroups which did not qualify as subfamily (4 insertions). The tree scale is indicated. (B) Expression of all MT2_Mm subfamilies across the different stages of preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot is the mean rpm of all insertions belonging to the indicated subfamily (rpm/insertion number) per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages. (C) Expression of single MT2_Mm insertions at the late 2-cell stage. Data was reanalyzed from (Oomen et al. 2025). Each point is the log2 transformed mean rpm value of a single insertion in all late 2-cell stage embryos.

Phylogenetic tree reconstruction revealed a tree with two major clades, similarly to ORR1A1 (**Figure 29A**). We set the minimum number of sequences for a new subfamily to 30, and could describe five new subfamilies. MT2_Mm_i was the most evolutionary distant subfamily, constituting by itself one of the two major clades (**Figure 29A**). This subfamily was characterized by longer internal branch length suggesting evolutionary more distant MT2_Mm sequences

(**Figure 29A**). The other four subfamilies were closer to each other, and the number of insertions per subfamily varied from 47 in the smallest to 1291 in the largest (**Figure 29A**).

Interestingly, the expression of these five newly defined subfamilies followed the same overall pattern as MT2_Mm (**Figure 29B and 6**). Notwithstanding, the expression intensity between different subfamilies varied, with the most evolutionary distant subfamily, MT2_Mm_i, showing lower expression compared to the others (**Figure 29B**). Similarly, when plotting individual insertion expression, we observed that MT2_Mm_i did not contain as many highly expressed insertions as MT2_Mm_ii to v (**Figure 29C**). Interestingly, TFBS were associated with specific subfamilies, and some were correlating with differences in expression (**Figure 30**). DUX binding site was present in nearly all MT2_Mm insertions (**Figure 30**). ELF3 and EHF were restricted to MT2_Mm_iv (**Figure 30**). SRF and GABPA appeared mutually exclusive, whereby SRF was only present in the evolutionary more distant subfamily (MT2_Mm_i) and GABPA was found in all other subfamilies (**Figure 30**). Therefore, the presence of the GABPA binding site appeared to be associated with higher MT2_Mm expression (**Figure 30**).

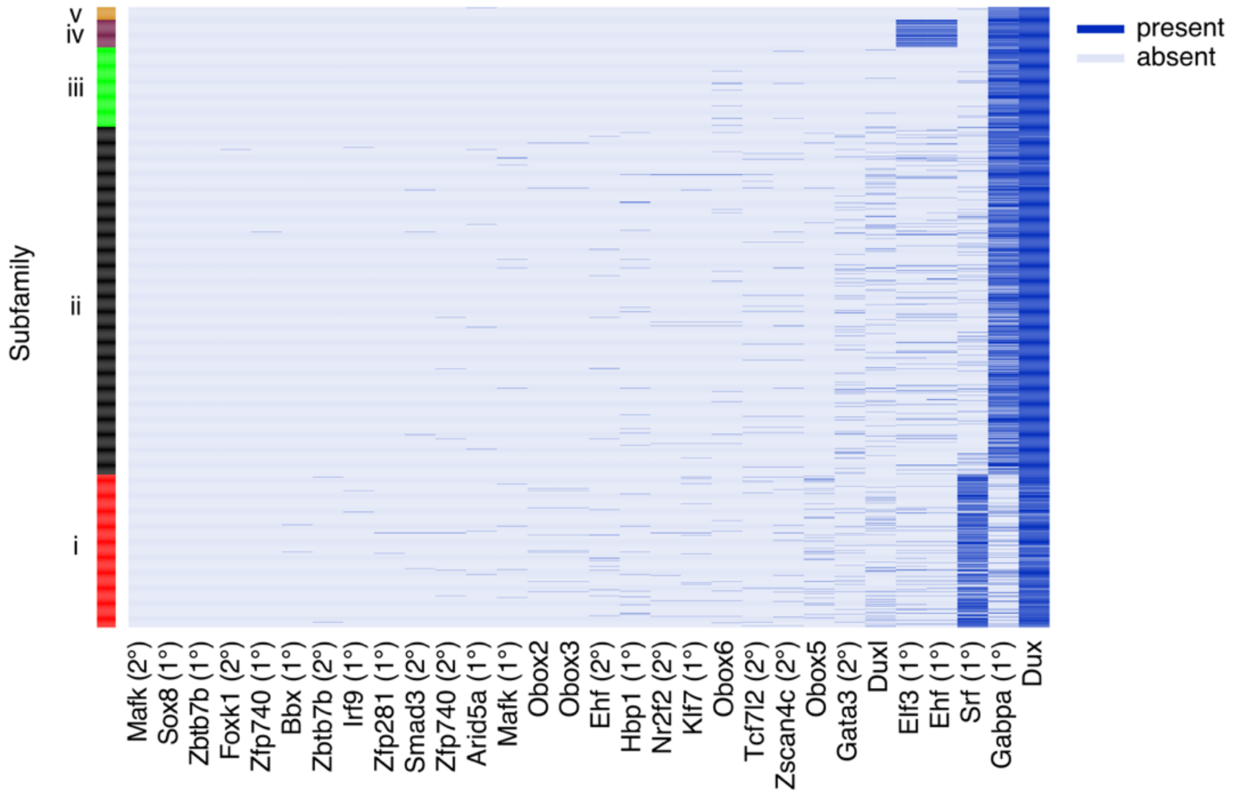


Figure 30. Heatmap showing the presence or absence of TFBS in MT2_Mm insertions. Insertions are ordered by subfamily indicated by the bar legend on the left, and TFs are clustered based on the number of insertions that contain the binding site. 1° and 2° refer to "primary" and "secondary" binding sites. We manually incorporated the DUX motif from (Hendrickson et al. 2017) since it is not present in the UniPROBE database.

In conclusion, phylogenetic analysis of five retrotransposon families revealed new subfamilies and highlighted previously unrecognized heterogeneity. We found that **FOXJ3** and **TBP** binding sites were nearly ubiquitous in B2_Mm1a insertions. In addition, we found that phylogeny and early embryonic expression of some subfamilies were linked to specific TFBS. Notably, the presence of **FOXJ3**, FOXK1 and **SRF** binding sites appeared to define the subfamilies identified through phylogenetic analysis of ORR1A1, while **SRF** and **GABPA** (also ELF3 and EHF) were associated with specific MT2_Mm subfamilies.

These findings suggest that the genomic expansion of these TE families may be linked to the acquisition and/or loss of these TFBS, and it provided a first foundation for refining the list of TFs selected for subsequent characterization.

4.2.4 Uncovering the evolutionary history of MT2_Mm

This part of my PhD was performed in collaboration with Carlos Michel Mourra-Díaz.

Given the pivotal role of MT2_Mm as a transcriptional regulator during early embryonic development (Peaston et al. 2004; Evsikov et al. 2004; Macfarlan et al. 2012), along with the distinct expression patterns and TFBS variations within MT2_Mm insertions across subfamilies, we sought to investigate deeper the evolutionary history of these elements.

It has been demonstrated for different TE families that younger elements tend to be more expressed than older ones, and this has, in some instances, been correlated to TFBS (DeBerardinis and Kazazian 1999; Goodier et al. 2001; Ito et al. 2017; Carter et al. 2022). Therefore, we decided to estimate the evolutionary age of each MT2_Mm insertion. We determined the genetic divergence of each insertion using an outgroup rooting method, whereby the phylogenetic tree is anchored with a known ancestor. This approach allowed us to estimate the genetic distance of each insertion to the root, thereby inferring their evolutionary age (Steel 2010; Kinene et al. 2016). MT2C_Mm was previously suggested as the closest ancestor of MT2_Mm (Franke et al. 2017), which we confirmed by computing the phylogenetic tree of the four main MT2 lineages (**Figure 31**).

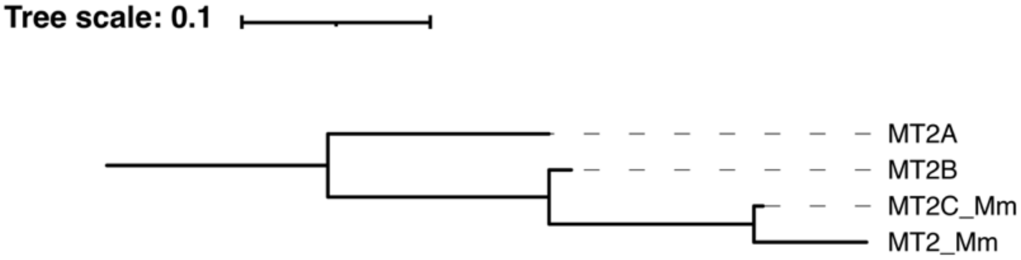


Figure 31. Rectangular phylogenetic tree of the consensus sequences of the main MT2 lineages.

Thus, we rooted the MT2_Mm phylogenetic tree using MT2C_Mm consensus sequence. The divergence analysis revealed that MT2_Mm_i, which we already knew was the most evolutionary distant subfamily (**Figure 29A**), was also the oldest MT2_Mm subfamily (**Figure 32**). MT2_Mm_i, along with MT2_Mm_ii and iii, also contained the most recent MT2_Mm insertions (**Figure 32**). This suggests that MT2_Mm_i was the first to expand, and prevailed within the mouse genome over evolutionary time. MT2_Mm_iv and v, though, seemed to have ceased expanding (**Figure 32**). Genomic expansion of MT2_Mm started with MT2_Mm_i, followed by that of MT2_Mm_ii, iv, v and finally iii (**Figure 32**).

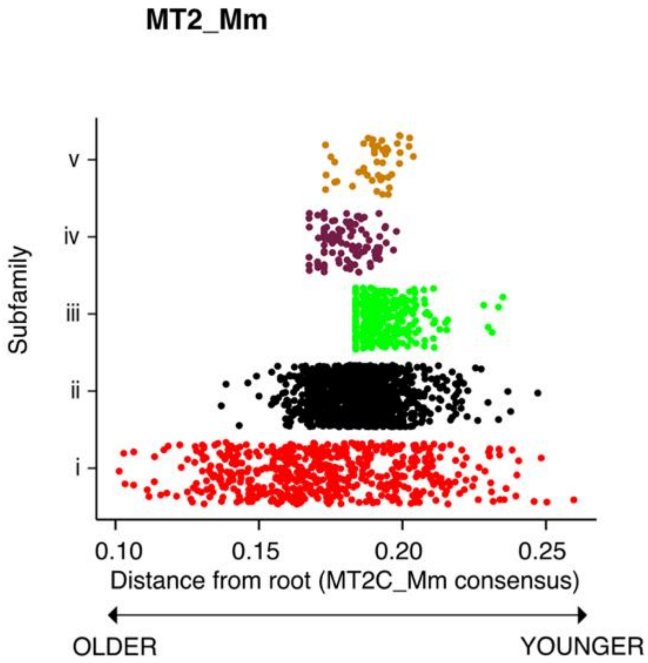


Figure 32. Divergence analysis of MT2_Mm insertions. Each dot on the plot is an individual MT2_Mm insertion, organized by subfamily. The position on the x axis is the genetic distance to the root of the phylogenetic tree (MT2C_Mm consensus sequence).

These substantial variation in subfamily spread over evolutionary time suggests differing colonization efficiencies. Interestingly, as opposed to what was shown and suggested with other TEs, we found no link between age and expression. Indeed, when plotting expression against age, we observed variation in insertion expression independent of their age (**Figure 33**). The evolutionary pattern that we describe for MT2_Mm contrasts from the sequential waves of expansions observed for L1 (Castro-Diaz et al. 2014). This suggests distinct evolutionary forces at play, whereby MT2_Mm_i may have experienced selective benefits that allowed it to keep expanding in parallel with other MT2_Mm subfamilies.

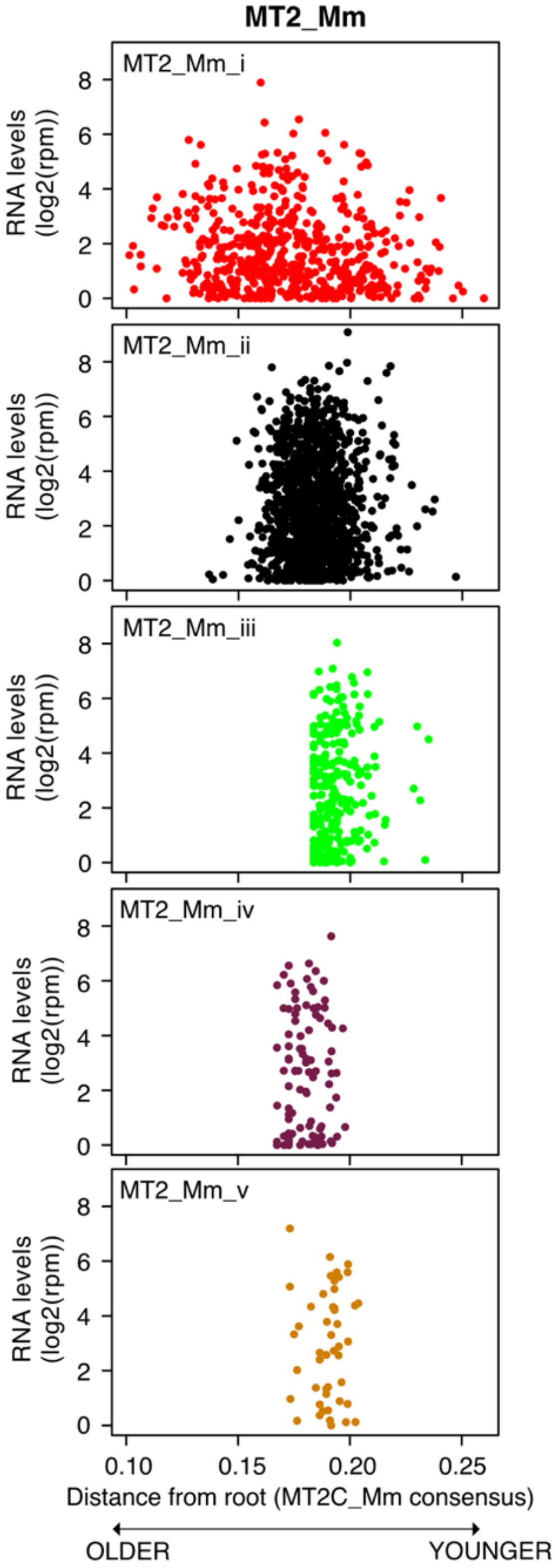


Figure 33. Expression against genetic divergence on single MT2_Mm insertions. Each dot on the plots is an individual MT2_Mm insertion. The position on the x axis corresponds to the genetic distance to the root (MT2C_Mm consensus sequence), while the position on the y axis is the log2 transformed mean rpm value of a single insertion in all late 2-cell stage embryos.

These observations prompted us to investigate further the evolution of the MERVL LTR and explore the closest ancestor of MT2_Mm: MT2C_Mm. MT2C_Mm is present only as solo LTRs in the mouse genome, likely due to its evolutionary older age, which may have contributed to the loss of internal sequences over time. We observed a similar pattern with MT2_Mm, where the oldest subfamily, MT2_Mm_i, is primarily composed of solo LTRs (80%), while in the younger subfamilies, solo LTRs comprise no more than 30% of insertions (**data not shown**). MT2C_Mm is transcribed during preimplantation development, with an induction of transcription also characteristic of activation upon major ZGA (**Figure 34A**). Using the same pipeline, we then conducted a size distribution analysis, which revealed the presence of a higher amount of truncated LTRs than observed for MT2_Mm, which may be attributed to the evolutionary age of MT2C_Mm (**Figure 34B**). We selected sequences between 385 and 565bp in size, resulting in a total of 1293 MT2C_Mm insertions (**Figure 34B**).

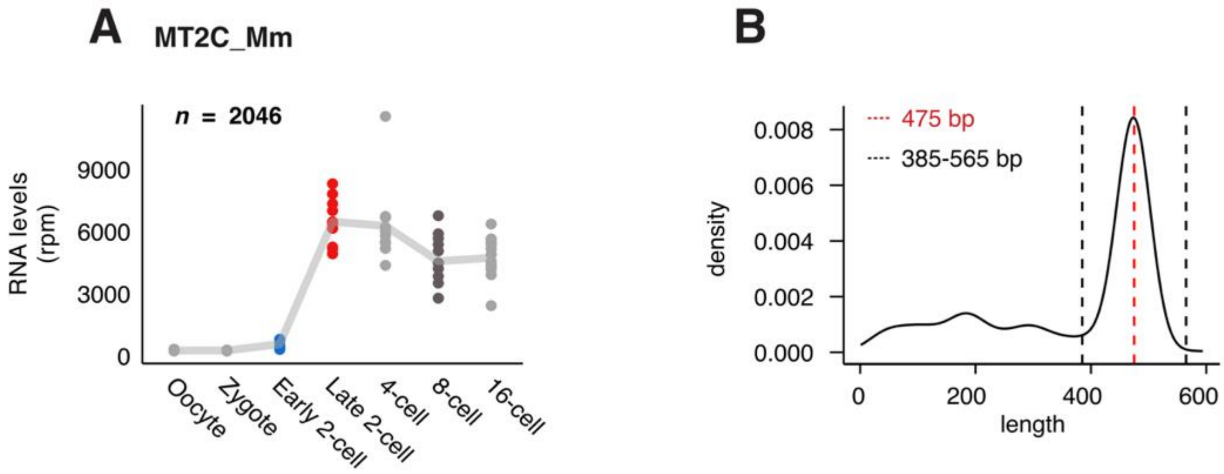


Figure 34. Expression pattern of MT2C_Mm across preimplantation development (A) and size density distribution plot of MT2C_Mm LTRs (B). (A) Data was reanalyzed from (Oomen et al. 2025). Each dot represents the sum rpm of all insertions (*n*) belonging to MT2C_Mm family per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages. (B) The consensus length is displayed in red dashed line. The selected size range in indicated in black dashed lines.

Phylogenetic tree reconstruction of MT2C_Mm insertions revealed 9 subfamilies. Importantly, in that case, the criteria used for subfamily definition were the same as the ones used for MT2_Mm, to enable comparisons (**Figure 35A**). The 9 obtained clades were clearly defined, and contained relatively homogenous number of insertions, from 45 in the smallest to 206 in the biggest subfamily (**Figure 35A**).

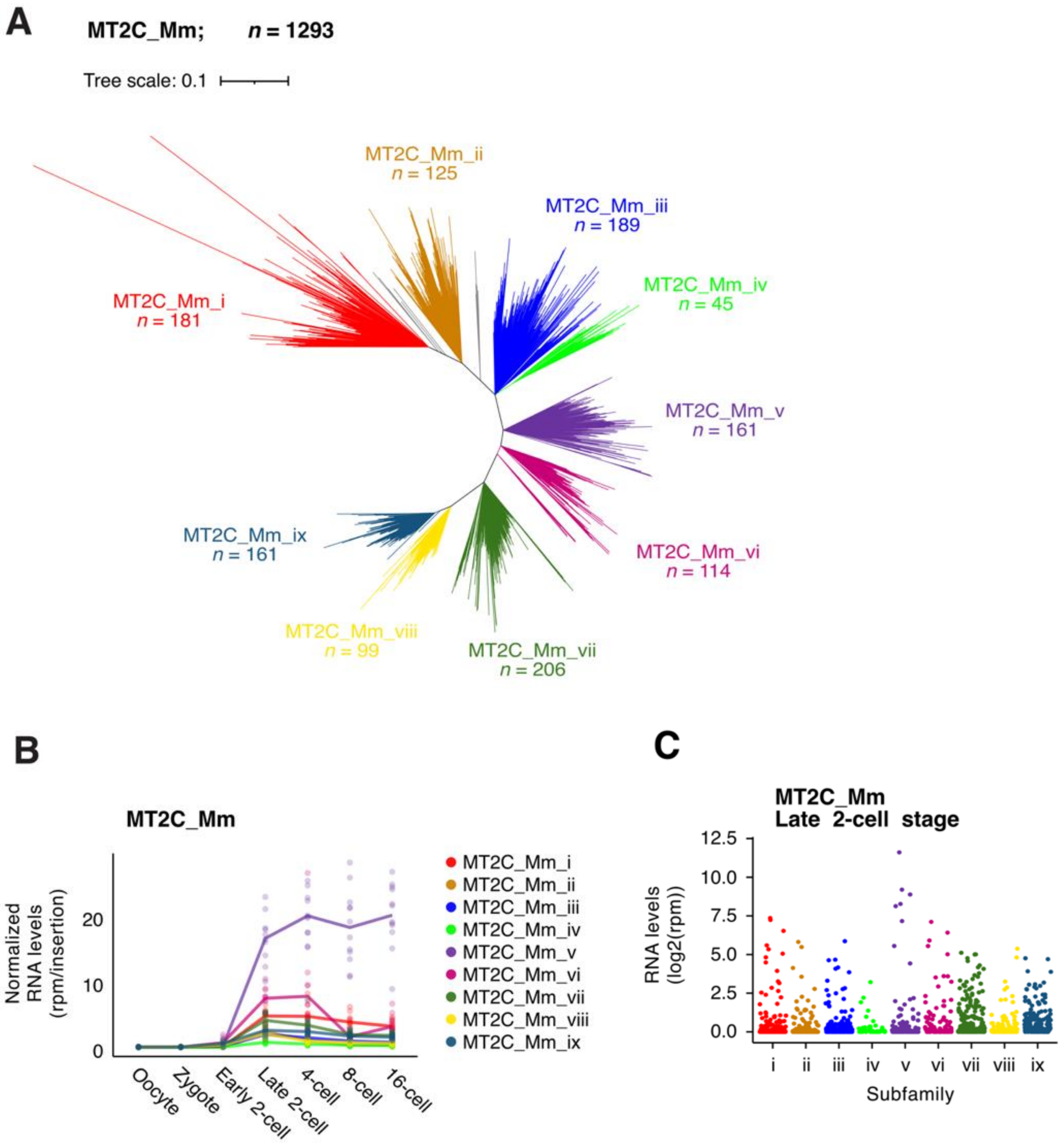


Figure 35. Phylogenetic analysis of MT2C_Mm. (A) Unrooted phylogenetic tree of MT2C_Mm insertions ranging between 385 and 565bp in length. *n* is the number of sequences per group indicated. The sequences in gray correspond to outgroups which did not qualify as subfamily (total 12 insertions). The tree scale is indicated. (B) Expression of all MT2C_Mm subfamilies across the different stages of preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot is the mean rpm of all insertions belonging to the indicated subfamily (rpm/insertion number) per single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages. (C) Expression of single MT2C_Mm insertions at the late 2-cell stage. Data was reanalyzed from (Oomen et al. 2025). Each point is the log2 transformed mean rpm value of a single insertion in all late 2-cell stage embryos.

Most of MT2C_Mm transcripts present at the late 2-cell stage derive from MT2C_Mm_v (**Figure 35B**). In fact, a single insertion was responsible for most MT2C_Mm_v transcripts in late 2-cell stage embryos (**Figure 35C**, *purple dot at log2(rpm) ~ 12*). All subfamilies contained insertions that were expressed at the late 2-cell stage (**Figure 35C**) and the patterns of expression during development were characterized by an increase in transcription upon ZGA for all subfamilies (**Figure 35B**). However, the transcript abundance differed between subfamilies, and some subfamilies remained expressed throughout preimplantation development while others were specifically only transcribed at the late 2-cell stage (**Figure 35B**). To pursue an evolutionary perspective in the characterization of MT2C_Mm subfamilies as an ancestor of MT2_Mm, we looked at TFBS within single insertions using the same TF list as for MT2_Mm. Many TFs exhibited scattered motifs across MT2C_Mm insertions, lacking distinct patterns (**Figure 36**). Unexpectedly, DUX binding sites were found in subfamilies viii and ix (**Figure 36**), indicating that these sites were already present in MT2C_Mm, which predates MT2_Mm. In addition, SRF binding sites were present throughout all MT2C_Mm insertions but became prominent in subfamilies vi to ix (**Figure 36**).

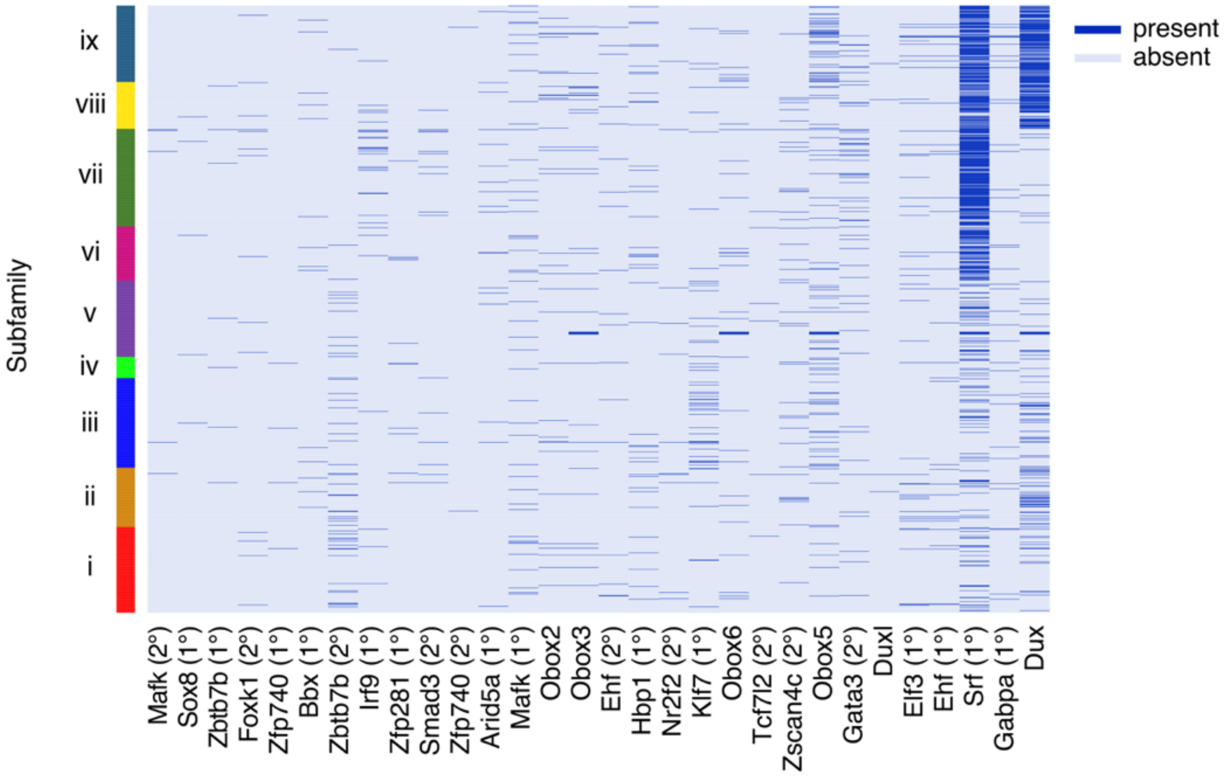


Figure 36. Heatmap showing the presence or absence of TFBS in MT2C_Mm insertions. Insertions are ordered by subfamily indicated by the bar legend on the left, and TFs are arranged in the same order as for MT2_Mm heatmap in Figure 30. 1° and 2° refer to “primary” and “secondary” binding sites.

This suggested that the binding sites for both DUX and SRF appeared prior to MT2_Mm expansion. To elaborate on the evolutionary pressures and the colonization strategies in relationship with TFBS, we computed the genetic divergence of MT2C_Mm insertions, using the same method as we used for MT2_Mm. In that case, the tree was rooted with MT2B consensus sequence, MT2C_Mm closest evolutionary predecessor (**Figure 31**).

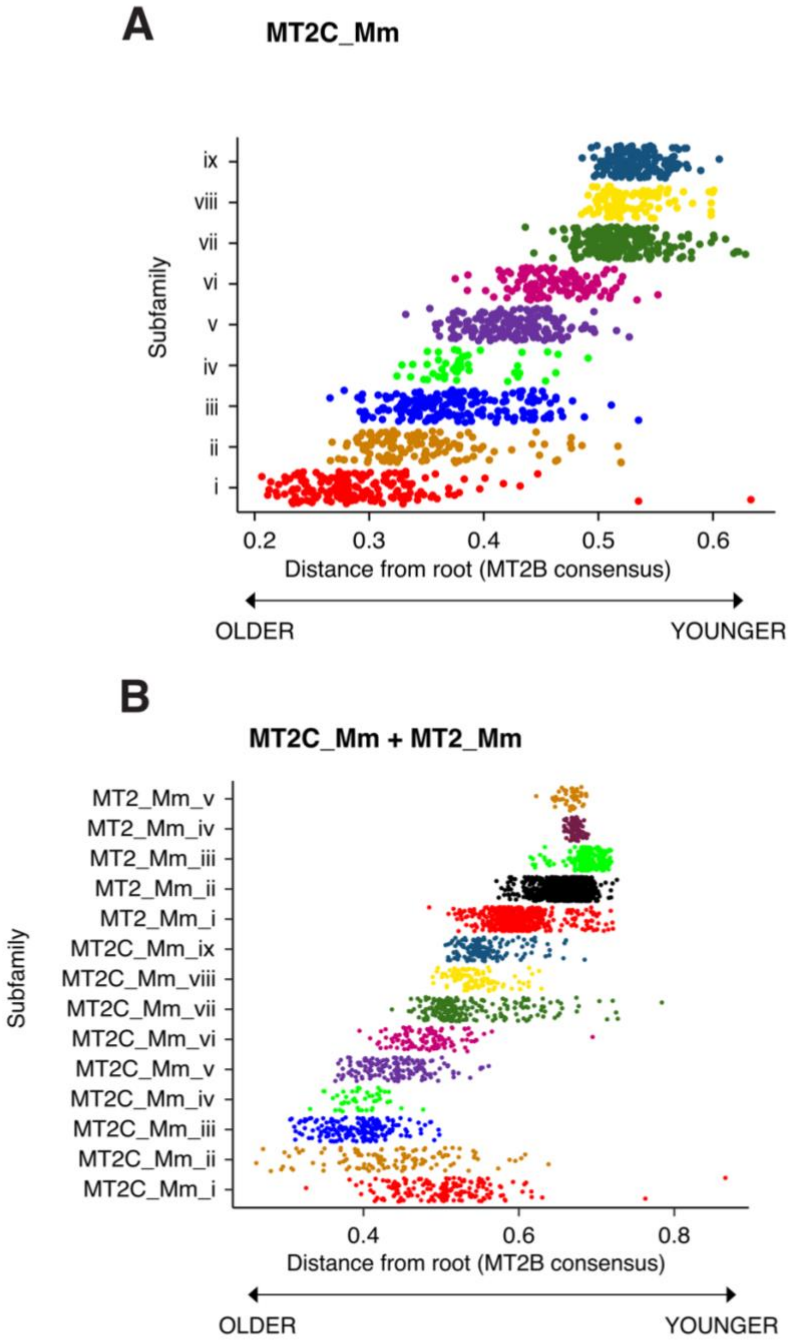


Figure 37. Divergence analysis of MT2C_Mm insertions (A) and combined MT2_Mm and MT2C_Mm insertions (B). (A, B) Each dot on the plots is an individual MT2C_Mm insertion (A) or MT2C_Mm or MT2_Mm insertion (B), organized by subfamily they belong to. The position on the x axis is the genetic distance to the root of the phylogenetic tree (MT2B consensus sequence).

The pattern of expansion of MT2C_Mm was completely different to that of MT2_Mm. We observed an evolution in sequential waves of expansion, more similar to what was observed for L1 (Castro-Diaz et al. 2014), from MT2C_Mm_i, the oldest subfamily, all the way to MT2C_Mm_ix (**Figure 37A**). To reconstruct the complete evolutionary history, we merged all MT2_Mm and MT2C_Mm insertions into a single tree and calculated the genetic divergence of each insertion (**Figure 37B**). Since genetic distance is relative, the results combining all insertions showed noticeable differences from MT2_Mm and MT2C_Mm insertions separately (**Figure 37A, B**). Specifically, incorporating MT2_Mm into the MT2C_Mm phylogenetic tree altered its structure, causing older MT2C_Mm insertions to cluster separately from MT2_Mm. On the combined tree, it appears that MT2C_Mm_i and MT2C_Mm_ii come from the same common ancestor, but MT2C_Mm_i goes out of the lineage, and the expansion of MT2C_Mm starts with MT2C_Mm_ii all the way to MT2C_Mm_ix eventually leading to the appearance of MT2_Mm (**Figure 37B**). Combining this information with the TFBS profiling shown in Figures 30 and 36, these data confirm that the well-described DUX binding site arose prior to MT2_Mm expansion, within the youngest MT2C_Mm subfamilies and was closely followed by the appearance of MT2_Mm_i. Furthermore, these results imply that the emergence of the SRF binding site predates that of DUX in the evolution of MT2C_Mm.

Analyzing individual insertions provides valuable insight into the actual sequences present in the genome, but it can become cumbersome for further analysis and deeper characterization of sequence features. To address this, we turned to consensus sequences. We generated new majority-rule consensus sequences for each of the five newly identified MT2_Mm subfamilies and the nine MT2C_Mm subfamilies and aligned the obtained 14 consensus sequences together (**Figure 38**). We identified a notable 9bp deletion that first appeared in MT2C_Mm_viii leading to the creation of the DUX binding motif (**Figure 38**, *purple*). We also observed two SNPs, a C to T leading to a predicted higher affinity binding motif for SRF at the transition from MT2C_Mm_vi to vii, and a T to A leading to a decrease in predicted affinity from MT2_Mm_i to ii (**Figure 38**, *cyan*). Finally, an A to G mutation between MT2_Mm_i and MT2_Mm_ii led to the emergence of a GABPA binding motif present in all in of the youngest MT2_Mm subfamilies (**Figure 38**, *orange*).

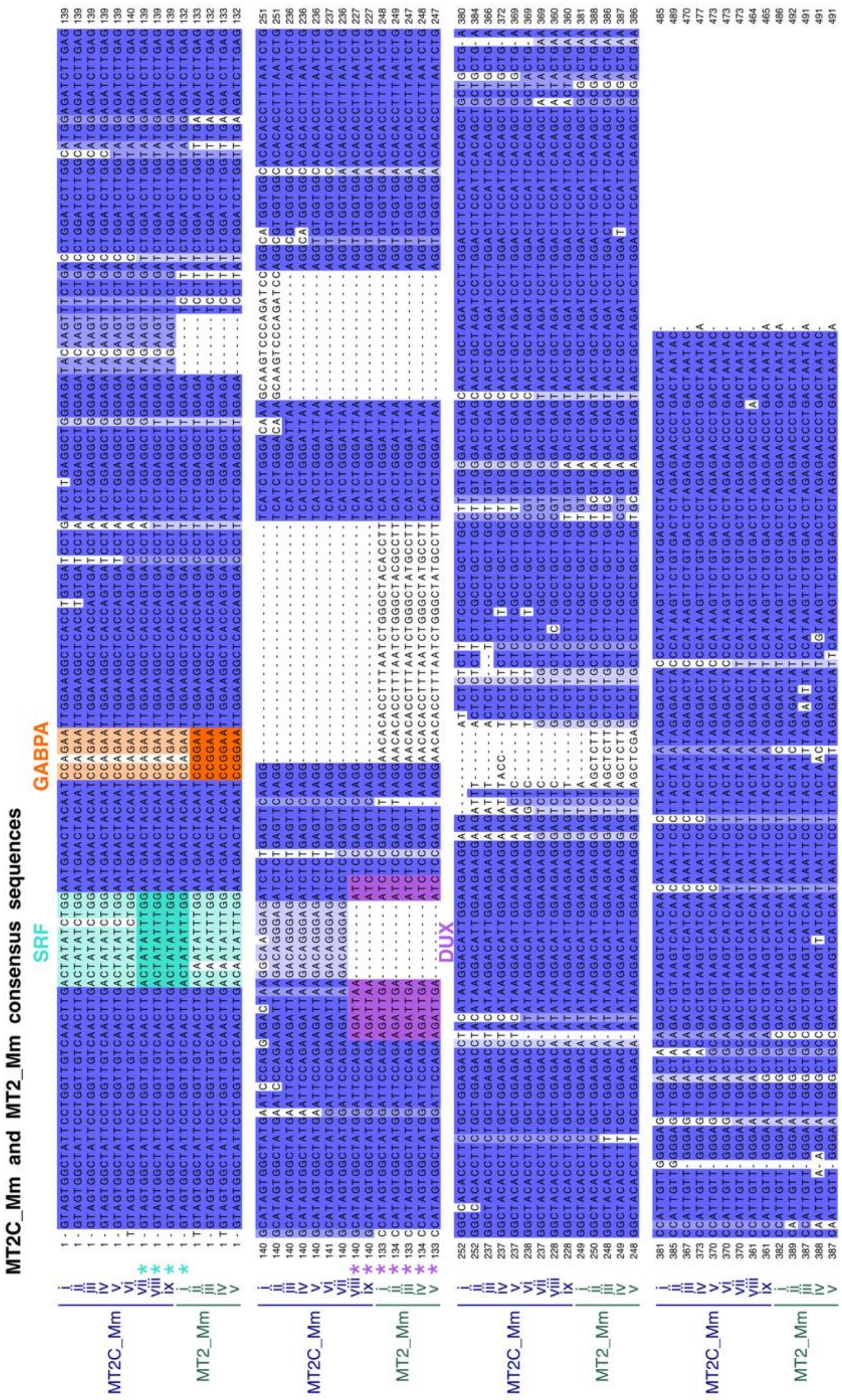


Figure 38. Multiple sequence alignment of consensus sequences from MT2_Mm and MT2C_Mm subfamilies. The positions of binding motifs for SRF, GABPA and DUX are highlighted in cyan, orange and purple, respectively. The SRF or GABPA higher confidence predicted binding site are indicated in darker cyan or orange, respectively, and the SNPs leading to lower-confidence prediction are in white. MT2_Mm subfamilies are displayed in dark green, MT2C_Mm in dark blue.

Using these consensus sequences, we conducted a median-joining network analysis, a method that estimates the most parsimonious path between sequences. To give this network analysis an evolutionary orientation, we added MT2B consensus sequence (**Figure 39**). Through this analysis, we reaffirmed the findings of the divergence analysis, revealing that MT2 evolution from MT2B initiated with a shared ancestor of MT2C_Mm_i and ii, and expanded progressively from MT2C_Mm_ii until the emergence of MT2_Mm_iii, the youngest subfamily (**Figure 39**). Within this evolutionary progression, we pinpointed the emergence of the DUX and GABPA binding sites, both of which appeared later in the evolutionary timeline than the increase in confidence prediction of the SRF binding site (**Figure 39**).

Therefore, we successfully reconstructed the full evolutionary history of MT2C_Mm and MT2_Mm, identifying the timing and emergence of the binding site for the well-known **DUX** TF. We discovered **GABPA** as a potential novel regulator of MT2_Mm, correlating with increased MT2_Mm subfamily expression. Furthermore, the emergence of the **SRF** binding site before DUX suggests SRF as a previously unidentified regulator of MT2 lineages.

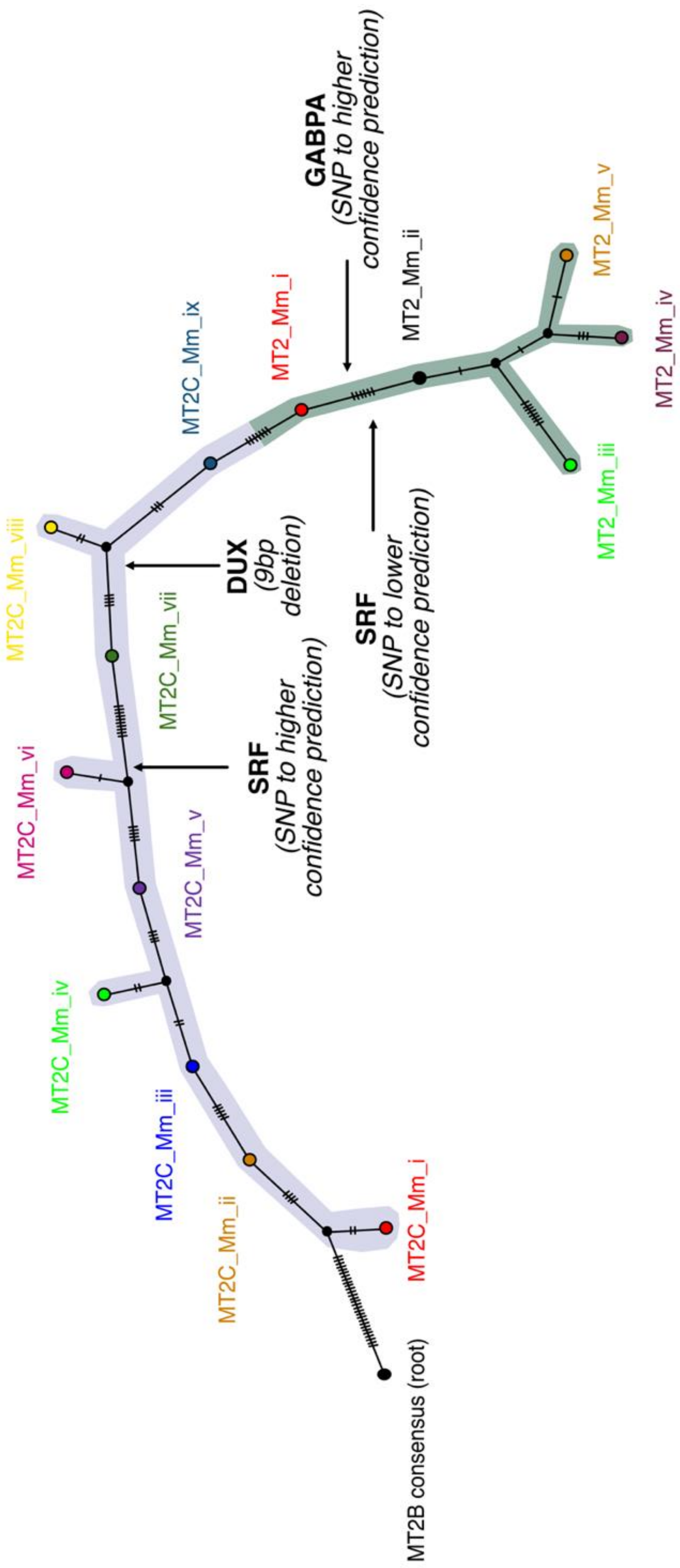


Figure 39. Median-joining network analysis of MT2C_Mm and MT2_Mm subfamily consensus sequences. The evolutionary orientation was inferred by rooting with the MT2B consensus sequence. The number of ticks represents the number of mutations at nongaps. The mutations leading to binding motifs for TFs are indicated. MT2_Mm subfamilies are highlighted in dark green, MT2C_Mm in dark blue.

4.3 Investigating TFs functions in regulating TEs in cell culture models

4.3.1 Targeted gain-of-function screen for TFs activating TE transcription in mESCs

Alongside the detailed phylogenetic analysis described in **section 4.2**, I worked to directly assess whether the TFs from **section 4.1** could drive TE transcription in mESCs. I selected a gain-of-function (GOF) screen in mESCs to test a larger set of TFs for their ability to activate TEs endogenously, which is not feasible at the same scale in embryos. Out of the 54 potential candidates identified in **section 4.1**, I chose to focus on 40 TFs. These 40 TFs were selected based on their expression profiles during preimplantation development, the quality of their motif match to the extracted footprint sequences, and relevant literature review. In addition, most of the TFs showed low expression in ESCs, except for *Gabpa* and *Atf1* (**Figure 40**).

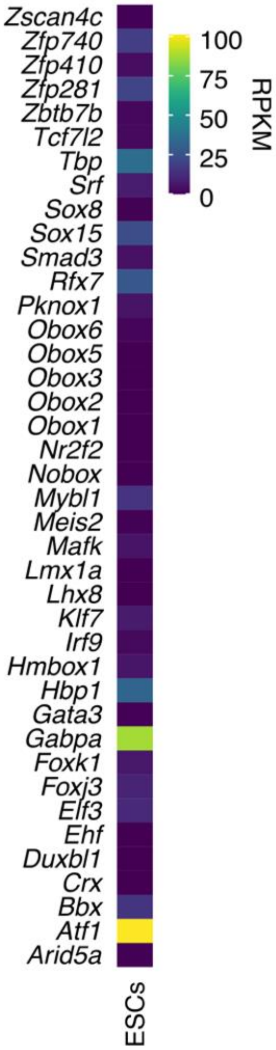


Figure 40. Expression of candidate TFs in mESCs. Heatmap showing the expression of the 40 candidates chosen for overexpression in mESCs. TFs are ordered by reverse alphabetical order. The values are rpkm.

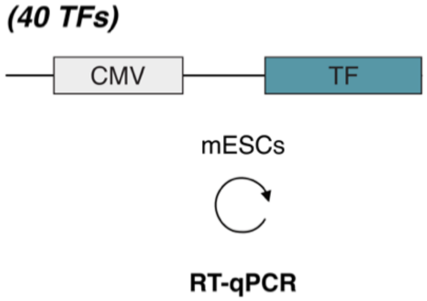


Figure 41. Schematic representation of the experimental design of the targeted gain-of-function screen strategy. The circular arrow refers to the transfection process in mESCs.

I cloned all 40 TFs in a vector containing a mammalian expression promoter and transfected each TF individually in mESCs (**Figure 41**). By quantitative real-time PCR (qPCR) in three technical replicates for two biological replicates, I measured TE expression. I analyzed the expression of IAPEZ, LINE1 (ORF1 and ORF2), MaLR MT, MaLR ORR (LTR and Int), MERVL (LTR and Int) and SINEB2 using qPCR primers. Except for MaLR ORRs, all primer sets were previously published and validated for specificity (Peaston et al. 2004; Fadloun et al. 2013; Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018; Zhang et al. 2019a). ORR1A0 and ORR1A1 LTRs were difficult to distinguish by qPCR, therefore I used a primer set amplifying both LTRs (MaLR_ORR_LTR). The "_int" primer set, though, specifically amplified ORR1A1.

The most reliable method for activating TE expression in mESCs is through TF overexpression, even though this approach does not accurately reflect the endogenous activity of TFs. Previous studies have shown that this approach effectively induces TE expression, such as DUX overexpression activating MERVL and its LTR in mESCs (Hendrickson et al. 2017). Therefore, I used DUX as a positive control to assess the reliability of this experimental design. DUX overexpression resulted in distinct, robust and specific upregulation of MERVL_int (both primer sets) and MT2_Mm (**Figure 42A**). In addition, I verified the overexpression efficiency by quantifying *Dux* RNA levels in the same samples (**Figure 42B**), validating the experimental design.

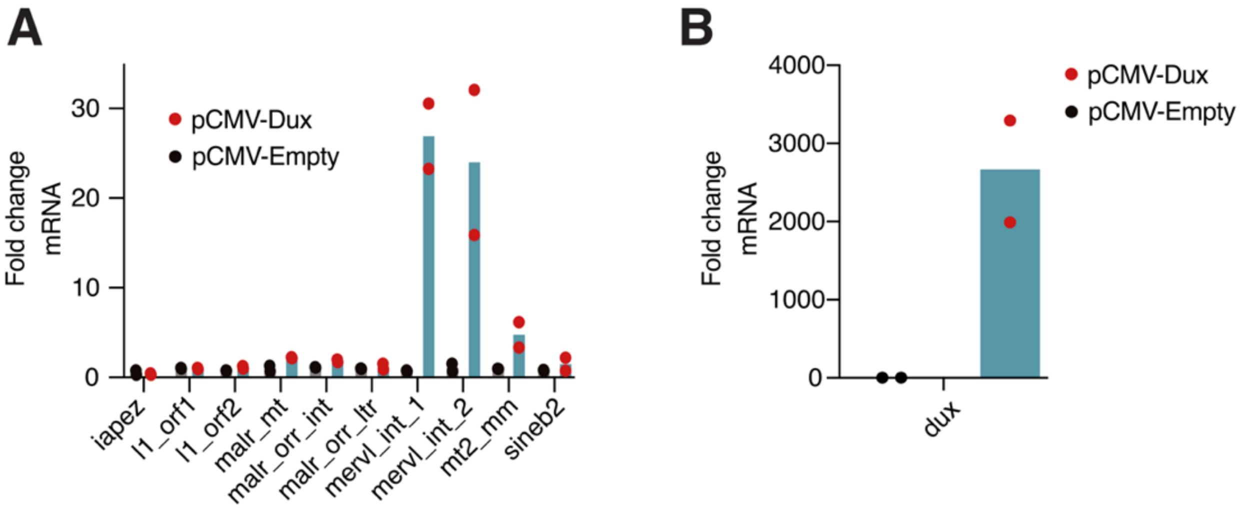


Figure 42. qPCR analysis of DUX overexpression in mESCs. (A, B) Fold change of the indicated TE families of interest RNA levels (A) and Dux mRNA levels (B) over no vector control by qPCR. The bar represents the mean over two biological replicates, the black and red dots represent the mean over three technical replicates for each biological replicate of transfected cells with either pCMV-Empty or pCMV-Dux, respectively.

Upon overexpression in mESCs, several TFs triggered the transcription of TEs, with some affecting multiple TE classes simultaneously (**Figure 43**). **FOXJ3**, **RFX7**, **TBP** and **SRF** were among the most prominent activators of multiple classes (**Figure 43**). FOXJ3 overexpression induced IAPEZ, LINE1, MERVL and SINEB2. RFX7 triggered IAPEZ, MaLR, MERVL and SINEB2 (**Figure 43**). Further, a broad range of TEs was activated by SRF and TBP, two conserved and ubiquitously expressed TFs. SRF, in particular, induced MERVL, MT2_Mm, L1 (ORF2) and IAPEZ (**Figure 43**). The most notable impact of TBP was observed on MaLR ORR (**Figure 43**). More specific effects were observed for **SMAD3** and **CRX**, both primarily activating MERVL, while **OBOX6** was specific to MaLR ORR. **LMX1A** and **GATA3** displayed moderate, yet distinct, effects on MERVL (**Figure 43**). Additionally, **ZFP410** broadly but subtly activated most TE families (**Figure 43**).

These data provided preliminary evidence of TFs driving TE transcriptional activation in mESCs. None matched DUX in induction capacity, possibly because none of these TFs possess pioneering activity, as seen with DUX (Eidahl et al. 2016). While indirect effects from TF overexpression cannot be ruled out, this targeted GOF screen identified **10 potential TFs** as promising candidates for further investigation.

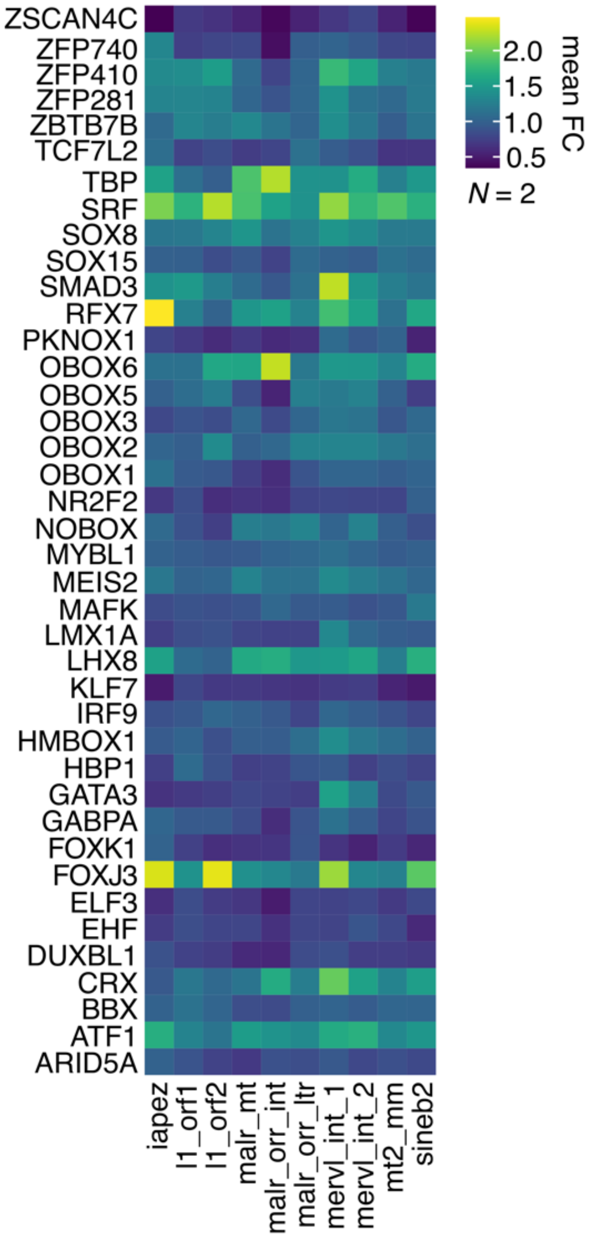


Figure 43. qPCR analysis of overexpression of 40 TFs in mESCs. Heatmap showing the average fold change of two biological replicates of the indicated TE families upon overexpression of each TF over no vector control by qPCR.

4.3.2 SRF activates MT2_Mm and MT2C_Mm in reporter assays

In addition to the targeted screen in mESCs and following the detailed phylogenetic analysis of MT2_Mm and MT2C_Mm in **section 4.2.4**, I sought to assess the ability of SRF and GABPA to drive the expression of these elements. I also investigated the validity of the DUX binding site prediction and whether DUX motif indeed emerged prior to MT2_Mm expansion. I leveraged the consensus sequences established in **section 4.2.4** in a heterologous system based on luciferase reporter constructs featuring MT2C_Mm or MT2_Mm subfamilies as promoters. These experiments were conducted in HEK293 cells to minimize potential confounding effects.

Given that the GABPA binding site was present in all MT2_Mm subfamilies except MT2_Mm_i (**Figure 38**), and was associated with higher expression at the 2-cell stage (**Figure 29B**), I first cloned MT2_Mm_i and MT2_Mm_ii into a luciferase vector (**Figure 44A**). I then co-transfected these vectors with GABPA expression vector to assess GABPA's ability to induce LTR transcription (**Figure 44A**). Since GABPA requires its obligatory partner, GABPB (Rosmarin et al. 2004), I co-transfected the cells with both TFs (**Figure 44A**).

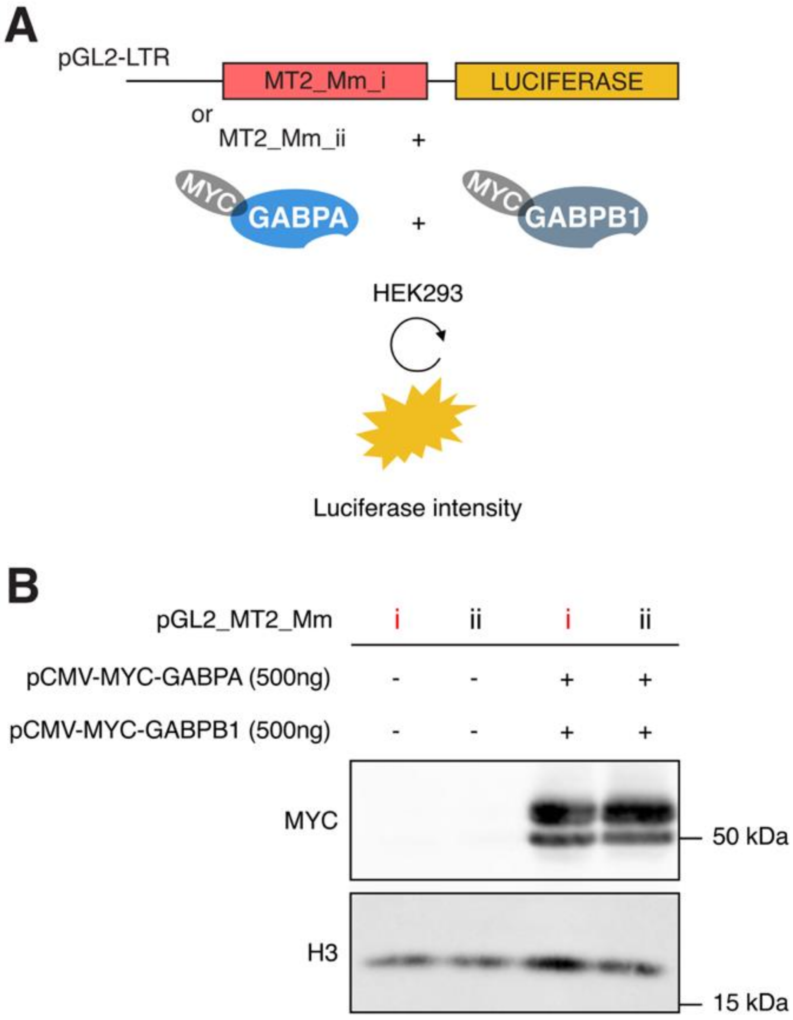


Figure 44. Setting up luciferase assay to test GABPA/B1 capacity to activate MT2_Mm. (A) Schematic representation of the experimental design used to test the capacity of GABPA/GABPB1 to drive transcription of different MT2_Mm subfamilies. (B) Western blot analysis of GABPA and GABPB1 upon overexpression of the MYC-tagged protein. *N* = 1.

I confirmed the overexpression efficiency by western blot for MYC-tagged GABPA and GABPB1 (**Figure 44B**). However, no increase in luciferase was observed upon GABPA/B1 overexpression for either MT2_Mm_i or MT2_Mm_ii (**Figure 45**). I concluded that, in this context, GABPA/B1 did not induce MT2_Mm transcription and decided to discontinue further investigation of this TF.

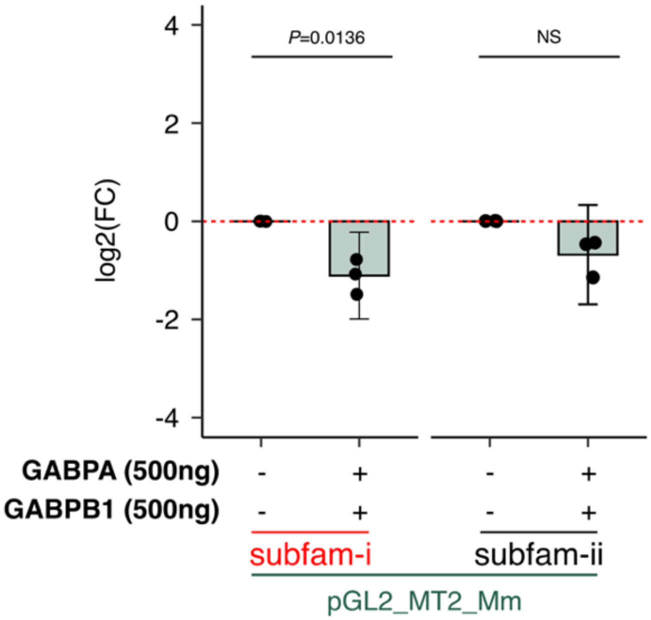


Figure 45. Luciferase activity of MT2_Mm in presence of GABPA/B1. Log2 transformed fold change of the normalized luciferase activity of MT2_Mm_i and MT2_Mm_ii in the presence of 500ng of pCMV-encoding GABPA and GABPB1 coding sequences over control (pCMV-empty). Bar plots show the mean +/- 95% confidence intervals (*N* = 3). For statistical analysis, preselected hypotheses compared each group to 0.

For SRF and DUX, I selected four subfamilies to clone into the luciferase vector. Based on the evolutionary trajectory of MT2C_Mm and MT2_Mm and the relationships with TFBS (**Figure 39**), I cloned MT2C_Mm_vii, which contains only the high-confidence predicted SRF binding motif, and MT2C_Mm_ix, which includes both the SRF motif and a deletion event that created the DUX binding motif. Additionally, I used the two main MT2_Mm subfamilies, previously cloned for GABPA/B1. MT2_Mm_i, like MT2C_Mm_ix, contains both binding sites, whereas MT2_Mm_ii lacks the high-confidence SRF motif because of a SNP (**Figures 38 and 46A**). The efficiency of TF overexpression was verified by western blot for the MYC tag (**Figure 46B**).

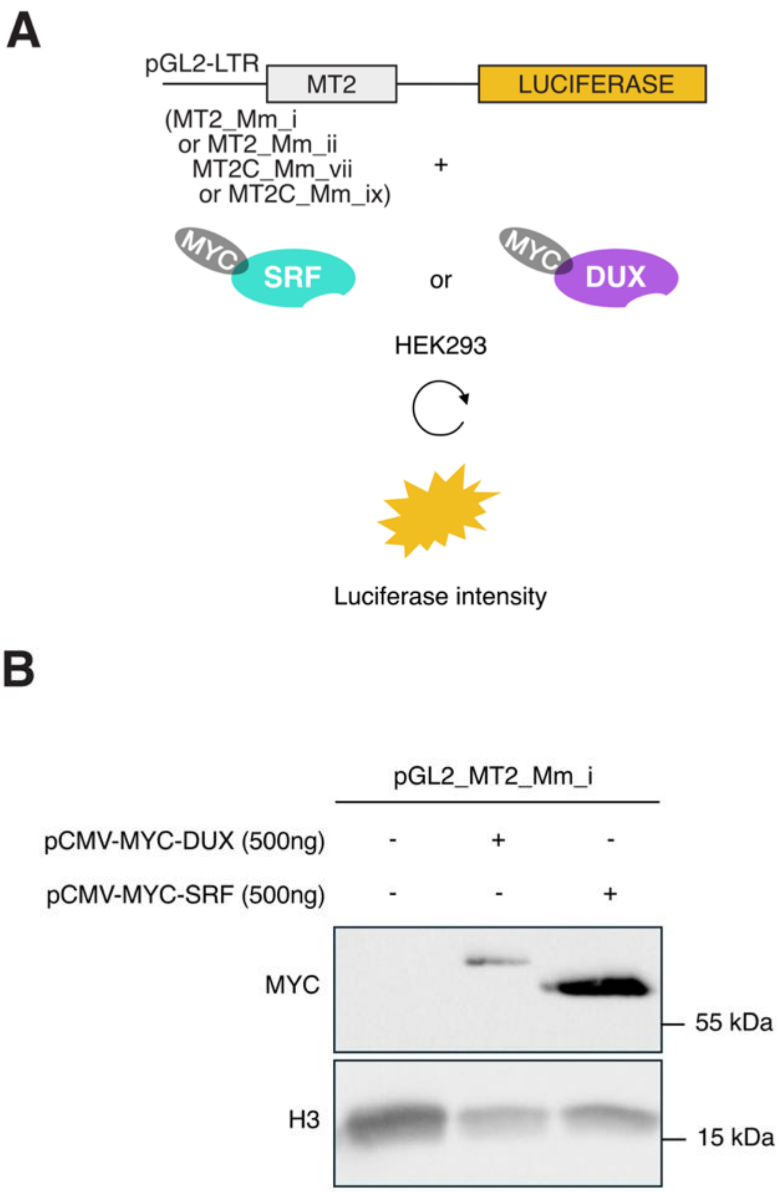


Figure 46. Setting up luciferase assay to test SRF and DUX capacity to activate different MT2C_Mm and MT2_Mm subfamilies. (A) Schematic representation of the experimental design decided to test the capacity of DUX and SRF to drive transcription of different MT2C_Mm and MT2_Mm subfamilies. (B) Western blot analysis of DUX and SRF upon overexpression of the MYC-tagged protein. *N* = 2.

Co-transfection of the SRF encoding vector with MT2C_Mm-driven luciferase constructs resulted in transactivation of both MT2C_Mm_vii and MT2C_Mm_ix at similar levels (**Figure 47**). It indicated that SRF can transactivate these subfamilies in a luciferase reporter assay, in cells. Notably, SRF also activated the expression of both MT2_Mm subfamilies. Subfamilies containing the higher-confidence predicted binding motif appear more responsive to SRF, as MT2_Mm_i, along with MT2C_Mm subfamilies, reached saturation at lower SRF concentrations than MT2_Mm_ii, in agreement with the TFBS analysis (**Figure 47**).

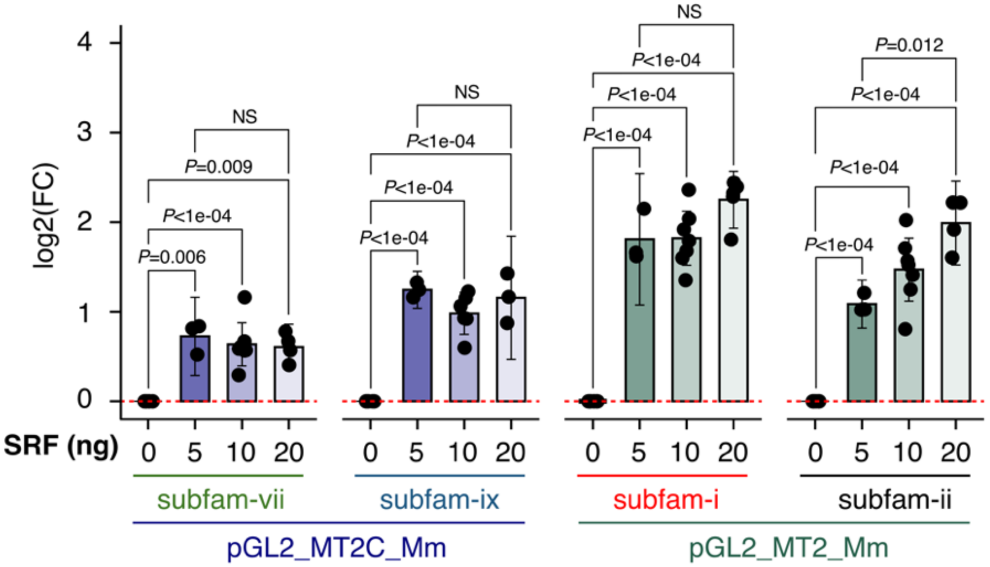


Figure 47. Luciferase activity of MT2_Mm and MT2C_Mm subfamilies in presence of SRF. Log2 transformed fold change of the normalized luciferase activity of MT2C_Mm_vii, MT2C_Mm_ix, MT2_Mm_i and MT2_Mm_ii, in the absence or presence of 5, 10 or 20 ng of pCMV-encoding SRF coding sequence over control (pCMV-empty). Bar plots show the mean +/- 95% confidence intervals (*N* ≥ 3). For statistical analysis, preselected hypotheses compared each group to 0, and 20 ng was compared to 5 ng for each subfamily.

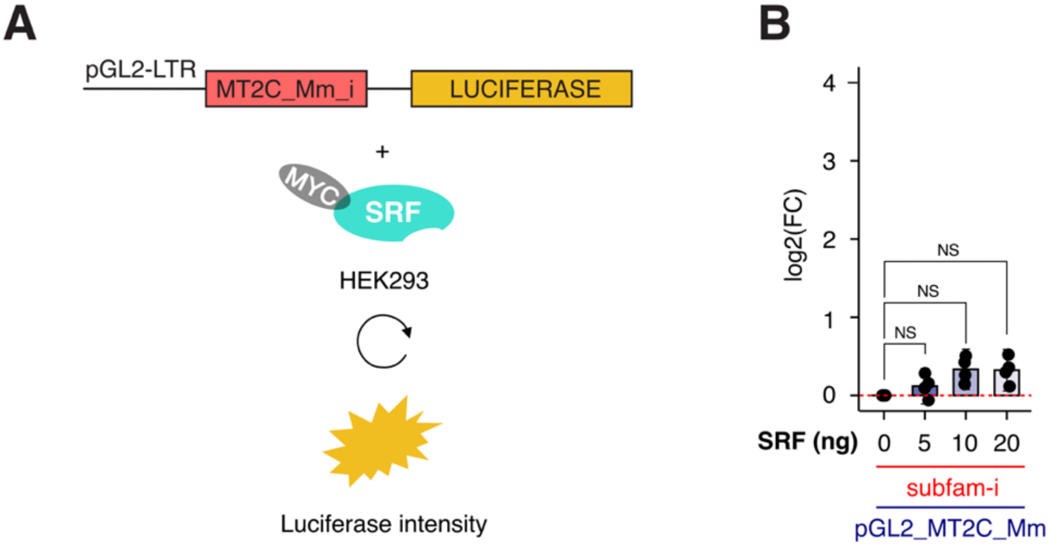


Figure 48. Luciferase activity of MT2C_Mm_i in presence of SRF. (A) Schematic representation of the experimental design decided to test the capacity of SRF to drive transcription MT2C_Mm_i. (B) Log2 transformed fold change of the normalized luciferase activity of MT2C_Mm_i in the absence or presence of 5, 10 or 20 ng of pCMV-encoding SRF coding sequence over control (pCMV-empty). Bar plots show the mean +/- 95% confidence intervals (*N* ≥ 3). For statistical analysis, preselected hypotheses compared each group to 0.

Importantly, an older MT2C_Mm subfamily which did not contain the higher-confidence predicted SRF motif (MT2C_Mm_i), showed no transcriptional activation upon SRF overexpression (**Figure 48**). These results are in line with the TFBS analysis, and suggest that the C to T mutation from MT2C_Mm_vi to vii plays a more significant role than the T to A from MT2_Mm_i to MT2_Mm_ii for SRF binding (**Figure 38**).

Finally, I observed transcriptional activation of both MT2C_Mm_vii and MT2C_Mm_ix upon co-transfection with the DUX encoding vector, though to varying degrees (**Figure 49**). Indeed, MT2C_Mm_ix was activated approximately 25 times more than MT2C_Mm_vii by DUX, suggesting that the 9bp deletion did create a functional DUX binding site (**Figure 49**). Furthermore, the transcriptional activation of MT2C_Mm_ix was comparable to that observed for MT2_Mm subfamilies, which are known to be activated by DUX (Hendrickson et al. 2017; Whiddon et al. 2017) (**Figure 49**). A difference in DUX-induced transcriptional activation strength between MT2_Mm_i and _ii was noted, correlating with differences in transcript abundance between MT2_Mm_i and _ii observed in embryos (**Figure 49**). These data suggest the presence of additional regulator(s) that bind to MT2_Mm_ii and interact with DUX to enhance transcriptional activation capacity.

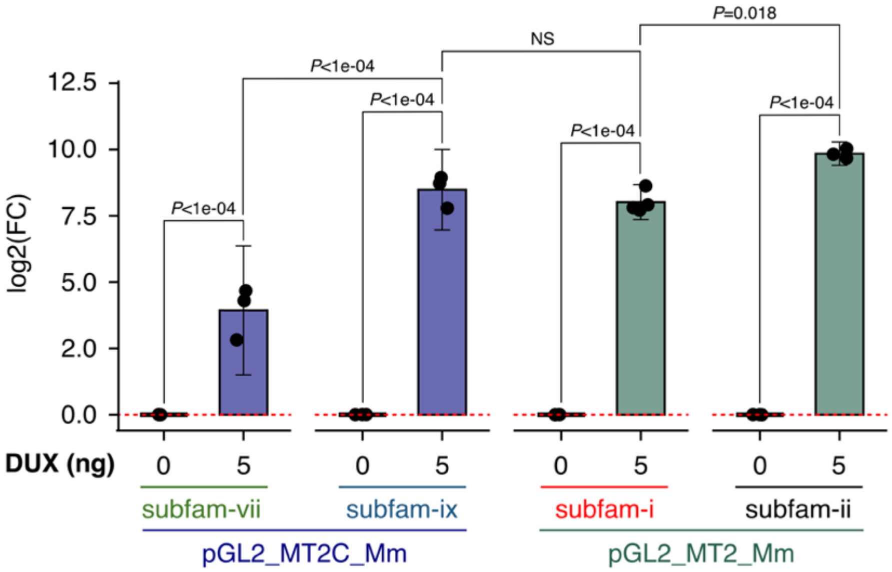


Figure 49. Luciferase activity of MT2_Mm and MT2C_Mm subfamilies in presence of DUX. Log2 transformed fold change of the normalized luciferase activity of MT2C_Mm_vii, MT2C_Mm_ix, MT2_Mm_i and MT2_Mm_ii, in the absence or presence of 5 ng of pCMV-encoding DUX coding sequence over control (pCMV-empty). Bar plots show the mean +/- 95% confidence intervals ($N \geq 3$). For statistical analysis, preselected hypotheses compared each group to 0 and log2(FC) with 5 ng of MT2_Mm_ii was compared to that of MT2_Mm_i, MT2_Mm_i to MT2C_Mm_ix and MT2C_Mm_ix to that of MT2C_Mm_vii.

We included controls using luciferase vectors lacking a promoter (pGL2_empty) and containing a scrambled LTR sequence (pGL2_scramble), which showed that neither SRF nor DUX overexpression led to transcriptional activation of either pGL2_empty or pGL2_scramble (**Figure 50**).

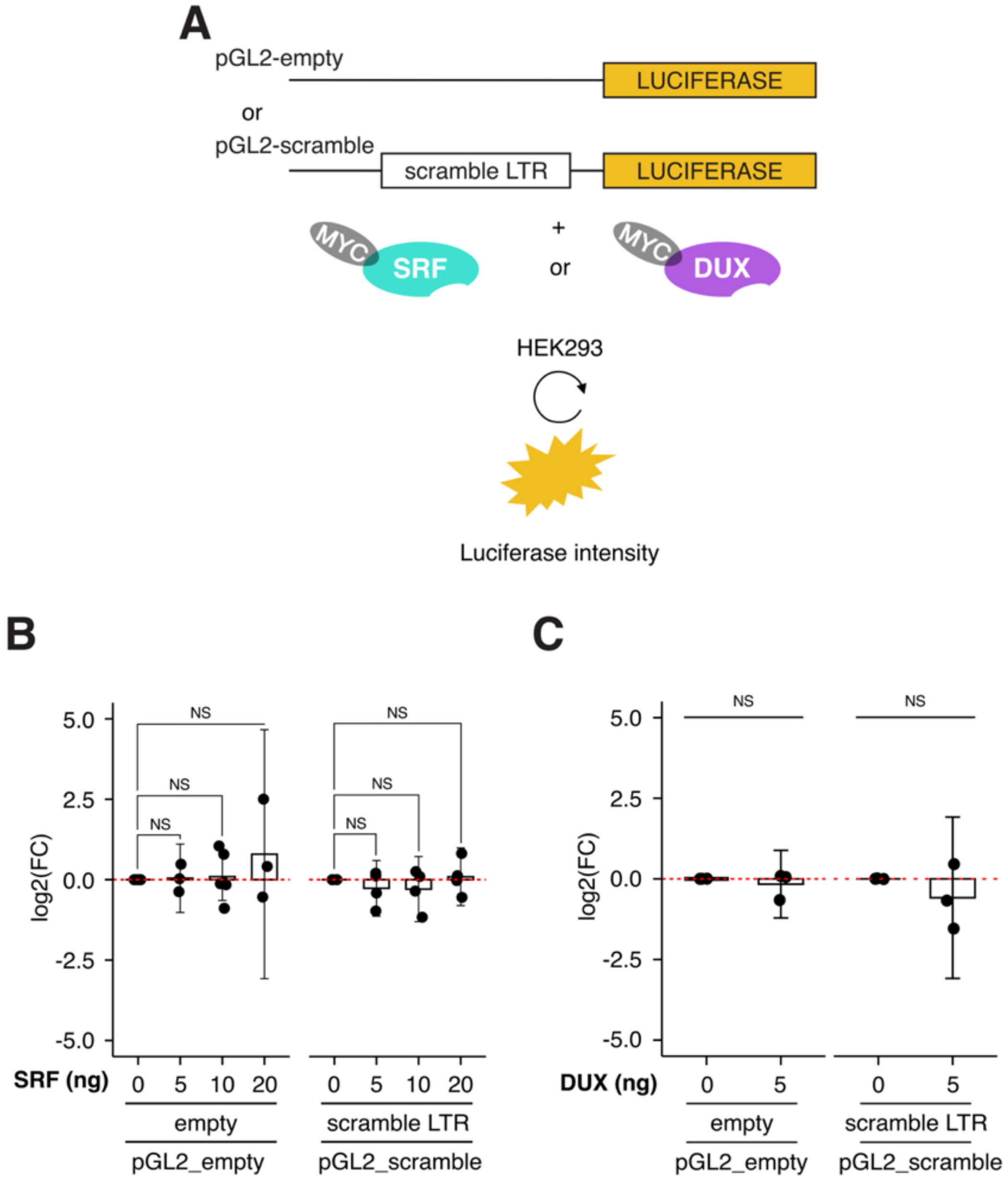


Figure 50. Luciferase activity of empty and scramble reporters in presence of DUX and SRF. (A) Schematic representation of the experimental design decided to test the capacity of DUX and SRF to drive transcription pGL2_empty and pGL2_scramble (B) Log2 transformed fold change of the normalized luciferase activity of pGL2_empty and pGL2_scramble in the absence or presence of 5, 10 or 20 ng of pCMV-encoding SRF coding sequence over control (pCMV-empty). Bar plots show the mean +/- 95% confidence intervals (*N* ≥ 3). For statistical analysis, preselected hypotheses compared each group to 0. (C) Log2 transformed fold change of the normalized luciferase activity of pGL2_empty and pGL2_scramble in the absence or presence of 5 ng of pCMV-encoding DUX coding sequence over control (pCMV-empty). Bar plots show the mean +/- 95% confidence intervals (*N* ≥ 3). For statistical analysis, preselected hypotheses compared each group to 0.

Altogether, these data confirmed the acquisition of a functional DUX motif through a 9bp deletion within MT2C_Mm, preceding MT2_Mm expansion, and established SRF as an ancient TF for MERVL, capable of activating the transcription of both MT2C_Mm and MT2_Mm in a cell culture model. These results were consistent with the observed impact of SRF on MERVL following transfection in mESCs (**Figure 43**). Based on the reporter assay data, GABPA was not pursued further. I therefore focused on the TFs identified in the targeted screen as potential TE activators in ESCs, with only SRF further validated in luciferase reporter assay. Of the 10 TFs highlighted in **section 4.3.1**, I chose to exclude only OBOX6. Although OBOX6 appeared to specifically affect MaLR ORR elements, I decided not to follow up on the OBOX family, as these factors have been extensively characterized elsewhere (Ji et al. 2023; Guo et al. 2024).

Thus, I focused on the remaining nine TFs to investigate their putative roles in preimplantation development: **FOXJ3, RFX7, TBP, SRF, SMAD3, CRX, LMX1A, GATA3 and ZFP410.**

4.4 A detailed expression profiling reveals that most candidate TFs are expressed in the embryo

With this refined selection of candidate TFs, I began investigating their putative role in preimplantation development by carefully characterizing their expression patterns at both the mRNA and protein levels. This provided valuable insights for subsequent manipulation of their expression, and uncovered new information, as most of these TFs have not been studied in the early embryo.

4.4.1 Expression profiles at the mRNA level

To investigate the mRNA expression levels of these TFs during preimplantation development, I used publicly available single embryo RNA-seq datasets from oocyte to the 16-cell stage (Oomen et al. 2025). While all TFs were expressed in early embryos, their expression patterns varied significantly. Based on these observations, I categorized the 9 TFs according to their expression profiles. Six of nine TFs; *Smad3*, *Rfx7*, *Foxj3*, *Tbp*, *Srf* and *Gata3* were present in the oocyte as maternal transcripts (**Figure 51**).

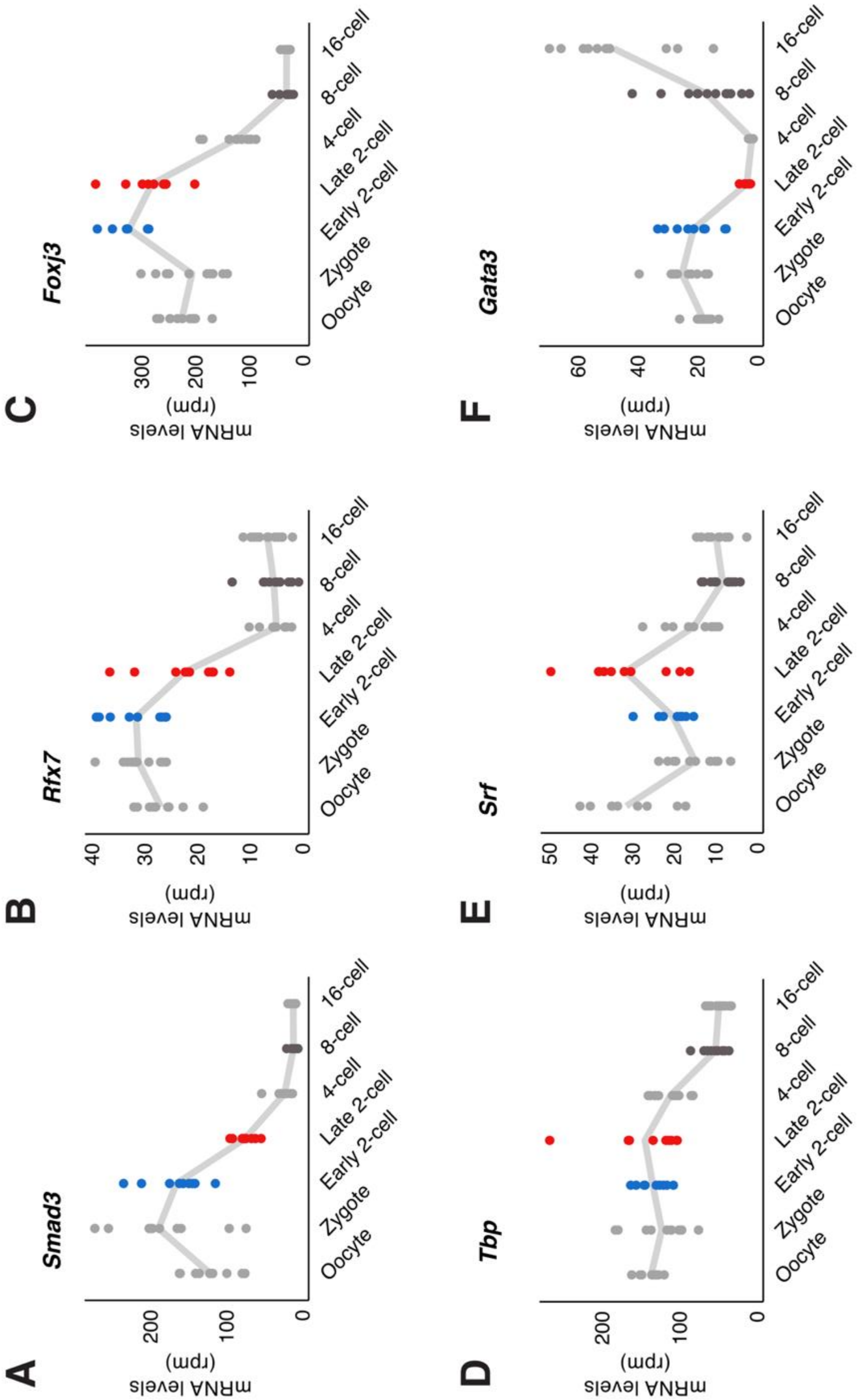


Figure 51. Expression profiles of maternally inherited TFs during preimplantation development. (A-F) Each dot represents *Smad3* (A), *Rfx7* (B), *Foxj3* (C), *Tbp* (D), *Srf* (E) and *Gata3* (E) mRNA levels (rpm) in individual embryos at the indicated stage. The trend line connects the mean values across embryos of individual stages.

Smad3 and *Rfx7* exhibited transient increases in mRNA levels in zygotes, becoming nearly undetectable by the 8-cell stage (**Figure 51A, B**). *Foxj3* showed a peak in expression at the early 2-cell stage, coinciding with minor ZGA, before rapidly diminishing to almost undetectable levels by the 8-cell stage (**Figure 51C**). In contrast, *Tbp* and *Srf* mRNA levels rose during the late 2-cell stage, coinciding with major ZGA, then quickly declined, but remained detectable throughout preimplantation development (**Figure 51D, E**). *Gata3* mRNA, after a modest increase during minor ZGA, dropped to near-zero levels by the 4-cell stage, before sharply rising to its peak at the 16-cell stage (**Figure 51F**). As a key marker of the trophectoderm, *Gata3* is known to be expressed from the 4- to 8-cell stage onward and to play a crucial role in the transition from morula to blastocyst (Home et al. 2009).

The remaining three TFs, *Lmx1a*, *Zfp410* and *Crx*, followed a major ZGA expression profile, with no detectable transcripts before the 2-cell stage and a peak of expression at the late 2-cell stage (**Figure 52**). While the mRNA levels of *Lmx1a* and *Crx* decreased after major ZGA (**Figure 52A, B**), the expression of Zfp410 remained elevated through to the 16-cell stage (**Figure 52C**). *Crx* transcript levels were low, and showed substantial variability between embryos (**Figure 52B**).

In conclusion, and in line with the candidate filtering performed in **section 4.1.2**, all the selected TFs were expressed at the mRNA level during preimplantation development. Importantly, their expression was prominent during the critical time of ZGA, highlighting their potential involvement in key developmental transitions. This mRNA expression at ZGA underscores the significance of these TFs in regulating early embryonic processes, laying the groundwork for further investigation into their roles in orchestrating transcription at these stages.

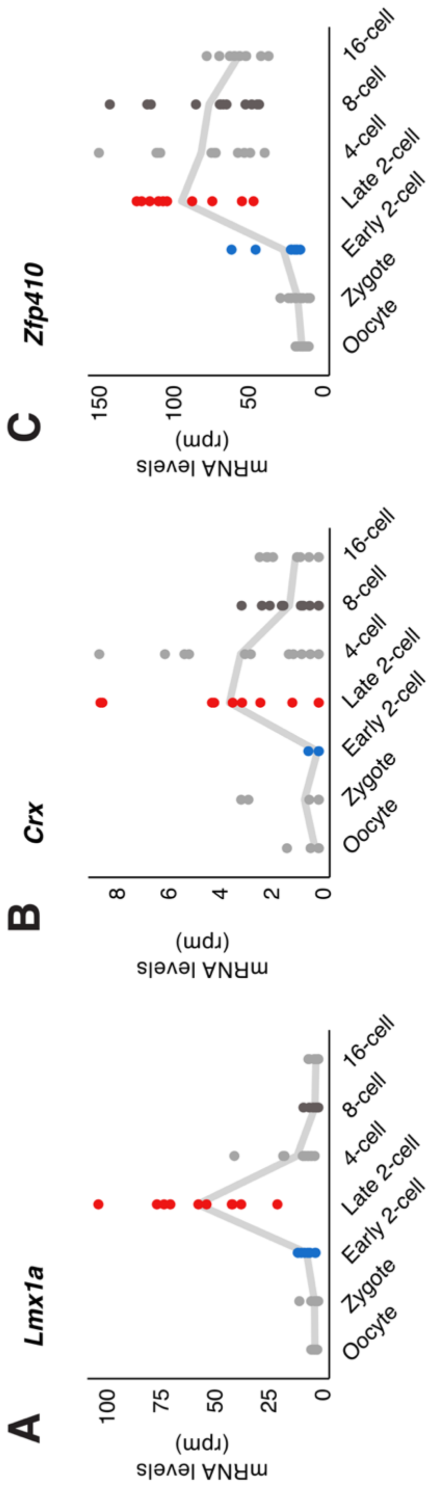


Figure 52. Expression profiles of zygotically expressed TFs during preimplantation development. (A-C) Each dot represents *Lmx1a* (A), *Crx* (B), *Zfp410* (C) mRNA levels (rpm) in individual embryos at the indicated stage. The trend line connects the mean values across embryos of individual stages

4.4.2 Expression profiles at the protein level

Some of the immunostainings were performed by Camille Noll.

During early embryogenesis, mRNA and protein expression are frequently disconnected from each other. Hence, I decided to also investigate the expression at the protein level of these nine TFs, by immunofluorescence (IF) at key preimplantation stages: zygote, early 2-cell, late 2-cell, 4-cell and 8-cell. As a matter of fact, I found CRX to be absent in embryos between zygote and 8-cell stages (**Figure 53A**). To rule out the possibility of technical issues with antibody specificity, I overexpressed CRX in mESCs and performed IF, successfully detecting nuclear CRX accumulation (**Figure 53B**). This experiment confirmed that the absence of CRX in early embryos was not due to a simple limitation in the experimental procedure.

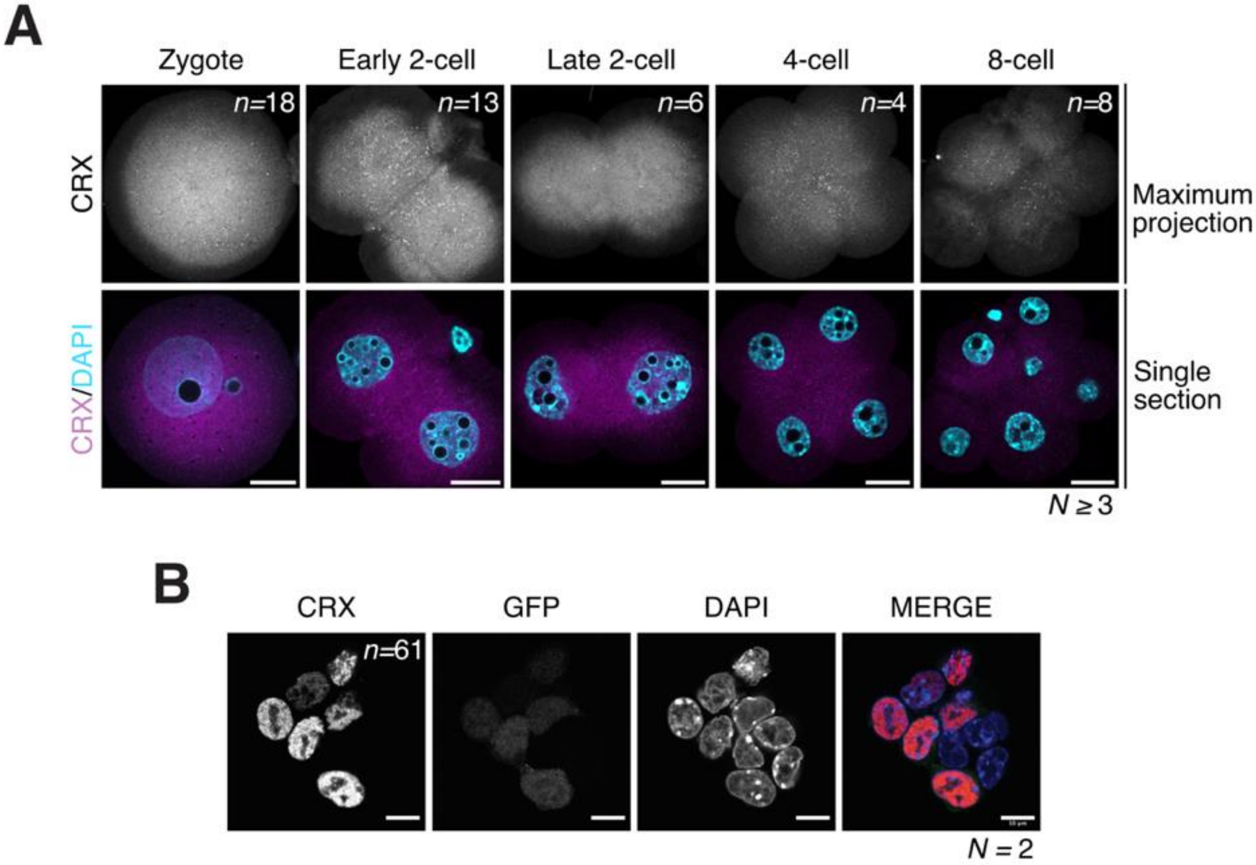


Figure 53. Expression profile of CRX at the protein level. (A) Representative images of CRX immunostainings at the indicated developmental stages. All embryos within each replicate were processed and acquired using the same conditions, hence the intensity of the fluorescent signal is comparable in all embryos. Top images are maximum intensity projections, bottom are merged images with DAPI staining shown as single confocal sections. *n* is the total number of embryos analyzed per stage. *N*, number of independent replicates. Scale bars, 20µm. (B) Representative images of CRX immunostaining and GFP fluorescence upon CRX overexpression in ESCs. *n* is the total number of cells that were displaying both GFP and CRX signal. *N*, number of independent replicates. Scale bars, 10µm.

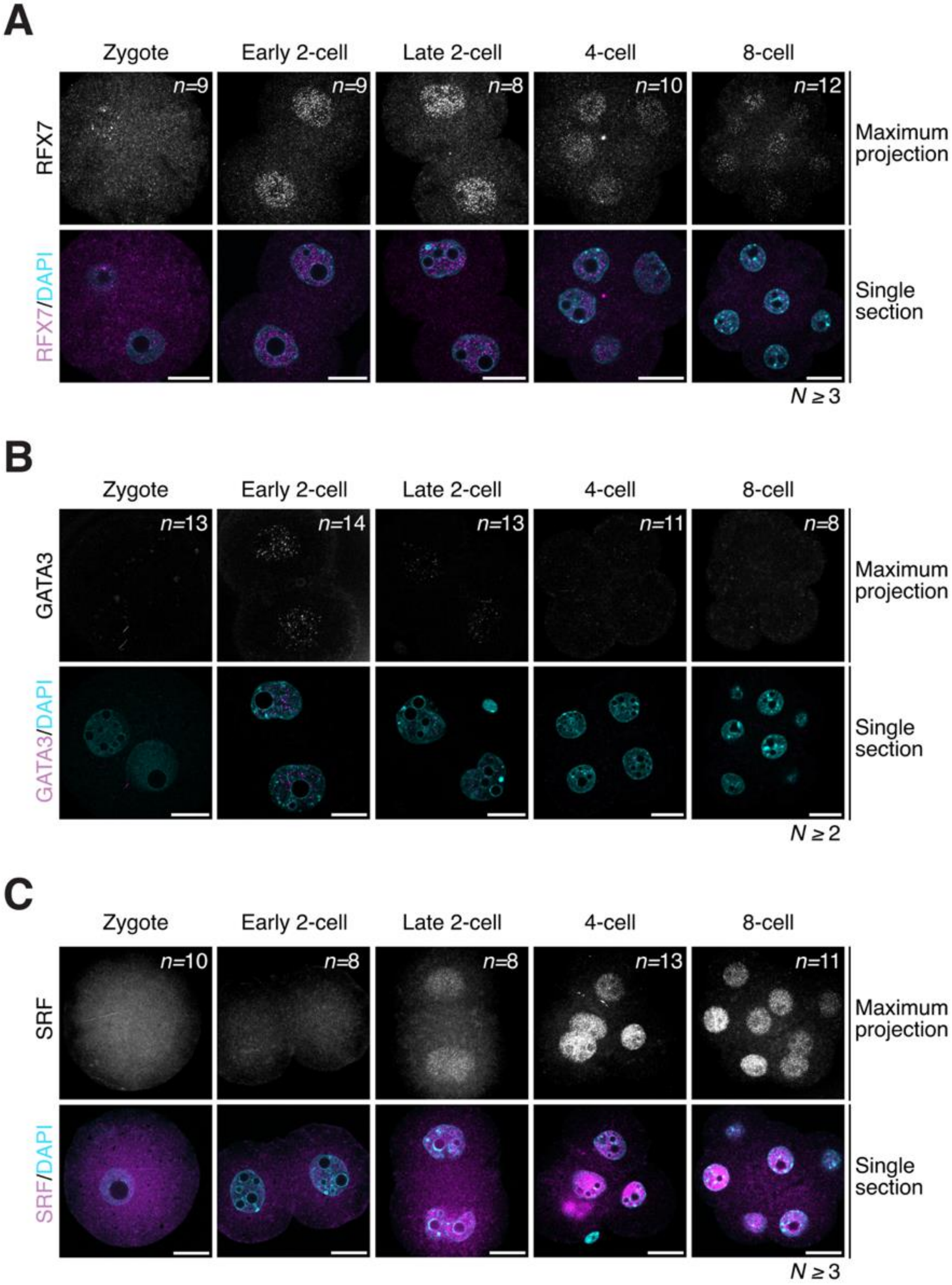


Figure 54. Expression profiles of zygotically expressed TFs at the protein level. (A-C) Representative images of RFX7 (A), GATA3 (B) and SRF (C) immunostainings at the indicated developmental stage. All embryos within each replicate were processed and acquired using the same conditions, hence the intensity of the fluorescent signal is comparable in all embryos. Top images are maximum intensity projections, bottom are merged images with DAPI staining shown as single confocal sections. *n* is the total number of embryos analyzed per stage. *N*, number of independent replicates. Scale bars, 20µm.

Three TFs were either absent or barely detectable in zygotes and then became transiently present at certain stages or remained detectable throughout the stages analyzed after their initial expression (**Figure 54**). For instance, RFX7 exhibited a faint, barely detectable signal in zygotes, but by the early 2-cell stage it displayed clear nuclear localization, which persisted, albeit fainter, at the 8-cell stage (**Figure 54A**). For GATA3, I detected very low levels of the protein, which transiently formed foci in the nuclei of early 2-cell stage embryos. This signal was still observed in the late 2-cell stage but was absent in all subsequent stages analyzed (**Figure 54B**). SRF, which was undetectable in zygotes, was not observed in the early 2-cell stage either but became readily detectable starting from the late 2-cell stage onward (**Figure 54C**). Based on these findings, I categorized these three TFs as zygotically expressed at the protein level even though their mRNA was maternally inherited (**Figure 51**), further confirming the uncoupling of mRNA and protein expression during early development.

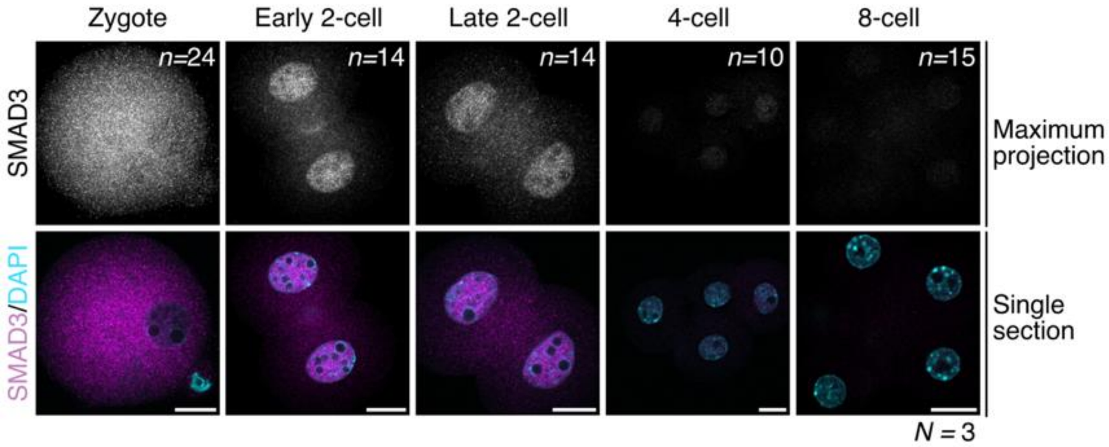


Figure 55. Expression profile of SMAD3 at the protein level. Representative images of SMAD3 immunostainings at the indicated developmental stage. All embryos within each replicate were processed and acquired using the same conditions, hence the intensity of the fluorescent signal is comparable in all embryos. Top images are maximum intensity projections, bottom are merged images with DAPI staining shown as single confocal sections. *n* is the total number of embryos analyzed per stage. *N*, number of independent replicates. Scale bars, 20µm.

In the case of SMAD3, the protein appeared to localize to the cytoplasm of zygotes (**Figure 55**). Importantly, I observed depletion of signal in the zygotic pronuclei, suggesting that the observed signal was genuine rather than cytoplasmic background (**Figure 55**). However, I was unable to definitely differentiate between the two possibilities. To further investigate SMAD3 expression and disentangle true cytoplasmic signal from background noise, I examined cultured embryos from F1 females, where ovulation was induced by hormone injection. This approach allowed me for more precise timing of the various developmental stages. I collected early zygotes and cultured them, fixing them at specific time points after hormone injection (**Figure 56**).

Unexpectedly, I did not observe any nuclear localization of SMAD3 at any of the stages analyzed (**Figure 56**). Instead, I observed consistent cytoplasmic signal and nuclear depletion across all stages analyzed, from the zygote to the late 2-cell stage, reminiscent of the nuclear depletion pattern observed in zygotes from CD1 mice shown in Figure 55 (**Figure 56**). Initially, I attributed these differences to the change in mice strains. To clarify, I directly collected late 2-cell stage embryos at around 48h post-hormone injection from females to assess whether SMAD3 nuclear enrichment would be observed in F1 non-cultured late 2-cell stage embryos (**Figure 57**).

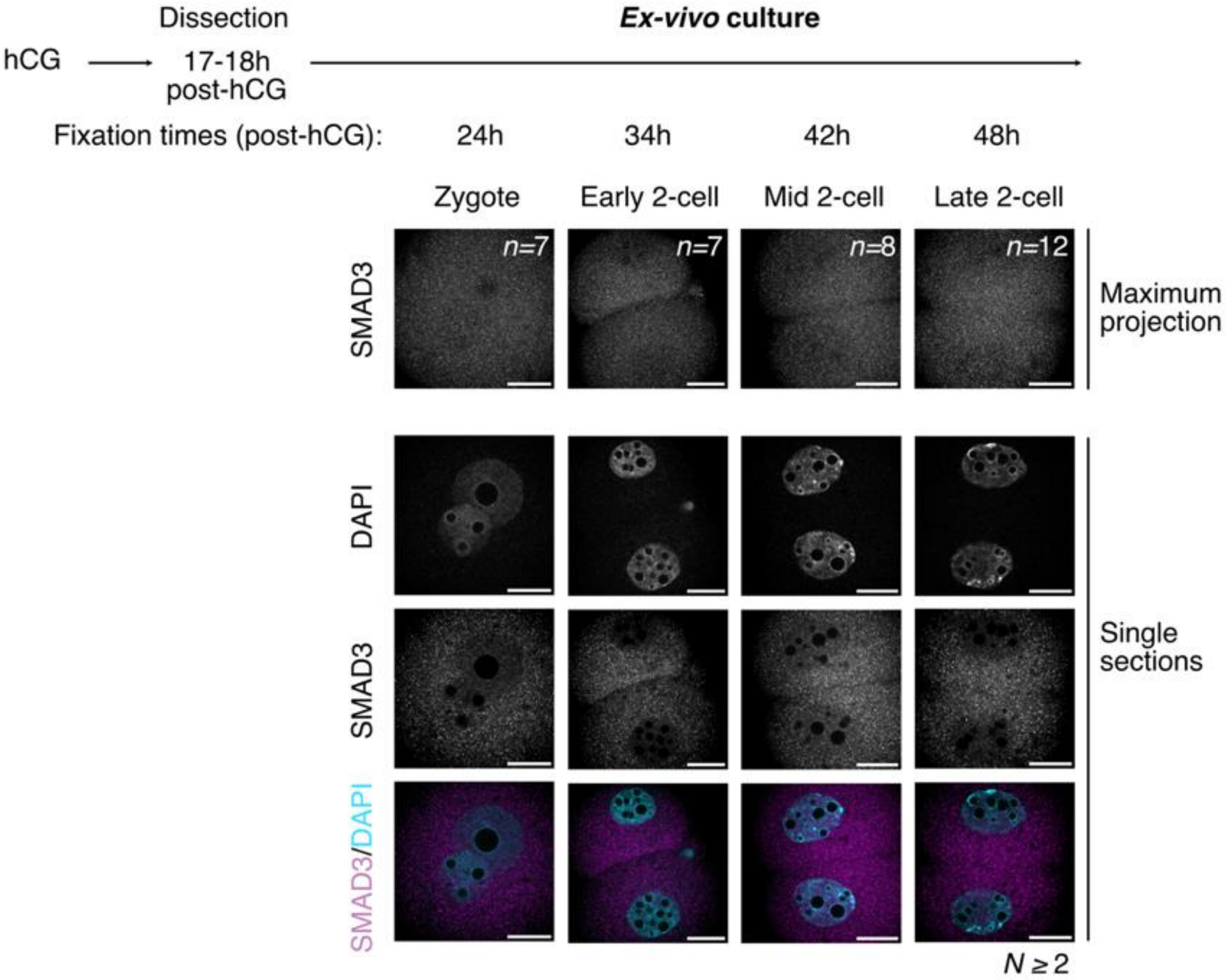


Figure 56. Expression profile of SMAD3 at the protein level in F1 cultured embryos. Representative images of SMAD3 immunostainings at the indicated developmental stage (timing post-hCG are indicated). All embryos within each replicate were processed and acquired using the same conditions, hence the intensity of the fluorescent signal is comparable in all embryos. Top images are maximum intensity projection, bottom are DAPI, SMAD3 and merged images with DAPI staining shown as single confocal sections. *n* is the total number of embryos analyzed per stage. *N*, number of independent replicates. Scale bars, 20µm.

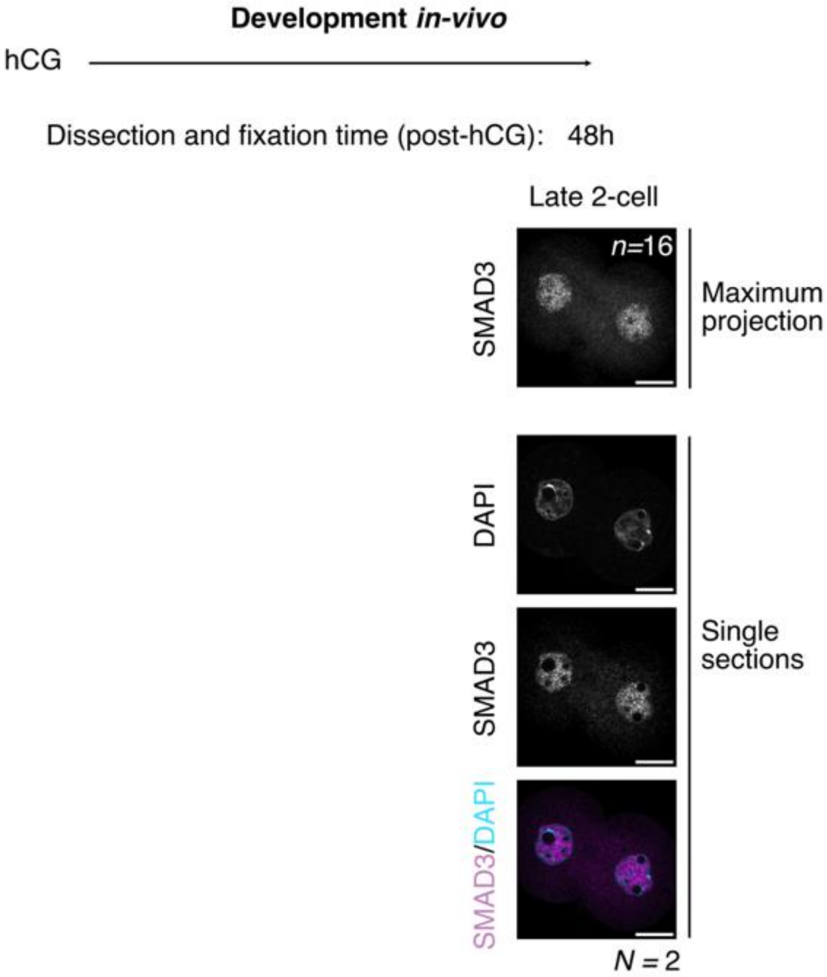
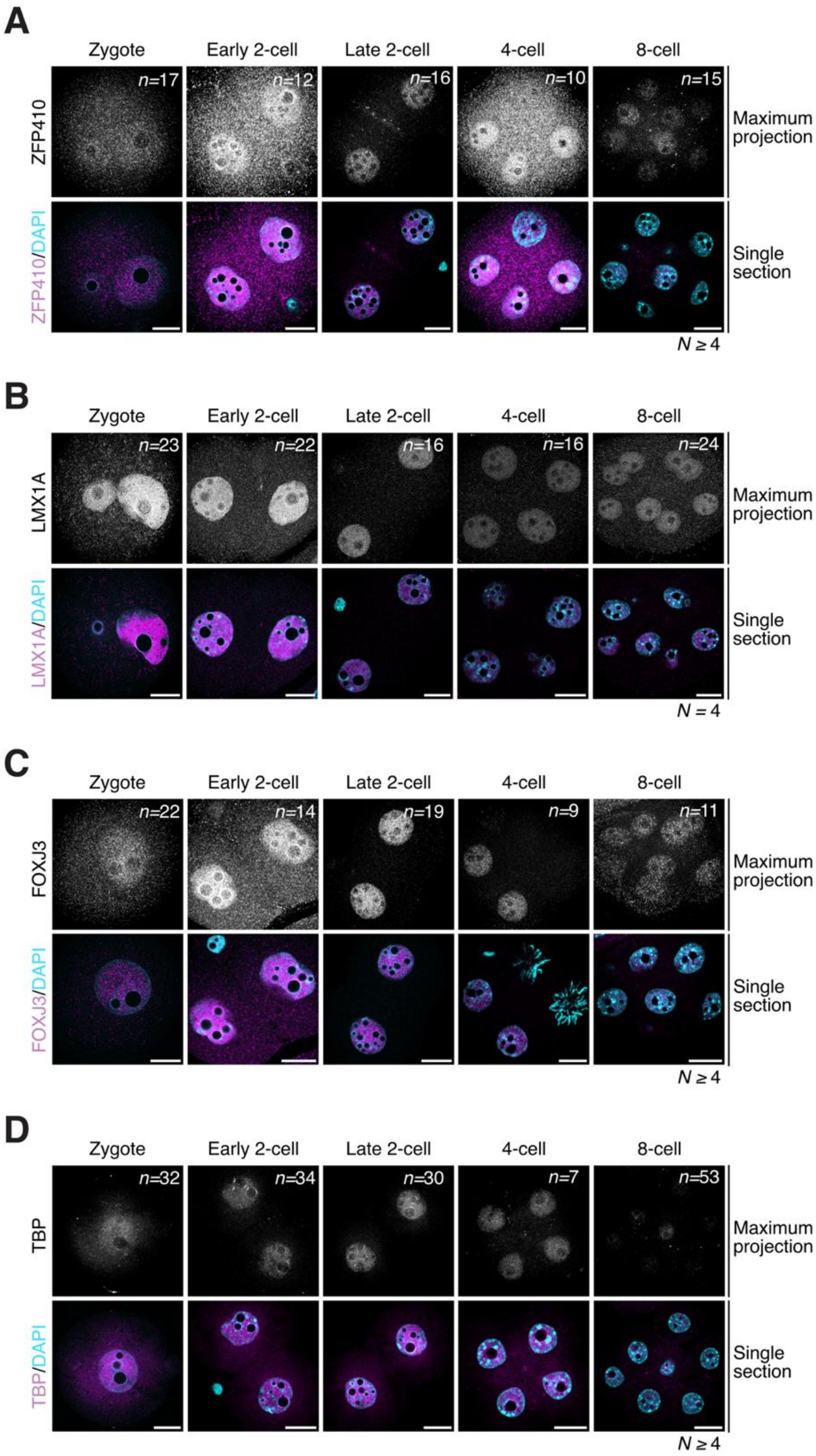


Figure 57. Expression of SMAD3 at the protein level in late 2-cell stage F1 non-cultured embryos. Representative images of SMAD3 immunostainings at the late 2-cell stage (48h post-hCG). Top image is a maximum intensity projection, bottom are DAPI, SMAD3 and merged images with DAPI staining shown as single confocal sections. *n* is the total number of embryos analyzed. *N*, number of independent replicates. Scale bars, 20µm.

Strikingly, I found that there was a clear SMAD3 nuclear enrichment in all the late 2-cell stage embryos analyzed (**Figure 57**), suggesting that the absence of nuclear SMAD3 in Figure 56 was a consequence of *ex-vivo* embryo culture. Given that SMAD3 is a downstream effector of TGFB signaling, this raises the possibility that TGFB signaling may be active in early embryos, in the females, facilitated by maternal tissues, but disrupted under culture conditions.

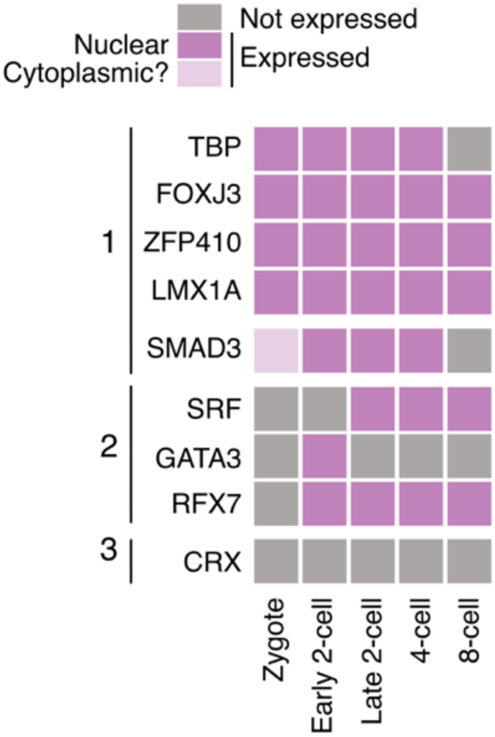
All other TFs were expressed throughout the stages assessed (**Figure 58**). ZFP410 was already visible in zygote pronuclei but exhibited clear nuclear localization from the early 2-cell stage (**Figure 58A**). LMX1A showed robust nuclear enrichment in zygotes which persisted until the 8-cell stage (**Figure 58B**). FOXJ3 and TBP were localized to the nuclei at all stages, though TBP became barely detectable by the 8-cell stage (**Figure 58C, D**).



(Legend on next page)

Figure 58. Expression profiles of maternally inherited TFs at the protein level. (A-D) Representative images of ZFP410 (A), LMX1A (B), FOXJ3 (C) and TBP (D) immunostainings at the indicated developmental stage. All embryos within each replicate were processed and acquired using the same conditions, hence the intensity of the fluorescent signal is comparable in all embryos. Top images are maximum intensity projections, bottom are merged images with DAPI staining shown as single confocal sections. *n* is the total number of embryos analyzed per stage. *N*, number of independent replicates. Scale bars, 20µm.

Thus, with the exception of CRX, whose transcript levels were low and variable across samples, all candidate TFs were expressed at the protein level in early mouse embryos, though exhibiting distinct temporal dynamics in their nuclear localization (**Figure 59**).



▨ Not expressed
Nuclear ■ | Expressed
Cytoplasmic? □ | Expressed

		Zygote	Early 2-cell	Late 2-cell	4-cell	8-cell
1	TBP	■	■	■	■	▨
	FOXJ3	■	■	■	■	■
	ZFP410	■	■	■	■	■
	LMX1A	■	■	■	■	■
	SMAD3	□	■	■	■	▨
2	SRF	▨	▨	■	■	■
	GATA3	▨	■	▨	▨	▨
	RFX7	▨	■	■	■	■
3	CRX	▨	▨	▨	▨	▨

Figure 59. Qualitative summary of expression profiles of 9 candidate TFs at the protein level. Gray represents not expressed, pink nuclear localization and light pink potentially cytoplasmic localization. TFs are arranged by patterns of expression observed: 1-maternally inherited protein, 2-zygotic expression, 3-virtually absent.

In conclusion, I have delineated the precise temporal expression patterns of nine TFs during the first few cell divisions of mouse embryonic development. Importantly, some of these TFs, specifically FOXJ3, TBP and SRF, were either ubiquitously present across insertions of certain TE families (FOXJ3 and TBP in B2_Mm1a) or their presence correlated with differential subfamily expression (SRF and FOXJ3 with MaLR ORR1A1) (**section 4.2**). In addition, FOXJ3 and TBP exerted specific effects on SINEB2 and ORR MaLR, respectively, upon overexpression in ESCs (**section 4.3.1**). Given that both TE families are understudied in preimplantation development, with their regulatory mechanisms largely unknown, I chose to focus on elucidating the role of these two TFs in transcriptional regulation during early embryonic development. Finally, the importance of SRF in the evolution of MT2 lineage (**section 4.2.4**) and its capacity to induce MERVL transcription both in ESCs upon overexpression (**section 4.3.1**), and in a luciferase reporter assay (**section 4.3.2**) prompted further investigations into its regulatory role in the embryo.

Even though the five other expressed TFs may also play a role in TE regulation during development, in the final phase of my PhD, I focused on investigating the roles of **SRF**, **FOXJ3** and **TBP** in mouse preimplantation development. I sought to characterize their function in regulating both TEs of interest and genes during ZGA, as their involvement in this context remains unexplored.

4.5 Addressing the function of TFs in regulating genes and TEs in the early embryo: SRF, FOXJ3 and TBP

Based on the expression profiles, I set-up two methods for TF depletion in mouse preimplantation embryos. There are different possible methods to deplete or alter the function of a protein in embryos. One possibility is the use of a dominant-negative, which competes with the endogenous protein for binding to the sequence of interest, without performing the transcription activation activity of the endogenous TF. Another possibility is to acutely deplete the protein in the embryo, by a Trim-Away system of depletion, which induces degradation of the target protein. For SRF, whose protein is expressed from the 2-cell stage, I initially attempted mRNA depletion using siRNAs, but this only resulted in depletion by the 4-cell stage (**data not shown**), making it unsuitable for analyzing effects on ZGA. Therefore, I turned to a dominant negative approach, which I will describe in the first part of this section. Since FOXJ3 and TBP are maternally inherited proteins, I employed a Trim-Away approach for their depletion, enabling acute protein loss in early embryos, which I will detail in the second part of this section.

Data analysis in this section of my PhD has been done by Carlos Michel Mourra-Díaz (with the help of Dr. Tamas Schauer) and Dr. Marlies E. Oomen.

4.5.1 SRF regulates genes and MERVL retrotransposons during ZGA

4.5.1.1 A dominant negative strategy for SRF loss-of-function

In mice, SRF is a protein consisting of 504 amino acids, encoding 7 exons in its largest isoform (**Figure 60**) (Deshpande et al. 2022). The C-terminal (C-ter) region contains the transactivation domain, which spans exons 4 to 6 (**Figure 60**), while the N-terminal (N-ter) region harbors the highly conserved DNA-binding and dimerization domain known as the MADS box, spanning exons 1 to 3 (**Figure 60**) (Deshpande et al. 2022). MADS is an acronym for the first TFs identified with this domain: MCM1, Agamous, Deficiens and SRF, which were discovered in yeast, plants and humans, demonstrating the conservation of this domain across different kingdoms of life (Passmore et al. 1989; Yanofsky et al. 1990; Sommer et al. 1990; Norman et al. 1988; Schwarz-Sommer et al. 1990). The MADS box binds to specific DNA sequences known as CArG motifs, which in mammals are referred to as Serum Response Elements (SREs). SRF is well-known for its involvement in controlling the immediate early response to extracellular stimuli and cell proliferation (Greenberg and Ziff 1984; Treisman 1986, 1987; Greenberg et al. 1987; Norman et al. 1988; Miano 2010). A study using a knockout mouse model in the late 1990s demonstrated that SRF is crucial for mesoderm formation in mammals (Arsenian et al. 1998). Although zygotic SRF knockout leads to gastrulation defects and embryonic lethality by E12.5 (Arsenian et al. 1998), the role of SRF at earlier stages of development remains unexplored, as no studies have addressed maternal SRF depletion.

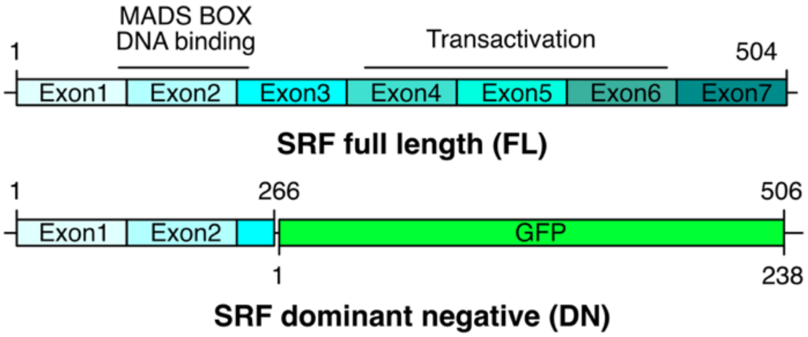


Figure 60. Schematic representation of the main domains of a full-length SRF protein (longest isoform) and the dominant negative (DN). The dominant negative contains only amino acids 1-266 fused to a GFP instead of the transactivation domain of SRF.

To induce a loss-of-function in the embryo, I turned to a dominant-negative approach by replacing the entire C-ter region of SRF, encompassing the transactivation domain, by GFP (**Figure 60**). This DN has been previously reported and used in the literature (Belaguli et al. 1999) and with the addition of GFP, was comparable in length to the longest SRF isoform (**Figure 60**). First, I validated that the DN construct successfully impaired SRF function using the established luciferase assay system on MT2_Mm_i (**Figure 61**).

Therefore, I microinjected the DN mRNA in zygote cytoplasm to induce LOF during ZGA and used dsRed mRNA as a positive control for the microinjection process (**Figure 62A, B**). GFP from the DN construct (**Figure 60**) allowed me to verify nuclear expression such that I only collected embryos with proper DN subcellular localization (**Figure 62B**). As a positive control for MERVL and ZGA effects, I injected embryos with an ASO targeting *Dux* (Guo et al. 2024), while scramble ASO and dsRed mRNA injected embryos served as negative controls for all experiments (**Figure 62A**). After microinjections, I cultured the embryos until the late 2-cell stage (48h post-hCG) and collected them for single embryo RNA sequencing (**Figure 62A**).

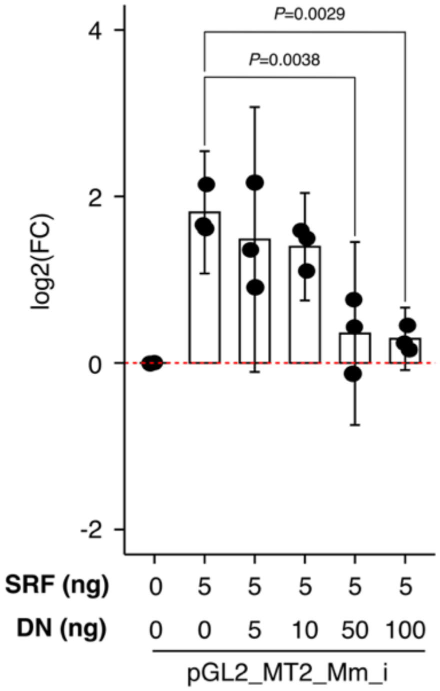


Figure 61. DN competition with SRF WT in luciferase assay in human HEK293 cells. Log2 transformed fold change of the normalized luciferase activity of MT2_Mm_i in the presence of 5 ng of pCMV-encoding SRF coding sequences with increasing amounts of pCMV encoding for the SRF dominant negative fused to GFP (DN) over control (pCMV-empty). Bar plots show the mean +/- 95% confidence intervals (*N* = 3). For statistical analysis, preselected hypotheses compared 5 ng of full-length SRF only, with 50 and 100 ng of DN construct transfected.

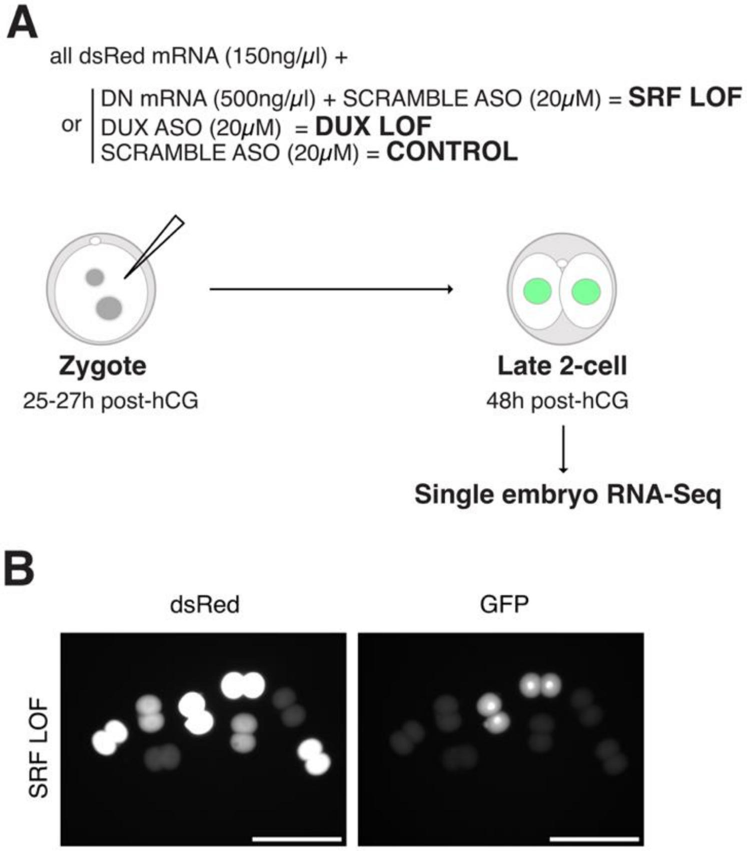


Figure 62. SRF and DUX LOF in embryos. (A) Schematic representation of the experimental design for LOF of DUX and SRF followed by single embryo RNA-seq at the late 2-cell stage. (B) Representative images of SRF LOF embryos. The dsRed is a control for microinjection. Embryos with a clear nuclear GFP signal were collected for single embryo RNA-seq. Scale bars, 200µm.

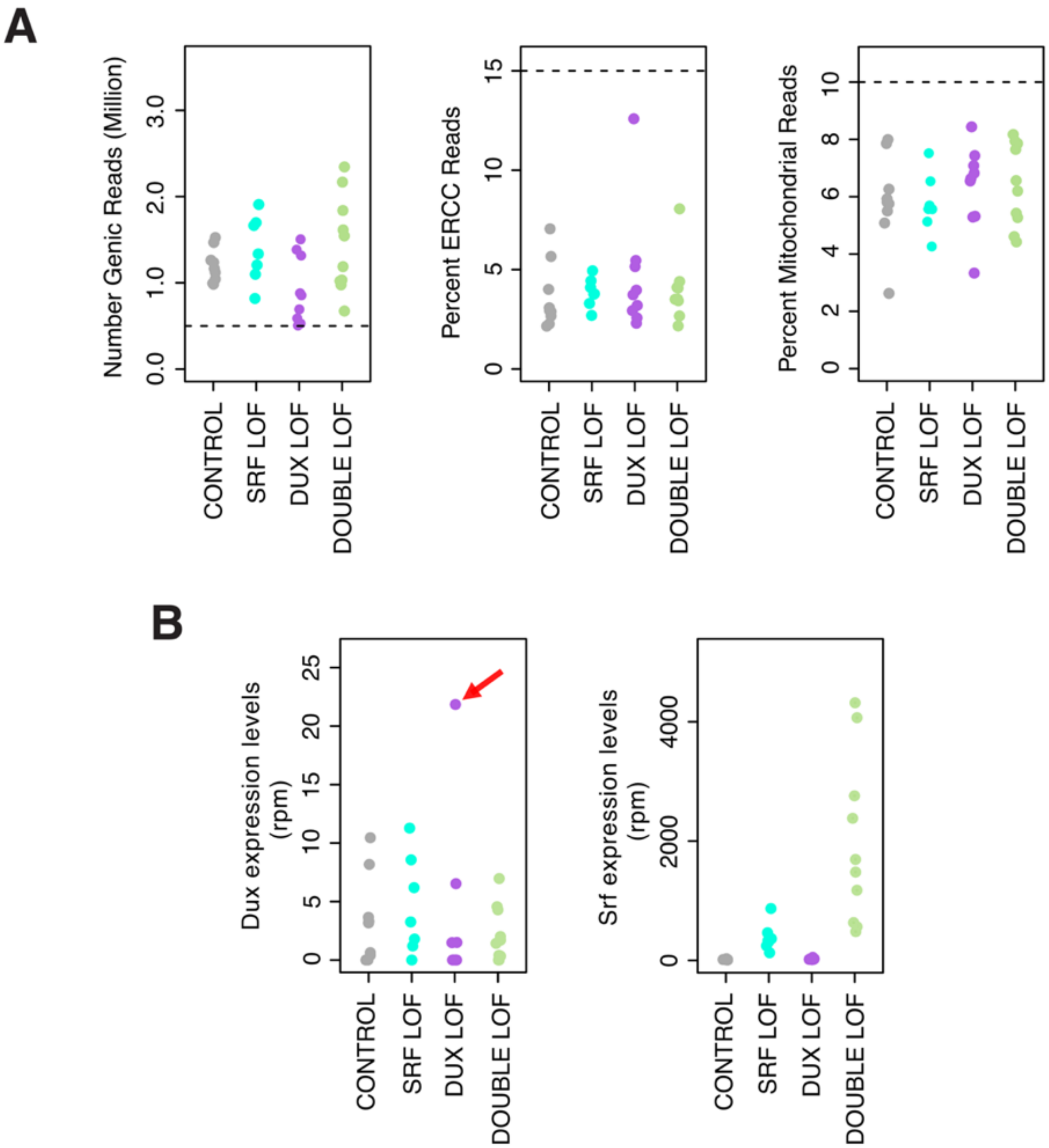


Figure 63. Single embryo RNA-seq quality controls. (A) Quality control dotplots showing the number of genic reads (Million), the percent of ERCC reads and of Mitochondrial reads in all experimental conditions as indicated. Each dot is an embryo. Dashed lines represent the threshold applied for quality control filtering. (B) Expression levels of *Dux* and *Srf* in all experimental conditions as indicated. Each dot is an embryo. The red arrow points to an embryo that was removed from further analysis due to over-accumulation of *Dux* transcripts, while *Dux* should be knocked-down.

4.5.1.2 SRF regulates genes and is essential for development

I collected a total of 7 SRF LOF, 9 DUX LOF, 10 DOUBLE LOF and 9 CONTROL embryos (**Figure 63**). While all libraries passed the general quality controls (QC) (**Figure 63A**), one DUX LOF embryo was excluded due to high *Dux* expression, indicating ineffective knockdown (**Figure 63B**, *red arrow*). Differential gene expression analysis revealed 763 differentially regulated genes in SRF LOF (**Figure 64A**), with slightly more downregulated (438) than upregulated (325).

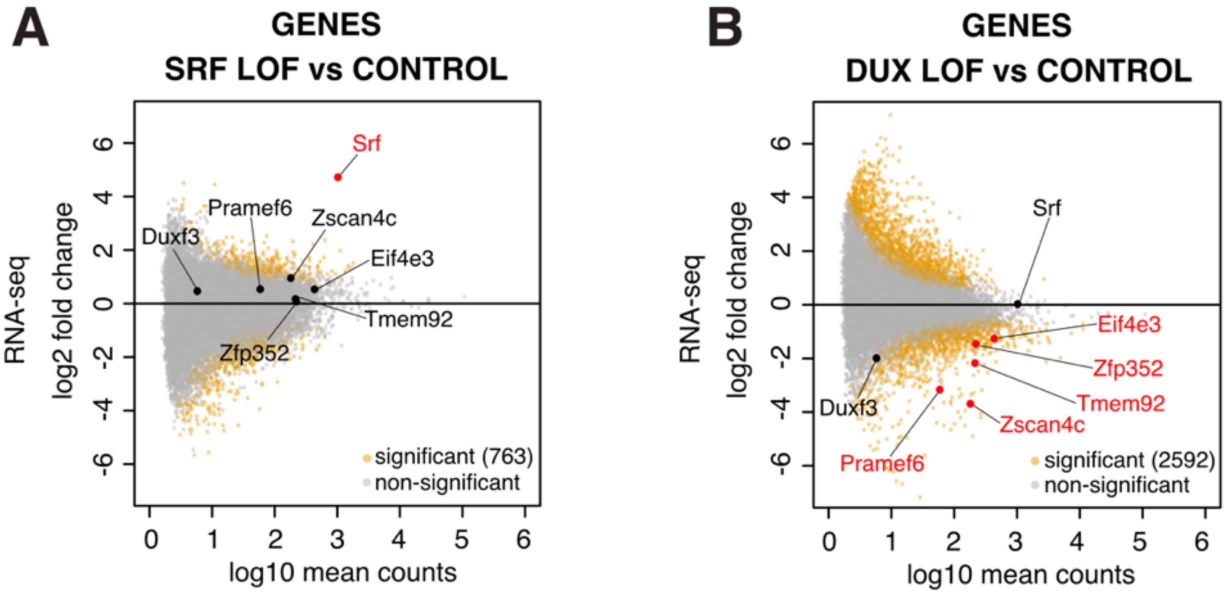


Figure 64. Differential gene expression analysis of SRF LOF (A) and DUX LOF (B). (A,B) MA plots comparing the log2 fold change in SRF LOF vs CONTROL (A) and DUX LOF vs CONTROL (B) embryos against the log10 RNA-seq mean counts. Differentially expressed genes are labelled in orange (padj < 0.05), non-differentially expressed genes are in gray. Examples of genes are labelled, significant in red, non-significant in black.

Upon DUX LOF approximately 2600 genes were differentially regulated, with 1049 downregulated and 1543 upregulated (**Figure 64B**). Notably, classic Dux targets such as *Zscan4* (*a-f*), *Zfp352*, *Pramef6*, *Tmem92*, *Eif4e3* were downregulated (**Figure 64B**), in agreement with previous reports (Hendrickson et al. 2017; De Iaco et al. 2017). None of these genes were affected in SRF LOF (**Figure 64A**). Comparing SRF and DUX LOF revealed minimal overlap in affected genes, with only 130 common to both (**Figure 65**), suggesting that SRF and DUX are regulating different gene sets during ZGA.

To further characterize SRF LOF, we were then curious to know the type of genes affected. Using the database of transcriptome in mouse early embryo (DBTMEE) (Park et al. 2015), we categorized the down and upregulated genes (**Figure 66**). Of the 438 significantly downregulated genes, roughly 19% were major ZGA genes, while a small subset (about 4%) were maternal RNAs (**Figure 66A**). In fact, most of the downregulated genes did not belong to any DBTMEE-defined categories (**Figure 66A**). These findings indicate that SRF orchestrates transcription at the late 2-cell stage during ZGA, exerting an influence on gene expression that extends beyond the regulation of strictly ZGA-specific genes.

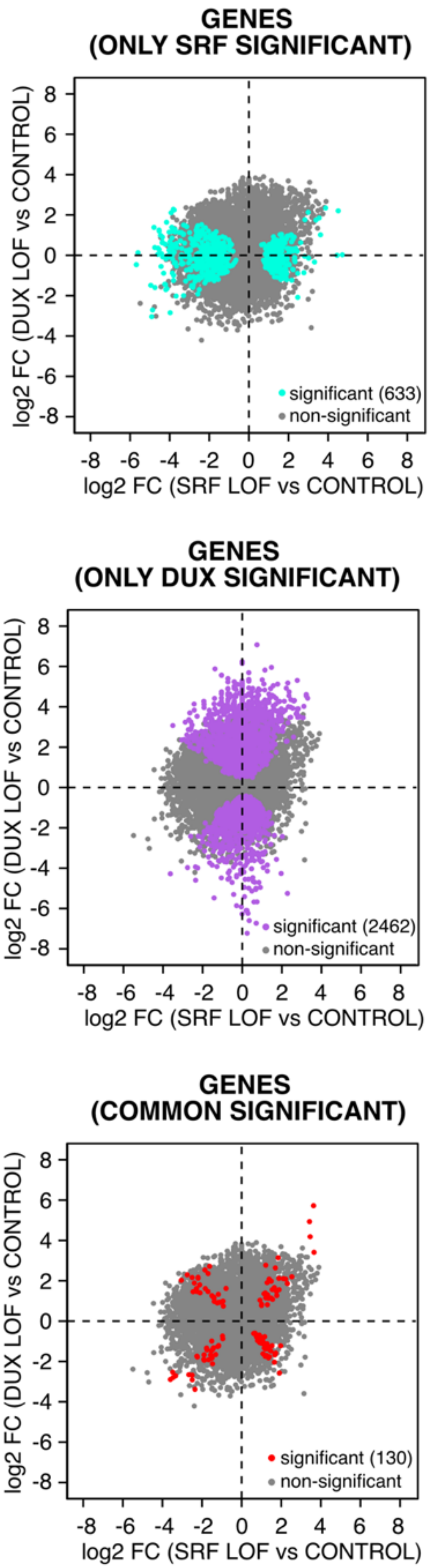


Figure 65. Comparison of differentially expressed genes upon SRF LOF vs DUX LOF. Scatterplot comparing the log2 fold change in SRF LOF vs CONTROL to DUX LOF vs CONTROL. Non-significant genes in neither of the two conditions are shown in gray. From top to bottom, significantly changed genes only in SRF LOF, DUX LOF and common to both LOFs are displayed in cyan, purple and red, respectively.

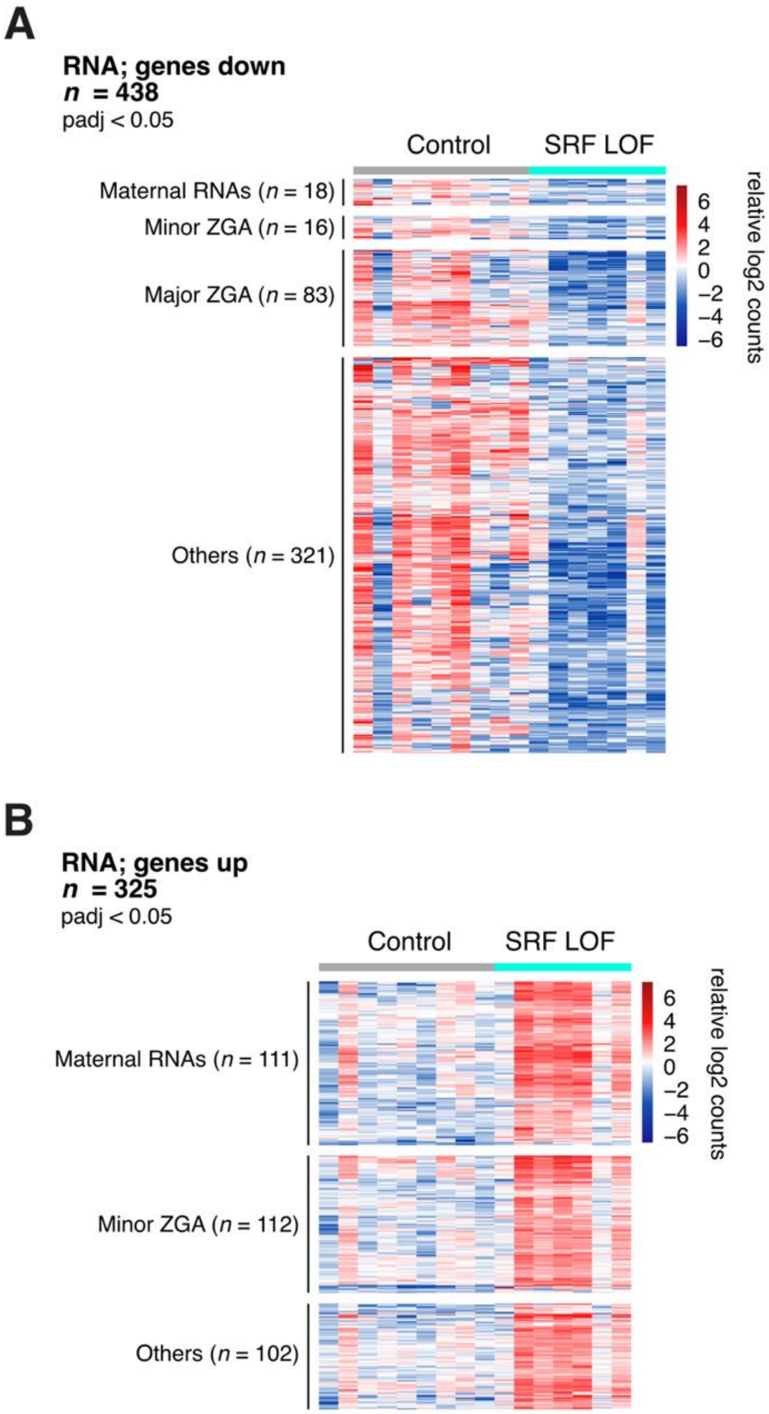


Figure 66. Classification of downregulated (A) and upregulated (B) genes in SRF LOF. (A,B) Heatmaps showing significantly downregulated (A) and upregulated (B) genes in SRF LOF ordered according to DBTMEE gene classification. Values are log2 transformed normalized counts centered on the row mean. Each column is an individual embryo, clustered by condition. *n* is the number of genes per category indicated.

Approximately one third of the upregulated genes were categorized as maternal RNAs, another third as minor ZGA genes, while the remaining third did not fit any defined category (**Figure 66B**). The elevated expression of maternal RNAs and minor ZGA genes at the late 2-cell stage suggested a developmental delay, implying an inability to properly degrade maternally inherited transcripts and suggestive of a disruption in the MZT. Collectively, these findings highlight SRF as a regulator of gene expression during ZGA, with its absence impairing the proper execution of the MZT.

We then conducted a principal component analysis (PCA) that included CONTROL, DUX LOF and SRF LOF embryos, alongside a non-manipulated RNA-seq dataset from embryos at various preimplantation developmental stages, all generated using the same protocol as employed in this study (**Figure 67**) (Oomen et al. 2025). As anticipated, control embryos clustered with late 2-cell non-manipulated embryos, whereas SRF LOF embryos positioned between early and late 2-cell non-manipulated embryos along PC1, suggesting that SRF LOF embryos indeed experienced a developmental delay (**Figure 67**). Interestingly, DUX LOF embryos grouped with control embryos (**Figure 67**).

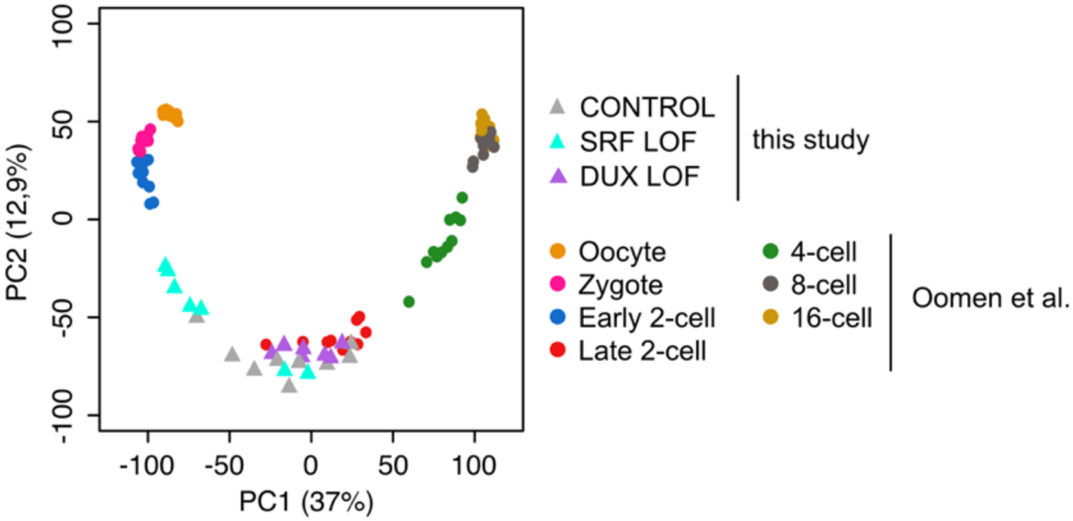


Figure 67. Principle component analysis of CONTROL, SRF LOF and DUX LOF embryos together with embryos from oocyte to 16-cell stage from (Oomen et al. 2025). Triangles are embryos from this study, color-coded based on the condition. Circles are embryos from Oomen et al. (2025), color-coded based on the stage. The proportions of variance explained by PC1 and PC2 are indicated.

The observed transcriptional phenotype, coupled with the failure to properly execute MZT, prompted us to investigate whether SRF was necessary for developmental progression. To address this, I performed a developmental assay experiment in which embryos were injected with either DN mRNA or GFP mRNA, and their development was monitored for the first four days of embryogenesis, up to the blastocyst stage (**Figure 68**). Remarkably, only half of the SRF LOF embryos reached the blastocyst stage (14/28), whereas most control embryos progressed to this stage (17/19) by day 4 (**Figure 68**). Primarily the second and third cell divisions appeared to be impaired: roughly 25% of SRF LOF embryos remained at the 2-cell stage on day 2, and 25% were still at the 4-cell stage, while 100% of control embryos reached the 8-cell stage by the same time (**Figure 68B**).

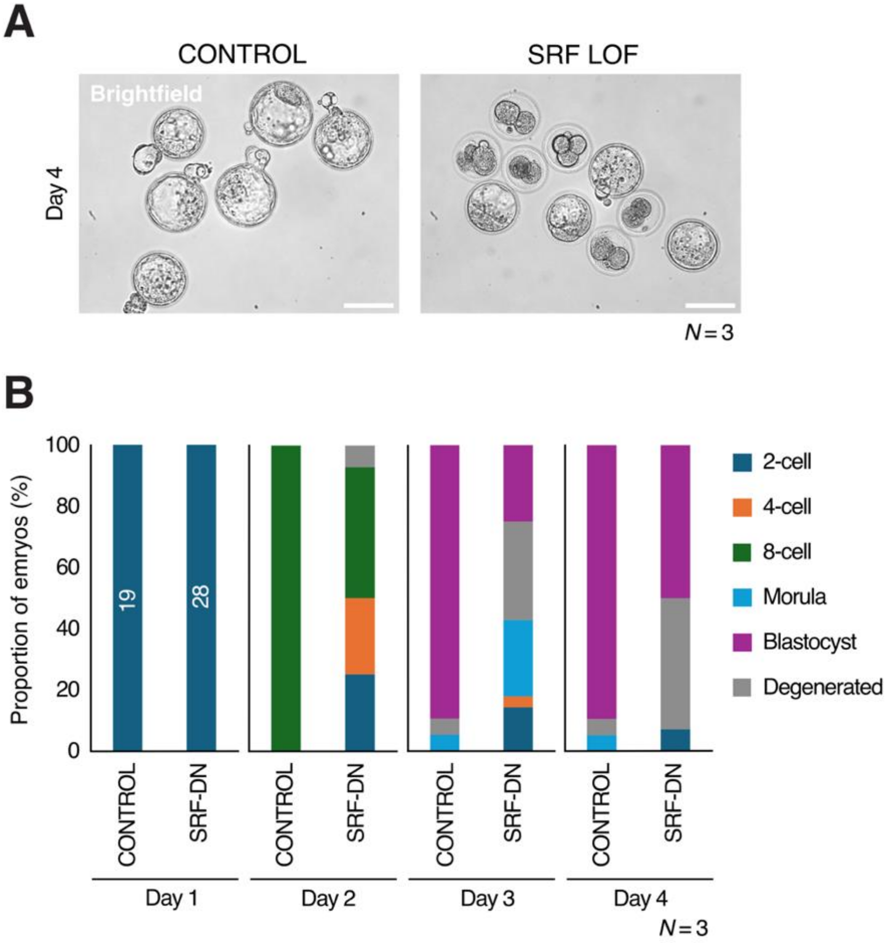


Figure 68. Developmental progression of control and SRF LOF embryos. (A) Brightfield images of CONTROL (left) and SRF LOF (right) embryos at day 4 (116h post-hCG (phCG)) of embryonic development. *N,* number of independent replicates. Scale bars, 100 µm. (B) Proportion of CONTROL and SRF LOF embryos in specific stage as indicated by the color code at day 1 (48h phCG), day 2 (72h phCG), day 3 (96h phCG) and day 4 (116h phCG). *N,* number of independent replicates. The numbers indicated at day 1 represent the total number of embryos analyzed over the different independent replicates.

Our findings suggest that SRF is critical for developmental progression, likely through its regulation of gene expression, though the precise mechanisms underlying this developmental phenotype cannot yet be definitively determined. Among the differentially regulated genes are those involved in the cell cycle, such as *Cyclin1* and *Cdkn1a*. Both are essential for proper cell cycle progression and may be implicated in the observed phenotype.

4.5.1.3 SRF controls TEs and host chimeric transcript expression

Subsequently, and perhaps most critically in the context of this project, we investigated whether SRF regulates TEs in late 2-cell stage embryos. 14 TEs were significantly differentially regulated in SRF LOF embryos compared to control (**Figure 69A**). All except two TEs were downregulated and this included different families of relatively old L1 (L1MCb, L1MA7), some ERVKs (RLTR31_Mur, RLTR9E, RMER6B, RMER13A1), some old MaLR families (ORR1C2-int, MLT1J-int, MLTN2), one SINEB1 family (B1_Mur1) and two DNA transposon families (Charlie1a, Tigger17c). Based on the results obtained in **section 4.3.2**, we looked more specifically at MT2_Mm and MT2C_Mm. At the family level, MT2_Mm was not affected in SRF LOF (padj = 0.99), whereas MT2C_Mm was downregulated about 3-fold but was not significant (padj = 0.06) (**Figure 69A**). DUX LOF led to about 85 significantly differentially regulated TEs, with about 70% of these being upregulated (62/85) (**Figure 69B**). MT2_Mm was among the most downregulated, showing a significant decrease (padj = 0.0008), as previously reported (**Figure 69B**) (De Iaco et al. 2017, 2020; Guo et al. 2019, 2024). MT2C_Mm though, was barely affected by removal of DUX (**Figure 69B**).

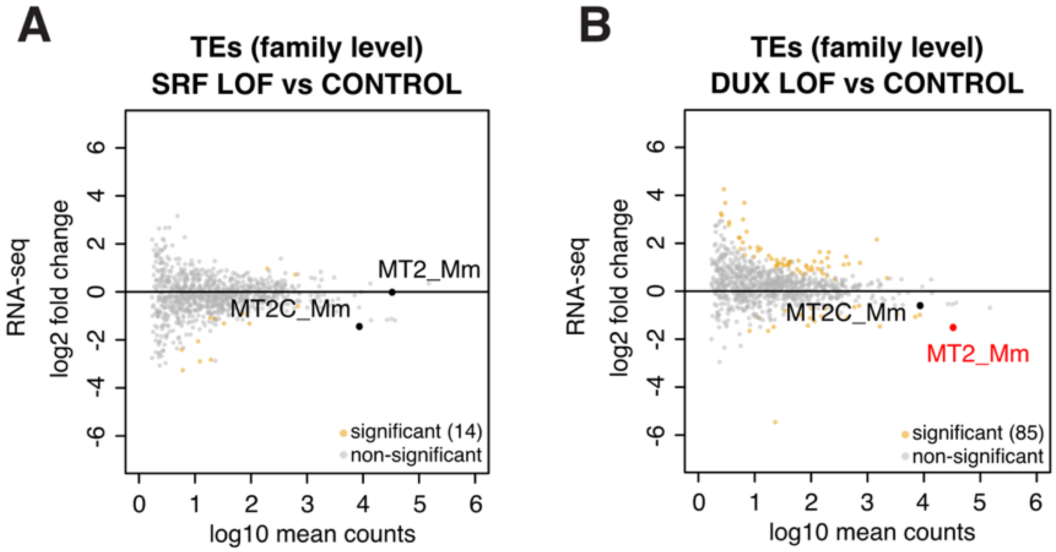


Figure 69. Differential TE expression analysis of SRF LOF (A) and DUX LOF (B). (A,B) MA plots comparing the log2 fold change in SRF LOF vs CONTROL (A) and DUX LOF vs CONTROL (B) embryos against the log10 RNA-seq mean counts. Differentially expressed TEs (from general DEseq object, see methods) are labelled in orange (padj < 0.05), non-differentially expressed TEs are in gray. Example of TEs (MT2_Mm and MT2C_Mm, considering only full-length insertions, as in phylogeny) are labelled, significant in red, non-significant in black.

To dissect more precisely the regulation of these TE families in the embryo, and in light of our phylogenetic and luciferase assay analysis for MT2C_Mm and MT2_Mm subfamilies, we decided to look at the effect of the LOFs on the newly established subfamilies. Strikingly, MT2C_Mm_vii was the strongest affected TE subfamily in SRF LOF, showing a downregulation of about 5-fold (**Figure 70A**). Importantly, none of the subfamilies belonging to any TE families other than MT2C_Mm_vii and MT2C_Mm_iv established by the phylogenetic analysis in **Section 4.2** were affected. MT2C_Mm_vii is the last MT2C_Mm subfamily during the course of its evolution that does not yet have a binding site for the TF DUX but has a high-confidence predicted binding site for SRF (**Figure 38**). When we looked at the effect of DUX removal on MT2 subfamilies, while we observed that all subfamilies of MT2_Mm were affected, as predicted by the presence of a DUX binding site in all of them, only the two youngest MT2C_Mm subfamilies, MT2C_Mm_viii and _ix, were affected (**Figure 70B**). This is in line with the emergence of a functional DUX binding site through a 9bp deletion within MT2C_Mm_viii (**Figure 38**). These observations suggest that indeed, the subfamilies of MT2C_Mm and MT2_Mm are associated with the occurrence of specific TFBS.

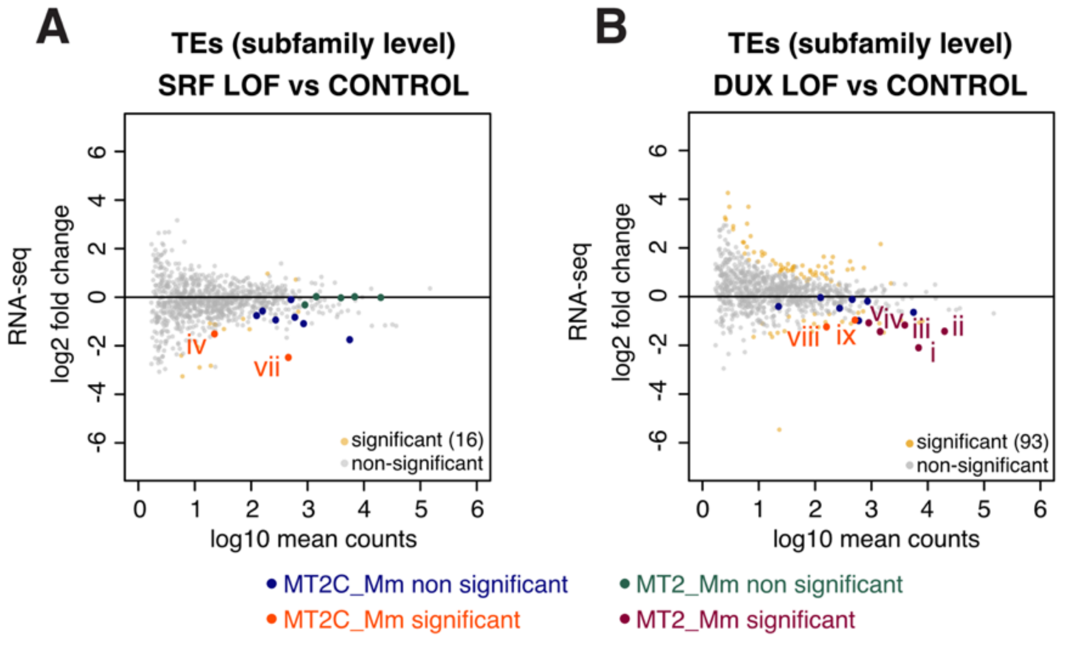


Figure 70. Differential TE expression analysis at the subfamily level of SRF LOF (A) and DUX LOF (B). (A,B) MA plots comparing the log2 fold change in SRF LOF vs CONTROL (A) and DUX LOF vs CONTROL (B) embryos against the log10 RNA-seq mean counts. Differentially expressed TEs (from subfamily DEseq object, see methods) are labelled in orange (padj < 0.05), non-differentially expressed TEs are in gray. Non-significant MT2_Mm and MT2C_Mm subfamilies are labelled in green and blue, respectively. Significant subfamilies of MT2_Mm and MT2C_Mm are indicated with their roman subfamily number in purple and orange, respectively.

To add an additional layer of evidence to the relevance of these findings, we addressed whether MT2C_Mm_vii, which does not have the binding for DUX, is expressed in the embryo using a reporter assay. *This experiment was performed with the help of Dr. Tsunetoshi Nakatani.*

We cloned MT2C_Mm_vii in front of an mRuby gene, and microinjected the plasmid into zygotes, cultured them until the late 2-cell stage and monitored the mRuby reporter fluorescence. We observed that MT2C_Mm_vii could drive transcription of the reporter at the 2-cell stage (**Figure 71**), while no fluorescence was observed in the negative control, without any promoter. As a positive control and as expected, we also cloned the major MT2_Mm subfamily, MT2_Mm_ii which was able to drive transcription of the reporter (**Figure 71**).

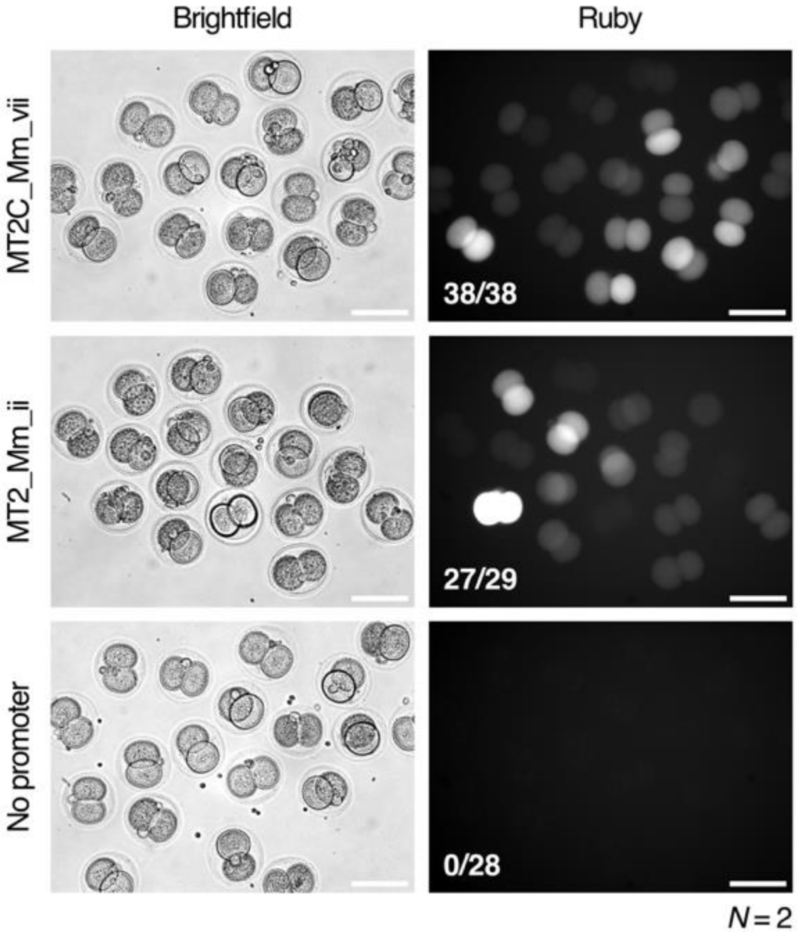


Figure 71. Reporter assay of MT2C_Mm_vii, MT2_Mm_ii and no promoter control in embryos. Representative brightfield (left) and Ruby (right) images of late 2-cell stage embryos after zygotic microinjection with the indicated reporter plasmid. Numbers indicate the total number of embryos with fluorescent signal observed at the late 2-cell stage. *N*, number of independent replicates. Scale bars, 200µm.

We conclude that MT2C_Mm_vii, which contains a TFBS for SRF but not for DUX, is transcriptionally active *in-vivo*, and regulated by SRF. Collectively, this suggests a shift in the dependency on different TFs during MT2 evolutionary trajectory. The transition from SRF to DUX dependency over time supports the hypothesis that the successive acquisition of activating TFBS contributed to waves of TE expansions across the genome. Our findings, therefore, suggest that SRF and DUX jointly orchestrate the regulation of MT2 expression at the late 2-cell stage.

MT2_Mm is known to function as a potent promoter for adjacent host genes, resulting in the formation of chimeric LTR-host gene transcripts and contributing to the activation of a portion of the 2-cell transcriptional program (Peaston et al. 2004; Evsikov et al. 2004; Macfarlan et al. 2012; Franke et al. 2017). Given that we have demonstrated that MT2C_Mm can drive transcription in embryos, and is regulated by SRF, we aimed to explore whether MT2C_Mm transcription is relevant to host biology by initiating genic transcription. As the RNA-seq protocol we used captures TSSs, we searched for chimeric TE-host gene transcripts initiated at TEs and extending into genes, using a tool called ChimeraTE (Oliveira et al. 2023). We quantified chimeric TE-gene transcripts in CONTROL, SRF LOF and DUX LOF embryos. In control embryos, chimeric TE-genes from all MT2C_Mm subfamilies were found, totaling 64 TE-gene events initiated in MT2C_Mm (**Figure 72**). Following SRF LOF, the number of chimeric TE-gene events decreased across all MT2C_Mm subfamilies, though the extent varied between subfamilies (**Figure 72**).

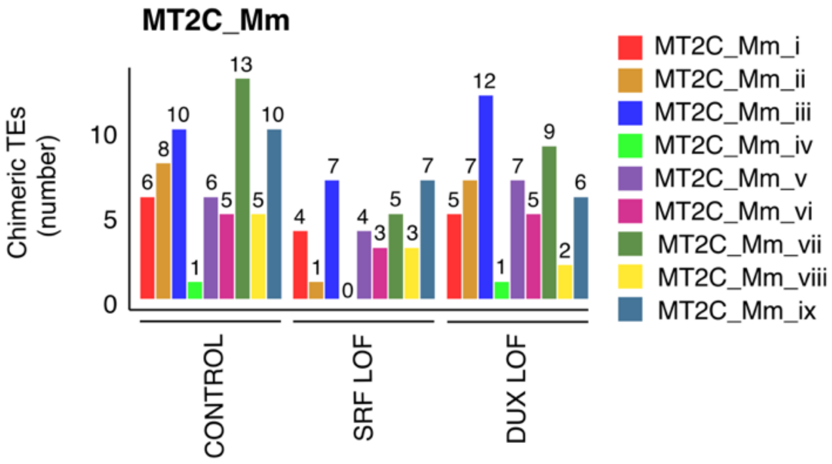


Figure 72. MT2C_Mm-derived chimeric TE-genes. Bar plot representing the number of chimeric TE-genes found in CONTROL, SRF LOF and DUX LOF organized according to the MT2C_Mm subfamily. Numbers are the chimeric TE-gene events from each subfamily in each condition.

We also identified chimeric TE-genes derived from all MT2_Mm subfamilies, with a notable predominance of MT2_Mm_i and MT2_Mm_ii, the two major subfamilies, resulting in a total of 134 chimeric events (**Figure 73**).

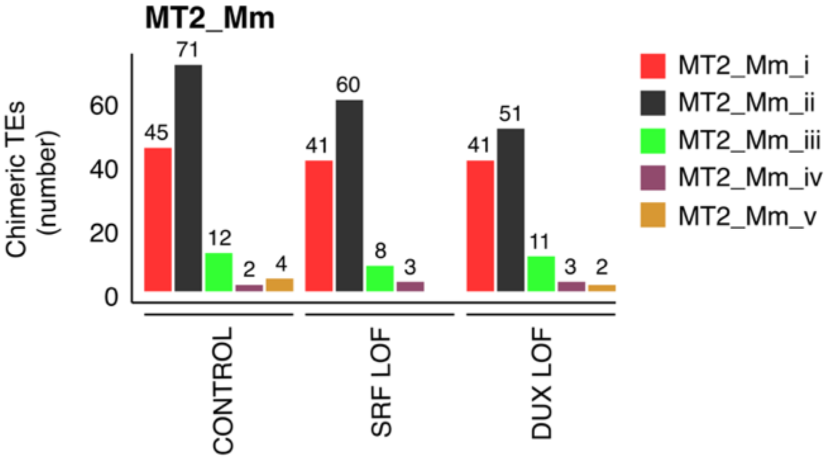


Figure 73. MT2_Mm-derived chimeric TE-genes. Bar plot representing the number of chimeric TE-genes found in CONTROL, SRF LOF and DUX LOF organized according to the MT2_Mm subfamily. Numbers are the chimeric TE-gene events from each subfamily in each condition.

To investigate the evolutionary shift in dependency between SRF and DUX in regulating chimeric TE-genes, we analyzed the expression levels of chimeric TE-genes identified in control embryos, categorized by TE subfamilies, in CONTROL, SRF LOF and DUX LOF embryos. We observed that MT2C_Mm_vii-driven chimeric transcript expression was primarily affected by SRF LOF (**Figure 74**). In contrast, MT2C_Mm_viii-driven transcript expression showed reduction upon both SRF and DUX LOF, with SRF having a slightly stronger effect (**Figure 74**). For MT2C_Mm_ix-driven transcript expression, SRF removal had a weaker impact compared to DUX removal (**Figure 74**). This effect was similarly observed on MT2_Mm_i-driven transcripts (**Figure 74**). Lastly, MT2_Mm_ii-driven transcripts were exclusively affected by DUX depletion (**Figure 74**). Thus, we identified a marked shift in the dependency of chimeric TE-gene expression on TFs along the MT2 evolutionary path. The emergence of the DUX binding site led to a transition, with DUX assuming a more dominant regulatory role over SRF.

After demonstrating that MT2C_Mm drives host gene expression via chimeric transcripts in embryos, and that SRF loss alters their expression, we investigated the promoter activity of these transcripts. We found that, in control 2-cell stage embryos, transcription of these chimeric TE-genes initiates within MT2C_Mm. For instance, in control embryos, the canonical promoter of *Borcs7* is used less frequently than the MT2C_Mm_vii alternative promoter (**Figure 75**). Remarkably, in SRF LOF embryos, the MT2C_Mm_vii promoter is nearly entirely silenced and that correlates with decreased *Borcs7* expression (**Figure 75**). SRF LOF overall either reduces the use of alternative promoters, of shifts expression to the canonical gene promoter. Two other examples of this are provided in **Appendix 3**.

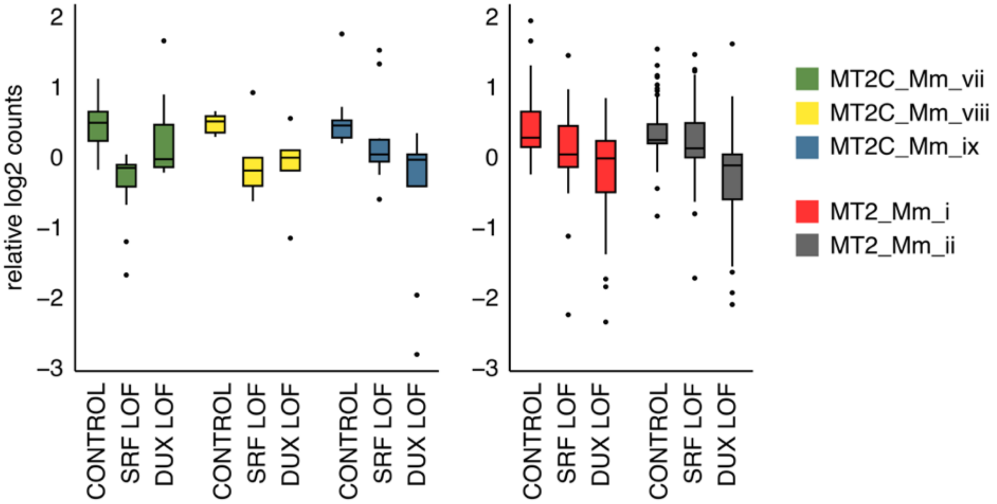


Figure 74. Expression of chimeric TE-gene transcripts. Boxplots showing the median and interquartile range of chimeric TE-gene expression in CONTROL, SRF LOF and DUX LOF by subfamilies. Values for each TE subfamily are the mean log2 transformed normalized counts for each condition centered on each chimeric TE-gene mean across conditions. Whiskers display the highest and lowest values within 1.5 times the interquartile range (IQR). Only chimeric TE-genes found in control embryos are shown.

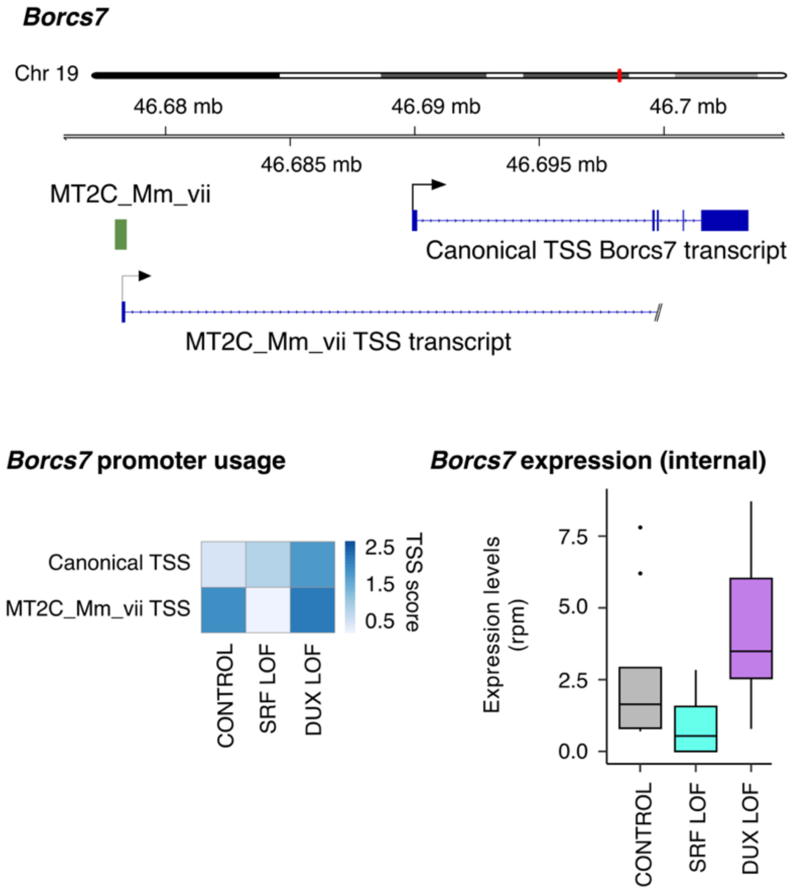


Figure 75. Expression and promoter usage of *Borcs7* gene at the late 2-cell stage. (top) Snapshot of the genomic regions of chromosome 19 containing the canonical *Borcs7* transcript and the alternative, TE-derived transcript (MT2C_Mm_vii-derived). (bottom) Heatmap displaying the TSS scores of both canonical and MT2C_Mm_vii-derived TSSs in CONTROL, SRF LOF and DUX LOF embryos. The boxplot shows the expression of *Borcs7*, as the median rpm and the interquartile range of *Borsc7* transcripts from internal sequencing reads. Whiskers display the highest and lowest value within 1.5 times the IQR.

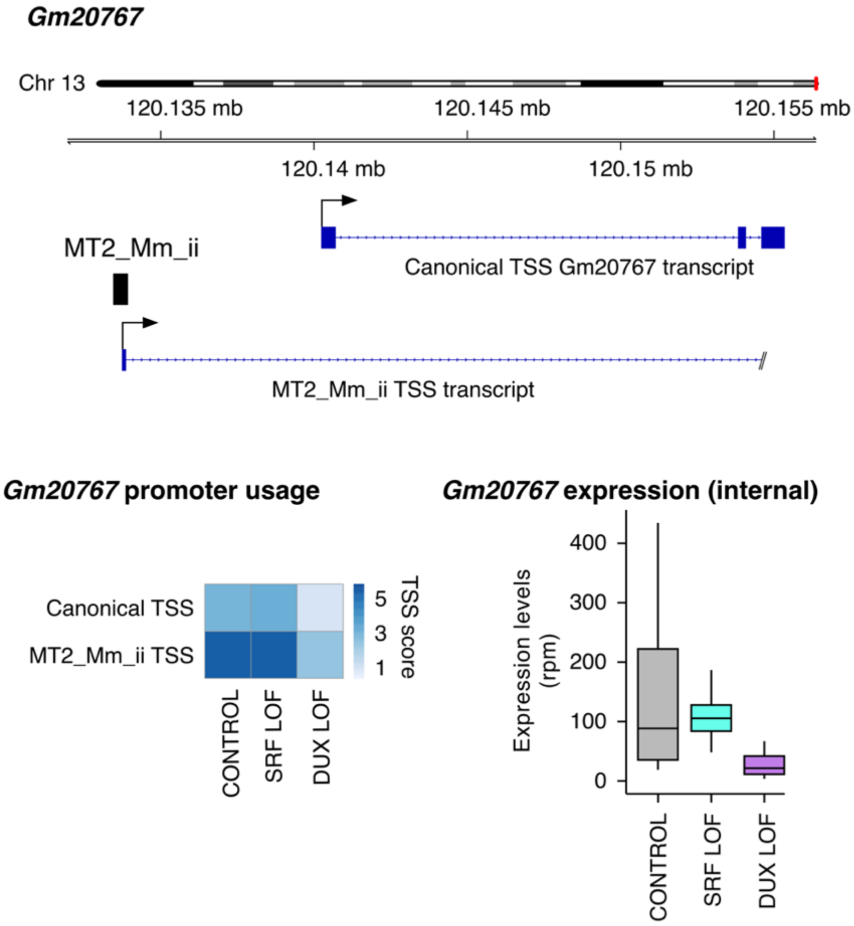


Figure 76. Expression and promoter usage of *Gm20767* gene at the late 2-cell stage. (top) Snapshot of the genomic regions of chromosome 13 containing the canonical *Gm20767* transcript and the alternative, TE-derived transcript (MT2_Mm_ii-derived). (bottom) Heatmap displaying the TSS scores of both canonical and MT2_Mm_ii-derived TSSs in CONTROL, SRF LOF and DUX LOF embryos. The boxplot shows the expression of *Gm20767*, as the median rpm and the interquartile range of *Gm20767* transcripts from internal sequencing reads. Whiskers display the highest and lowest value within 1.5 times the IQR.

Similarly, we observed MT2_Mm promoter usage over canonical in control embryos, for example from an insertion belonging to subfamily MT2_Mm_ii driving expression of *Gm20767*. DUX LOF results in strong reduction of the MT2_Mm_ii-derived promoter usage and consequently, decrease in *Gm20767* expression (**Figure 76**).

In summary, we have identified SRF as a novel regulator of MERVL at the 2-cell stage. SRF and DUX modulate host genome expression, in part by influencing chimeric transcript levels. Both MT2_Mm and MT2C_Mm serve as platforms for alternative promoter activity, regulated by DUX and SRF. These TFs interact with specific MT2 subfamilies within a regulatory framework shaped by successive waves of genomic expansion and evolutionary pressures experienced by TEs.

4.5.2 TBP regulates genes and MaLR during ZGA

4.5.2.1 Establishing Trim-Away for TBP and FOXJ3 in the embryo

Finally, I aimed to investigate the role of TBP and FOXJ3 in the regulation of genes and TEs during early embryonic development. Both TBP and FOXJ3 are maternally inherited proteins (**Figure 58C, D**), hence RNA-based depletion was unsuitable. Instead, I employed an acute protein depletion technique called Trim-Away. This antibody-based method, makes use of exogenous Trim21 to target and degrade the protein of interest (Clift et al. 2017). Trim-Away was shown to achieve rapid protein depletion, making it ideal for use in embryos (Clift et al. 2017). I microinjected early zygotes with antibodies against TBP or FOXJ3 (along with their respective IgG controls), as well as dextran for injection control. After culturing the embryos for approximately 4 hours, I performed a second microinjection of mRNA encoding Trim21-mCherry (**Figure 77**). The embryos were cultured until the late 2-cell stage, then collected for single-embryo RNA sequencing, using the protocol for TSS capture (Oomen et al. 2025) (**Figure 77**).

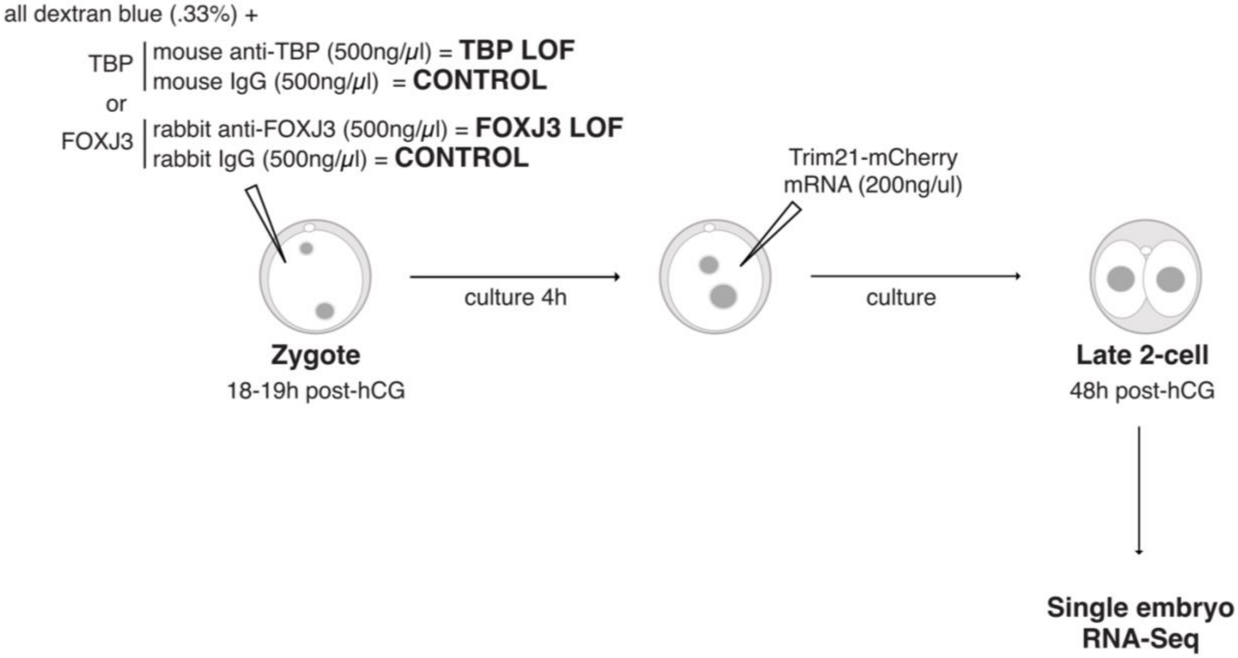

all dextran blue (.33%) +
TBP
mouse anti-TBP (500ng/μl) = TBP LOF
mouse IgG (500ng/μl) = CONTROL
or
FOXJ3
rabbit anti-FOXJ3 (500ng/μl) = FOXJ3 LOF
rabbit IgG (500ng/μl) = CONTROL
Trim21-mCherry
mRNA (200ng/ul)
culture 4h
culture
Zygote
18-19h post-hCG
Late 2-cell
48h post-hCG
Single embryo
RNA-Seq

Figure 77. Schematic representation of the experimental design used for FOXJ3 LOF and TBP LOF in the early embryo followed by single embryo RNA-seq at the late 2-cell stage.

I first verified the experimental setup and fixed embryos at the late 2-cell stage (same timing as for RNA-sequencing collection) (**Figure 77**). Immunofluorescence confirmed the nuclear depletion of both TBP and FOXJ3 at the late 2-cell stage, following Trim-Away (**Figure 78**).

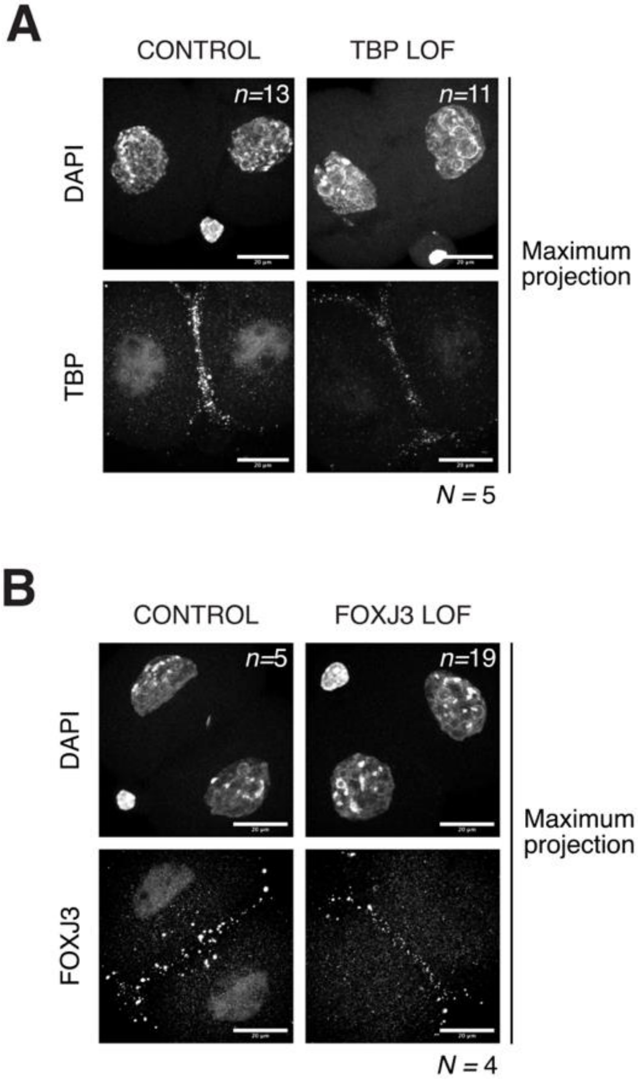


Figure 78. Trim-Away efficiency controls at the late 2-cell stage. (A,B) Representative images of immunofluorescence of TBP (A) and FOXJ3 (B) following Trim-Away. All images are maximum intensity projections. *n* is the total number of embryos analyzed per condition. *N*, number of independent replicates. Scale bars, 20μm.

4.5.2.2 Loss of TBP disrupts ZGA and delays development

I collected a total of 10 TBP LOF and 11 controls, as well as 17 FOXJ3 LOF embryos and 13 controls. Two embryos did not meet general quality control standards (**Figure 79A, B**), including one control for TBP LOF embryo and one FOXJ3 LOF embryo. Consequently, we proceeded with a total of 49 embryos (10+10+16+13) for further analysis (**Figure 79A, B**).

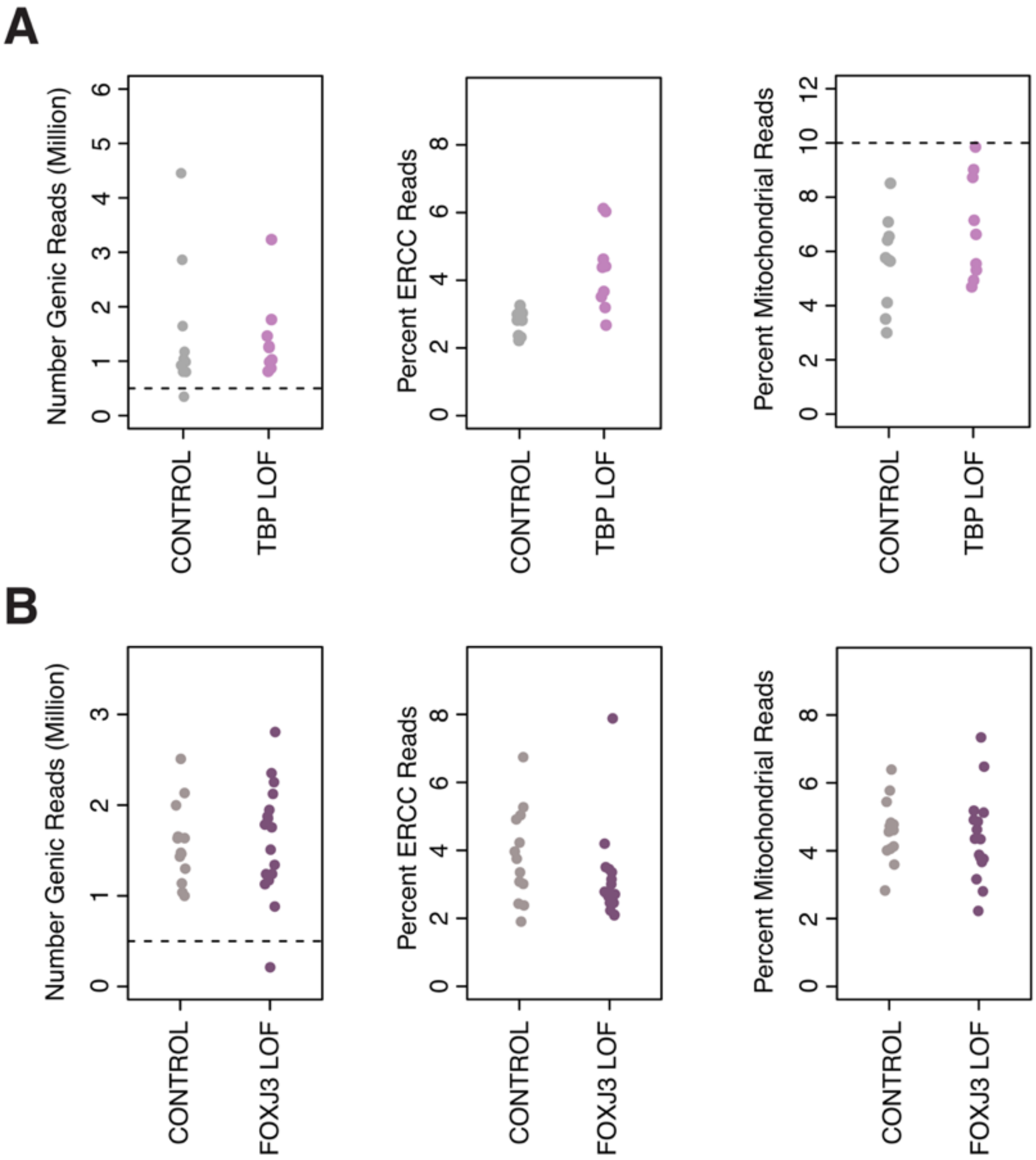


Figure 79. Single embryo RNA seq quality control for TBP LOF (A) and FOXJ3 LOF (B). (A,B) Quality control dotplots showing the number of genic reads (Million), the percent of ERCC reads and of Mitochondrial reads in all experimental conditions as indicated. Each dot is an embryo. Dashed lines represent the threshold applied for quality control filtering.

Surprisingly, differential gene expression analysis after FOXJ3 depletion showed minimal effect on gene expression (**Figure 80A**), with only four genes (*Letm1*, *Ino80b*, *Gm20274, 2310039H08Rik*) exhibiting significant changes. This suggests that FOXJ3 is largely dispensable for gene expression during ZGA. In stark contrast, TBP depletion caused a profound effect on the 2-cell transcriptome (**Figure 80A**), with 5861 genes significantly affected, 96% of which were downregulated, underscoring TBP as a critical regulator of transcription during ZGA. These results are consistent with the central role of TBP in initiating transcription and functioning as a general TF (Butler and Kadonaga 2002). TBP forms part of the multisubunit complex general transcription factor (GTF) TFIID, which, along with at least 13 conserved TBP-associated factors (TAFs), engages gene promoters (Reinberg et al. 1987; Buratowski et al. 1989; Dynlacht et al. 1991; Poon and Weil 1993). This event triggers a sequential recruitment of additional GTFs, eventually leading to the attachment of RNAPII (Reinberg et al. 1987; Buratowski et al. 1989, 1991).

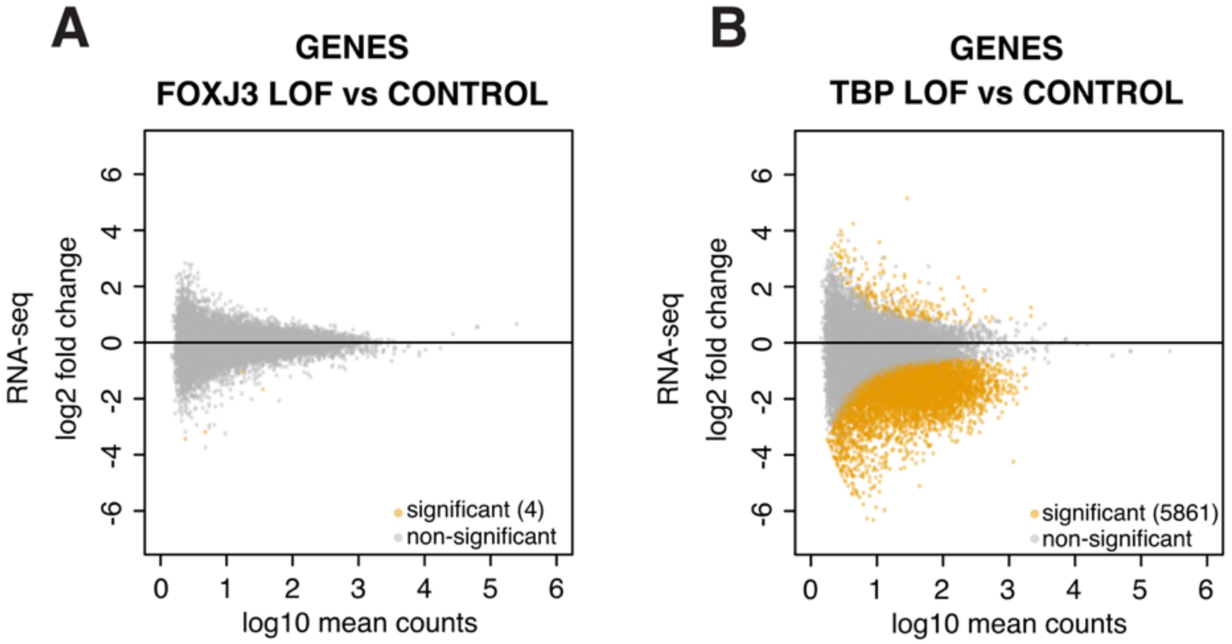


Figure 80. Differential gene expression analysis of FOXJ3 LOF (A) and TBP LOF (B). (A,B) MA plots comparing the log2 fold change in FOXJ3 LOF vs CONTROL (A) and TBP LOF vs CONTROL (B) embryos against the log10 RNA-seq mean counts. Differentially expressed genes are labelled in orange (padj < 0.05), non-differentially expressed genes are in gray.

A knockout mouse model from the early 21st century demonstrated that embryos deficient in TBP developed normally until the blastocyst stage, at which point they ceased to grow and underwent apoptosis (Martianov et al. 2002). However, the absence of maternal knockout prevented from concluding on the role of TBP in the first few hours following fertilization. Our data indicated that TBP is important for transcription in late 2-cell stage embryos. Therefore, we sought to further investigate the affected genes, focusing particularly on those that were downregulated. Using the DBTMEE classification system (Park et al. 2015), we found that approximately 50% of the genes annotated as major ZGA gene were significantly downregulated in TBP LOF, indicating that TBP impaired major ZGA (**Figure 81**). Interestingly, upon closer examination of the genes affected by TBP loss, particularly on typical “2C” markers, including the DUX target genes highlighted in Figure 64, none appeared to be affected by TBP depletion (**Appendix 4**).

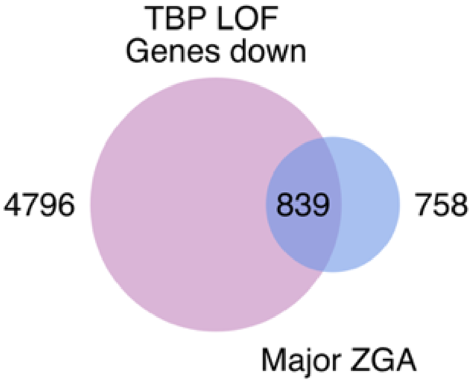


Figure 81. Overlap of genes downregulated in TBP LOF with major ZGA genes. Venn diagram showing the overlap between downregulated genes upon TBP LOF relative to CONTROL embryos (padj < 0.05) and major ZGA genes as defined by the DBTMEE database (Park et al. 2015).

CDK9, a key component of pTEFb, serves an important regulator of transcription elongation by modulating the activity of negative elongation factors (NELF) (Chen et al. 2018) and phosphorylating Ser 2 residue of RNAPII C-terminal domain (CTD) and SPT5 subunit of DRB Sensitivity inducing factor (DSIF) (Kwak and Lis 2013). In line with a major role for TBP in regulating RNAPII-mediated transcription, we observed an extensive gene overlap between those downregulated by CDK9 inhibition or SPT5 depletion, and those affected by TBP depletion (60% and 79%, respectively) (**Figure 82**) (Abe et al. 2022). Thus, similarly to CDK9 and SPT5 (Abe et al. 2022), TBP plays a key role in regulating global transcription during ZGA.

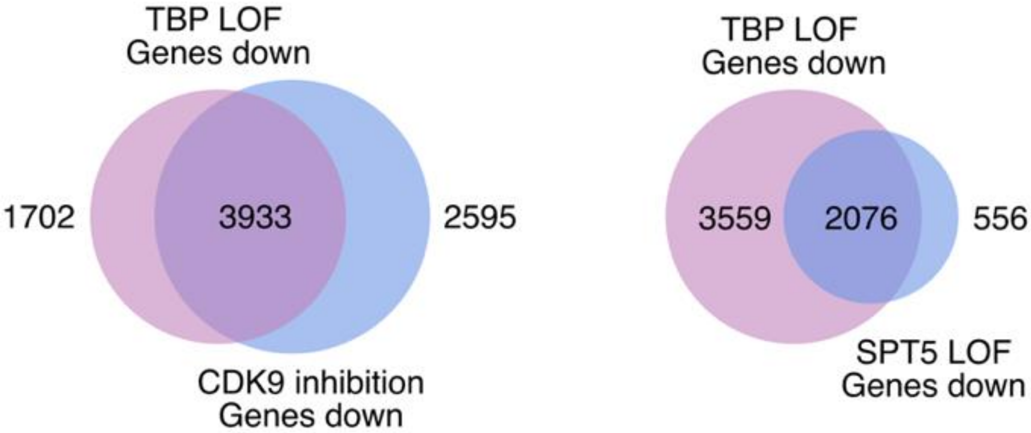


Figure 82. Overlap of genes downregulated in TBP LOF and genes affected by CDK9 and SPT5 disruptions. Venn diagrams showing the overlap between downregulated genes upon TBP LOF relative to CONTROL embryos (padj < 0.05) and genes downregulated in CDK9 inhibition relative to control (left) or upon SPT5 LOF relative to control (right) (Abe et al. 2022).

We then conducted a principal component analysis (PCA) including CONTROL and TBP LOF embryos, alongside non-manipulated RNA-seq dataset from embryos at various preimplantation development stages, generated with the same protocol (**Figure 83**) (Oomen et al. 2025).

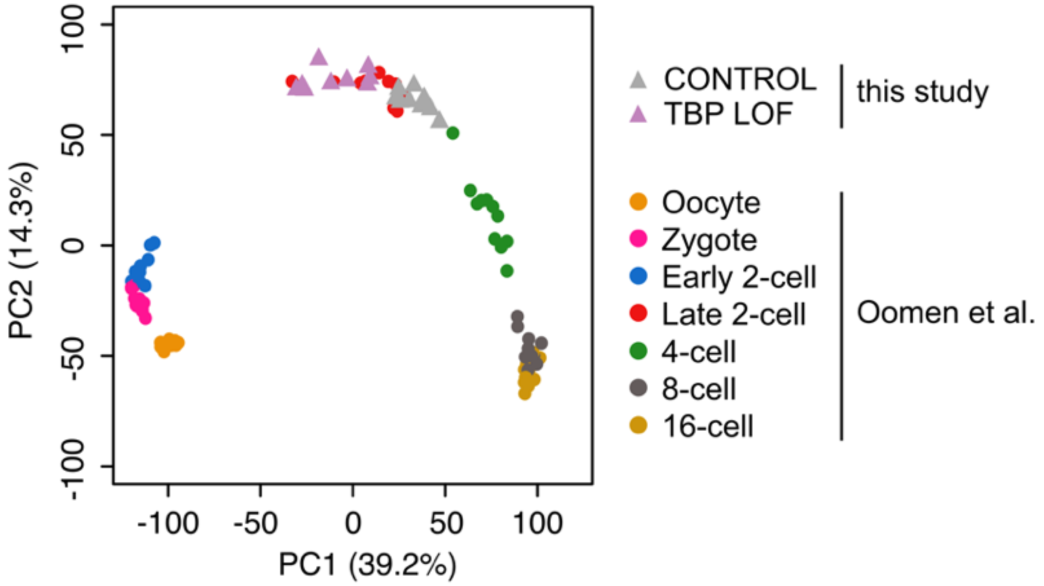


Figure 83. Principle component analysis of CONTROL, TBP LOF together with embryos from oocyte to 16-cell stage from (Oomen et al. 2025). Triangles are embryos from this study, color-coded based on the condition. Circles are embryos from Oomen et al. (2025), color-coded based on the stage. The proportions of variance explained by PC1 and PC2 are indicated.

The PCA revealed a slight shift in TBP embryos compared to control embryos along the PC1 axis, suggesting a developmental delay (**Figure 83**). This led us to investigate whether these transcriptional changes affected developmental progression, specifically whether TBP is essential for proper development. To address this, we microinjected embryos with the Trim-Away system for TBP and its IgG control, and monitored development up to the blastocyst stage (*which was performed with the help of Mrinmoy Pal)* (**Figure 84**). Cell division between the 4- to 8-cell stages, as well as blastocyst formation, were delayed (**Figure 84**). Despite significant transcriptomic changes, the embryos were not arrested at the 2-cell stage and successfully underwent the second cell division. Our observed phenotype sharply contrasted with the penetrance of 2-cell arrest induced by SPT5 LOF, despite substantial gene overlap (**Figures 82, 84**) (Abe et al. 2022). Although the transient effect of protein depletion by Trim-Away cannot be entirely excluded, these results suggested that the 2-cell arrest phenotype may be due to specific downregulated genes rather than a global transcriptional defect.

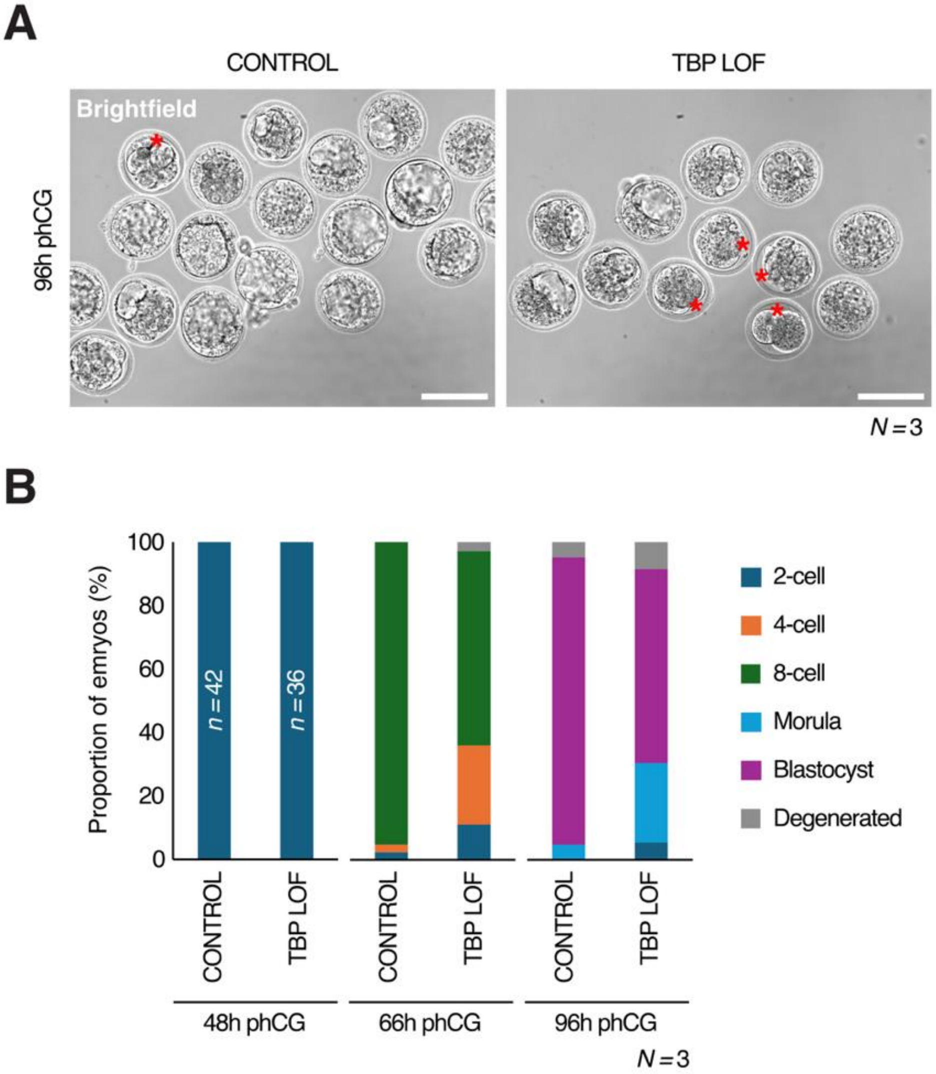


Figure 84. Developmental progression of CONTROL and TBP LOF embryos. (A) Brightfield images of CONTROL (left) and TBP LOF (right) embryos at 96h phCG. Red asterisks highlight delayed embryos. *N,* number of independent replicates. Scale bars, 100 µm. (B) Proportion of control and TBP LOF embryos in specific stage as indicated by the color code at 48h, 66h and 96h phCG. *N,* number of independent replicates. The numbers indicated in the bars at 48h represent the total number of embryos analyzed over the different independent replicates.

Finally, we sought to determine whether the genes affected by TBP loss were directly regulated by TBP, or if these changes were indirect. To do so, I performed CUT&Tag for TBP in late 2-cell stage embryos, with four replicates for TBP and two IgG control replicates*.* Upon analyzing each replicate individually, we found that while the first replicate showed better signal and coverage than the others, the four replicates were largely concordant (**Appendix 5**). Therefore we combined all the reads for further analysis. We found that most of the downregulated genes were indeed bound by TBP, suggesting direct transcriptional regulation (**Figure 85**).

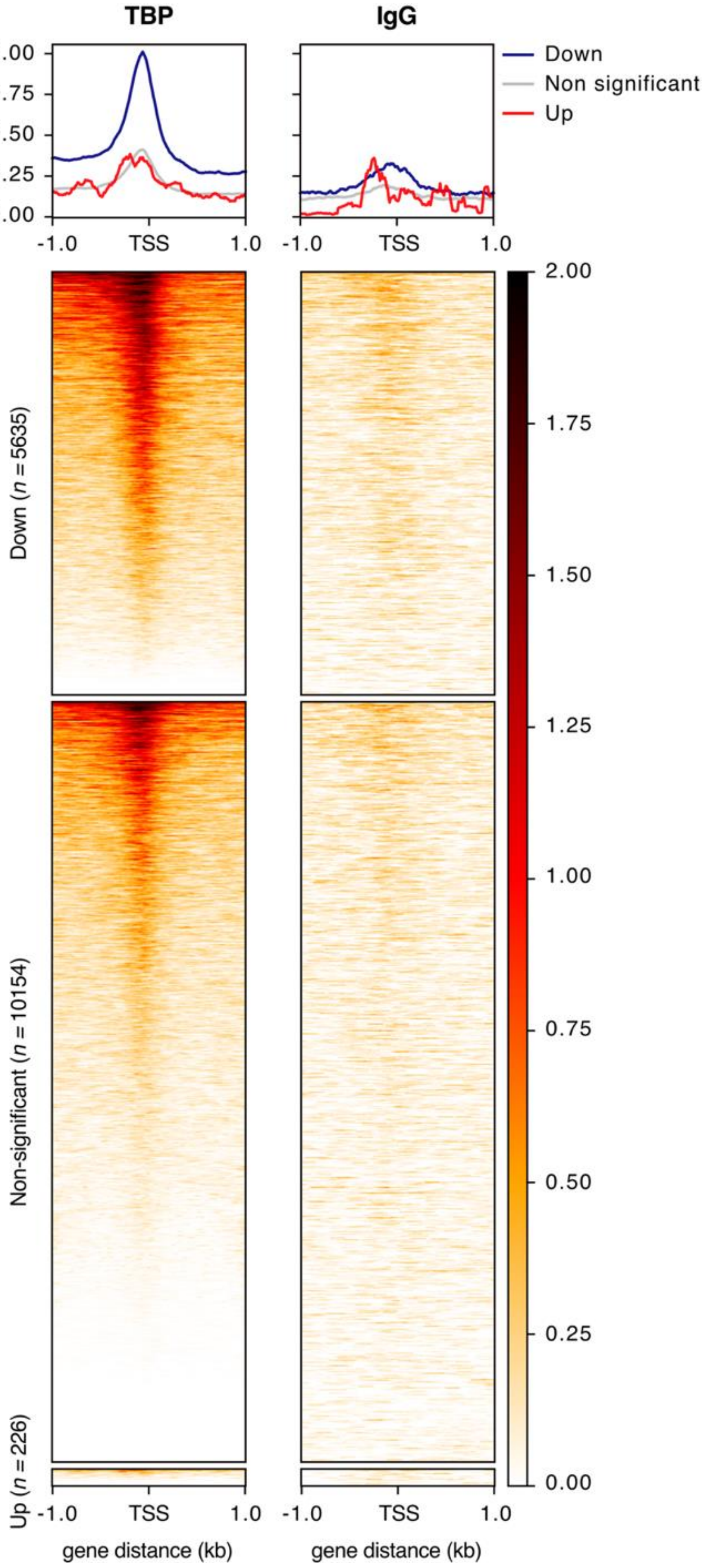


Figure 85. TBP enrichment at genes in 2-cell stage embryos. Signal aggregate plots and heatmaps of TBP enrichment (left) and IgG control (right) from CUT&Tag reads pooled from the different biological replicates (*N* = 4) over the TSS of down-regulated (padj < 0.05), non-significant and up-regulated (padj < 0.05) genes upon TBP LOF relative to CONTROL embryos. *n* is the number of genes per indicated category.

Importantly, the enrichment of TBP to these genes was clear compared to IgG control (**Figure 85**). We also observed that some genes which were not transcriptionally affected by TBP removal, were also bound by TBP (**Figure 85**). This could be due to residual TBP binding following Trim-Away, or because TBP binds to these genes without regulating their transcription. On the other hand, the upregulated genes were not bound by TBP (**Figure 85**), suggesting that these effects are largely indirect, potentially resulting from the slight developmental delay observed.

In conclusion, FOXJ3 appears to be non-essential for ZGA, whereas TBP directly regulates nearly half of major ZGA genes, and deregulation of these targets leads to a developmental delay but does not prevent blastocyst formation.

4.5.2.3 TBP directly regulates MaLR in the mouse early embryo

Ultimately, we investigated whether FOXJ3 and TBP are regulating TEs at the late 2-cell stage. FOXJ3 removal had virtually no impact on TE expression, as no TE family was significantly differentially regulated upon its depletion (**Figure 86A**). Notwithstanding, 278 TE families were significantly differentially expressed upon TBP loss (**Figure 86B**). All except one family of DNA transposons (Eulor6D) were significantly downregulated, suggesting that TBP acts mostly as a transcriptional activator of TEs in the embryo. Several TE families were downregulated, including evolutionary old families of L1 (L1MEa, L1MCb, L1M8) and some SINEB1 elements (B1Mus1, B1F, B1_Mur1, B1_Mm) (**data not shown**). Among the TEs of interest, SINEB2 B3A family was significantly downregulated (**Figure 86B**). Consistent with our targeted GOF screen results in ESCs (**Figure 43**), ORR1A0 and ORR1A1 were significantly downregulated upon TBP depletion in embryos. This effect extended beyond ORR1A0 and ORR1A1, affecting other MaLR ORR families, including ORR1A2 and ORR1A3 (**data not shown**). Notably, the downregulation appeared specific to MaLR ORR elements, as MT2_Mm remained unchanged (**Figure 86B**), and MaLR elements from the MT family such as MTA_Mm and MTB_Mm showed no significant changes (**data not shown**).

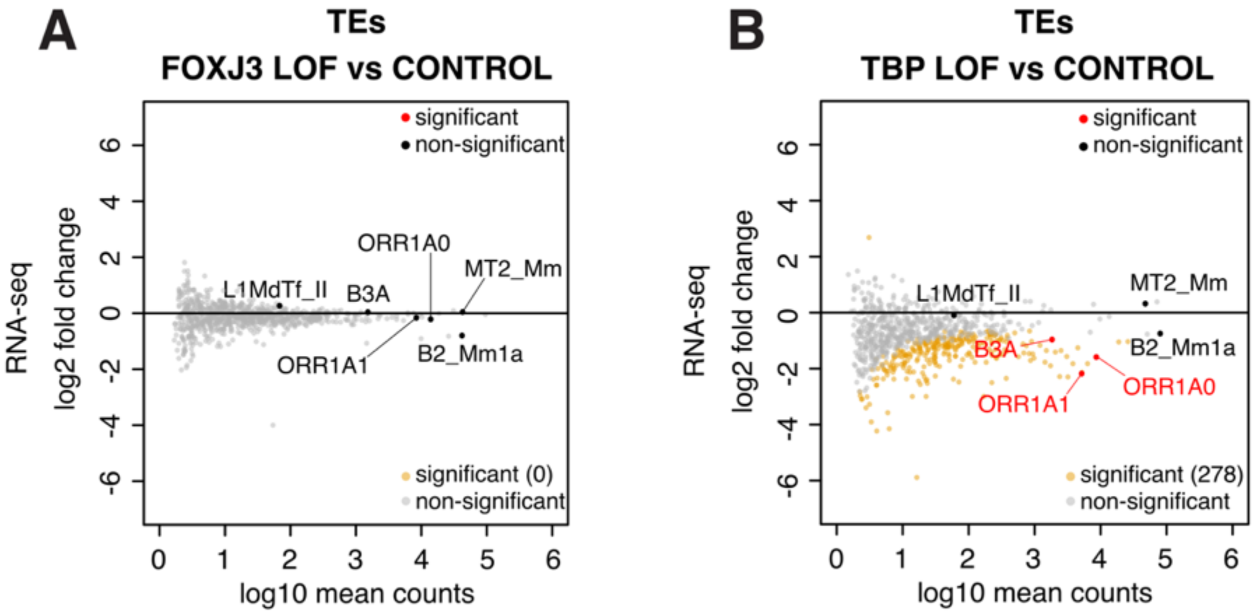


Figure 86. Differential TE expression analysis of FOXJ3 LOF (A) and TBP LOF (B). (A,B) MA plots comparing the log2 fold change in FOXJ3 LOF vs CONTROL (A) and TBP LOF vs CONTROL (B) embryos against the log10 RNA-seq mean counts. Differentially expressed TEs are labelled in orange (padj < 0.05), non-differentially expressed genes are in gray. The six TE families of interest in this project (all insertions included) are labelled, significant in red, non-significant in black.

We next wondered whether TBP was directly regulating the transcription of ORR1A0 and ORR1A1. We analyzed our CUT&Tag data for TBP at the late 2-cell stage to assess whether TBP binds to these elements. Both ORR1A0 and ORR1A1 were found to be bound by TBP, compared to IgG control (**Figure 87**), suggesting that TBP directly binds to and regulates the transcription of these TEs in the embryo. However, when examining the enrichment over MT2_Mm, whose transcription was not affected by TBP depletion, we unexpectedly observed TBP CUT&Tag signal compared to IgG (**Figure 88**).

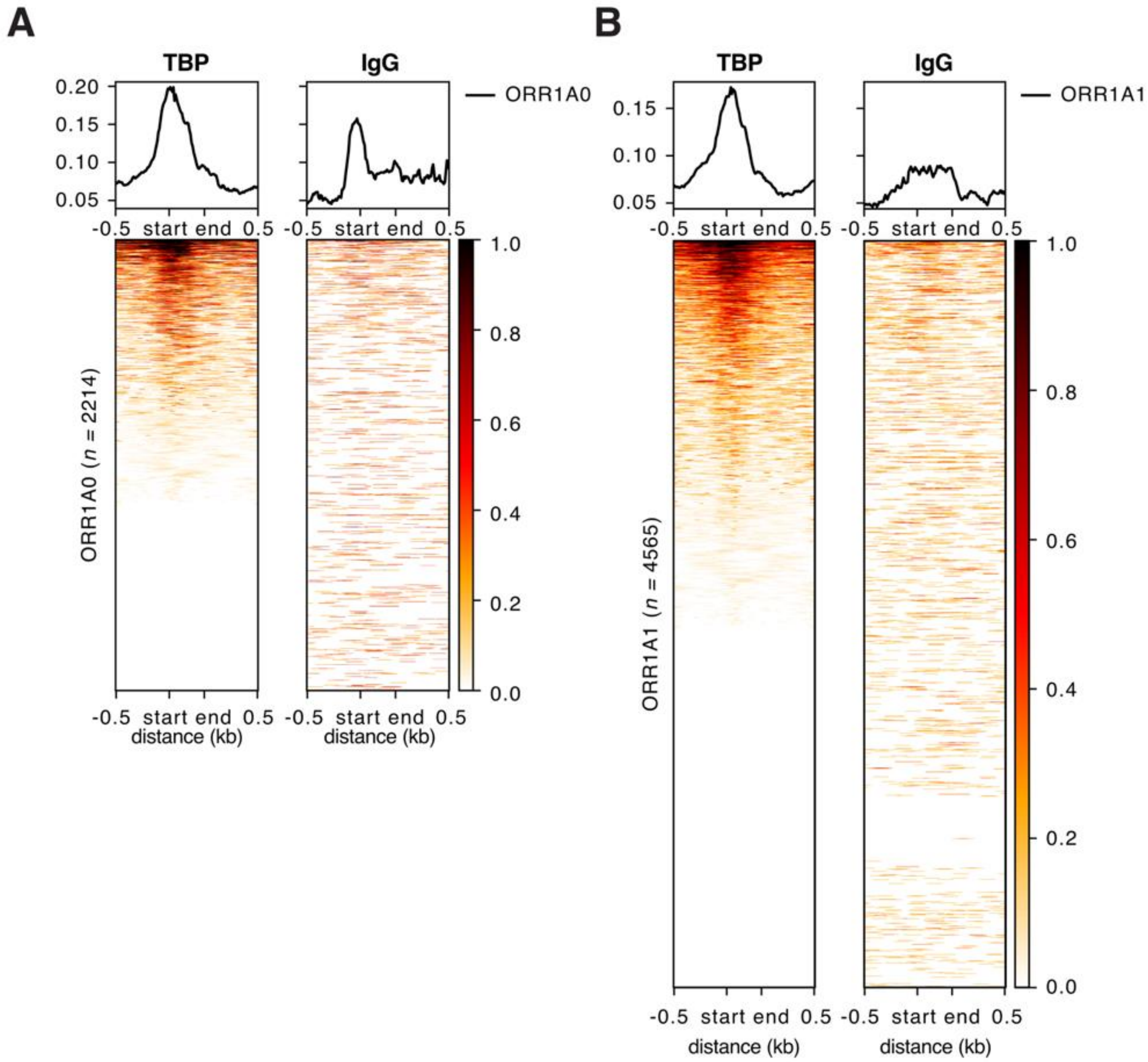


Figure 87. TBP enrichment at ORR1A0 (A) and ORR1A1 (B) insertions in 2-cell stage embryos. (A, B) Signal aggregate plots and heatmaps of TBP enrichment (left) and IgG control (right) from CUT&Tag reads pooled from the different biological replicates (*N* = 4) over all ORR1A0 (A) and ORR1A1 (B) insertions. Start and end refer to the position of the LTR. *n* is the number of insertions per TE family as indicated.

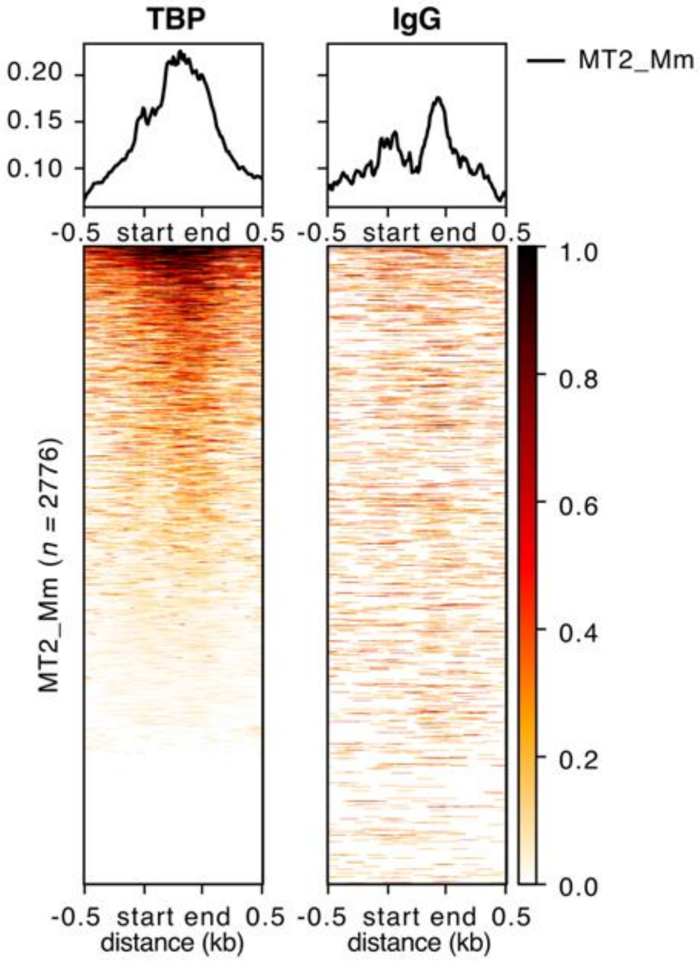


Figure 88. TBP enrichment at MT2_Mm insertions in 2-cell stage embryos. Signal aggregate plots and heatmaps of TBP enrichment (left) and IgG control (right) from CUT&Tag reads pooled from the different biological replicates (*N* = 4) over all MT2_Mm insertions. Start and end refer to the position of the LTR. *n* is the number of MT2_Mm insertions.

We analyzed individual insertion expression using TElocal (Jin et al. 2015). We found that while there are 2214, 4565 and 2776 ORR1A0, ORR1A1 and MT2_Mm insertions within the mouse genome, respectively, 582, 424 and 1841 are considered expressed (See method for threshold). Differential expression analysis revealed that approximately 35% of expressed ORR1A0 insertions were significantly downregulated upon TBP LOF, with two insertions upregulated (**Figure 89A**). Similarly, 50% of expressed ORR1A1 insertions were significantly downregulated (**Figure 89B**). In contrast, only 3% of expressed MT2_Mm insertions were downregulated (**Figure 89C**). We further evaluated TBP binding in relation to the impact of TBP removal on expression. We found that several significantly downregulated ORR1A0 and ORR1A1 insertions were bound by TBP, suggesting that TBP directly activates the transcription of these elements in the embryo (**Figure 90**).

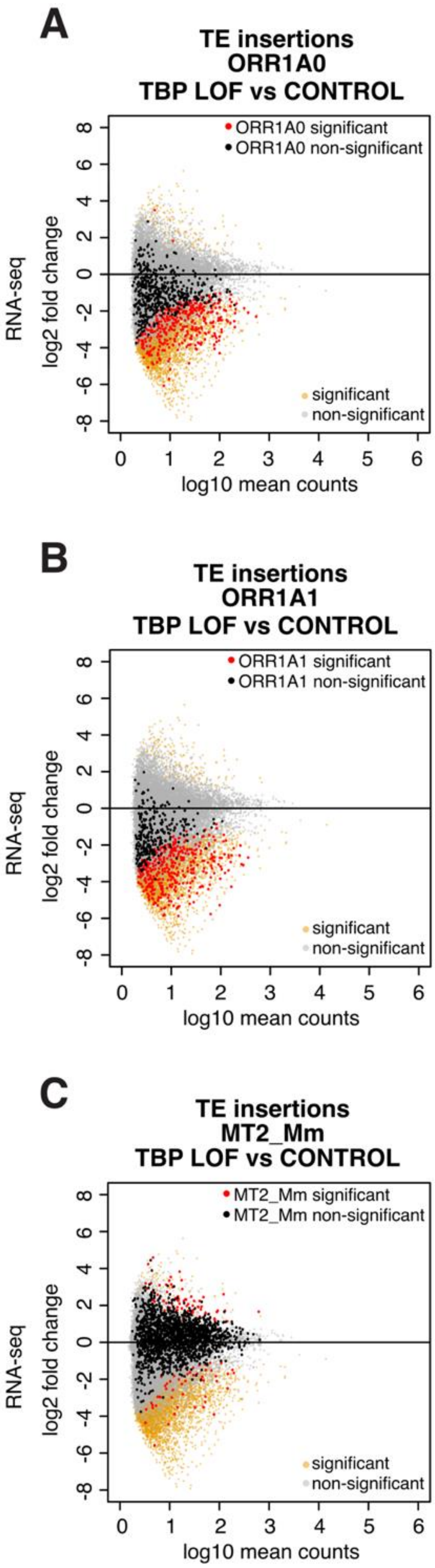


Figure 89. Differential expression analysis of single ORR1A0 (A) ORR1A1 (B) and MT2_Mm (C) insertions upon TBP LOF. (A-C) MA plots comparing log2 fold change in TBP LOF vs CONTROL embryos against log10 RNA-seq mean counts. Differentially expressed individual TE insertions are displayed in orange (padj < 0.05) and non-differentially expressed individual TE insertions are displayed in gray. Non-significant individual ORR1A0 insertions (A), ORR1A1 insertions (B) and MT2_Mm insertions (C) are labelled in black, significantly differentially expressed individual insertions of the same family are labelled in red.

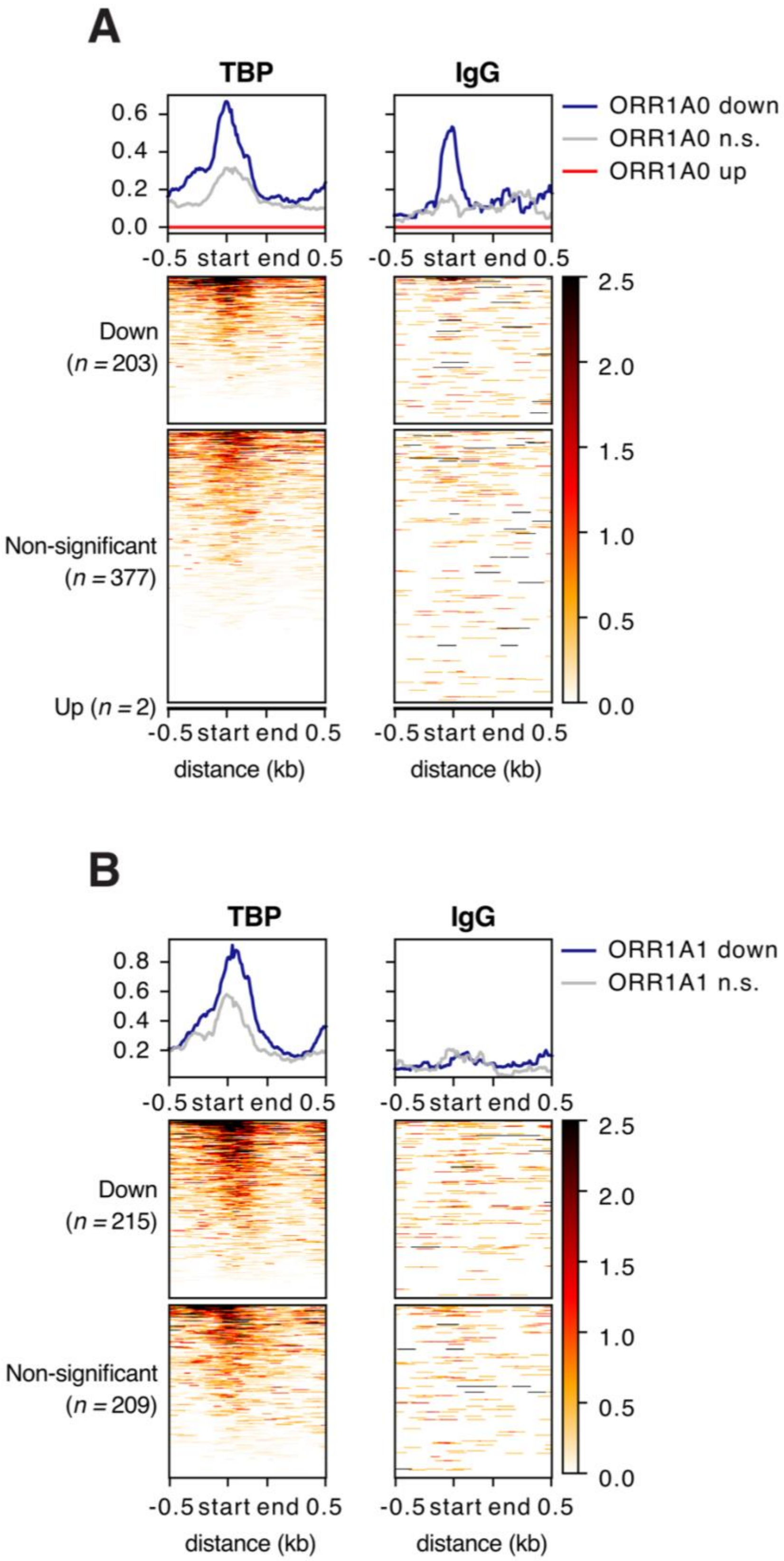


Figure 90. TBP enrichment at expressed ORR1A0 (A) and ORR1A1 (B) insertions in 2-cell stage embryos. (A, B) Signal aggregate plots and heatmaps of TBP enrichment (left) and IgG control (right) from CUT&Tag reads pooled from the different biological replicates (*N* = 4) over the downregulated (padj < 0.05), non-significant and upregulated (padj < 0.05) ORR1A0 (A) and ORR1A1 (B) insertions. Start and end refer to the position of the LTR. *n* is the number of insertions per category as indicated.

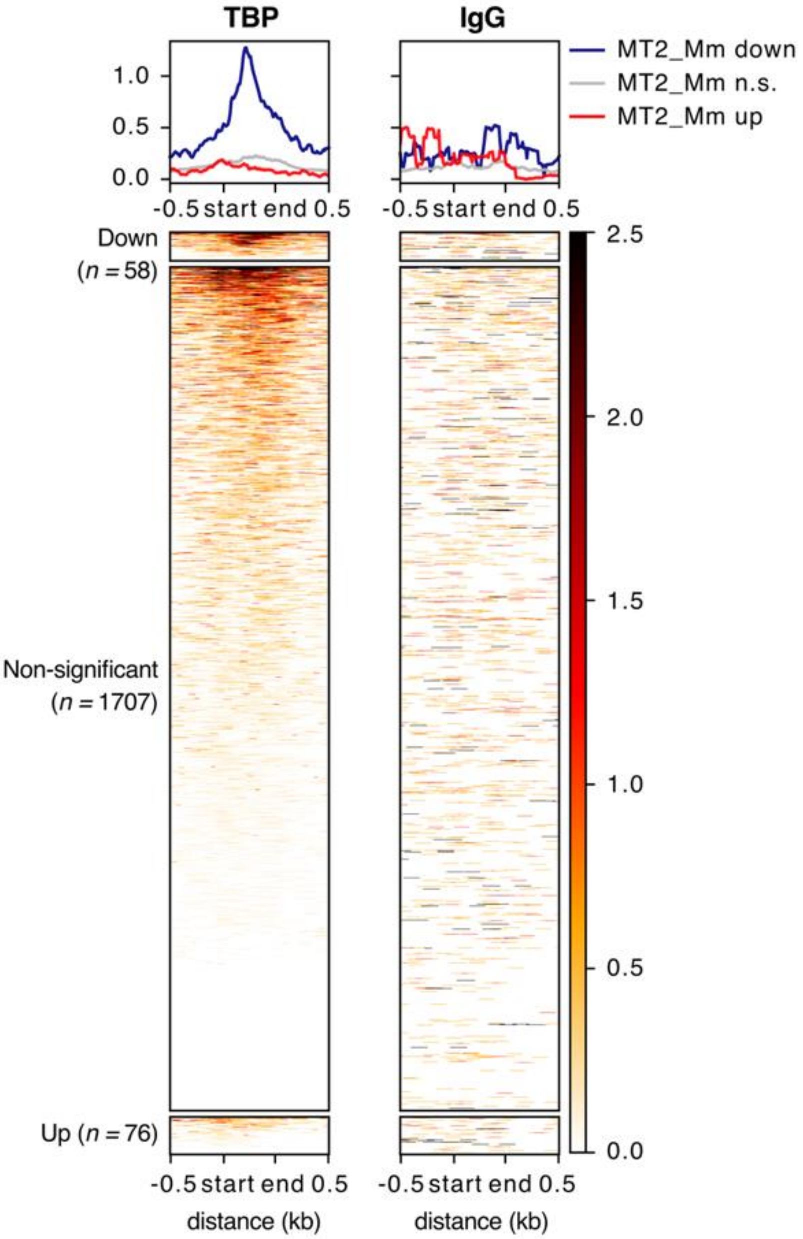


Figure 91. TBP enrichement at expressed MT2_Mm insertions in 2-cell stage embryos. Signal aggregate plots and heatmaps of TBP enrichment (left) and IgG control (right) from CUT&Tag reads pooled from the different biological replicates (*N* = 4) over the downregulated (padj < 0.05), non-significant and upregulated (padj < 0.05) MT2_Mm insertions. Start and end refer to the position of the LTR. *n* is the number of insertions per category as indicated.

Importantly, analyzing MT2_Mm revealed that the insertions bound by TBP corresponded to those significantly downregulated upon TBP removal (**Figure 91**). Therefore, only a small fraction (3%) of MT2_Mm are bound and transcriptionally activated by TBP in the early embryo (**Figure 91**). Upon closer inspection of these 58 MT2_Mm insertions, we observed that nearly 70% (40/58) belonged to the MT2_Mm_i subfamily (**Appendix 6**). This prompted us to investigate TBP binding to the different MT2_Mm subfamilies, and we found that TBP was predominantly bound to MT2_Mm_i, the evolutionary oldest subfamily (**Figure 92**).

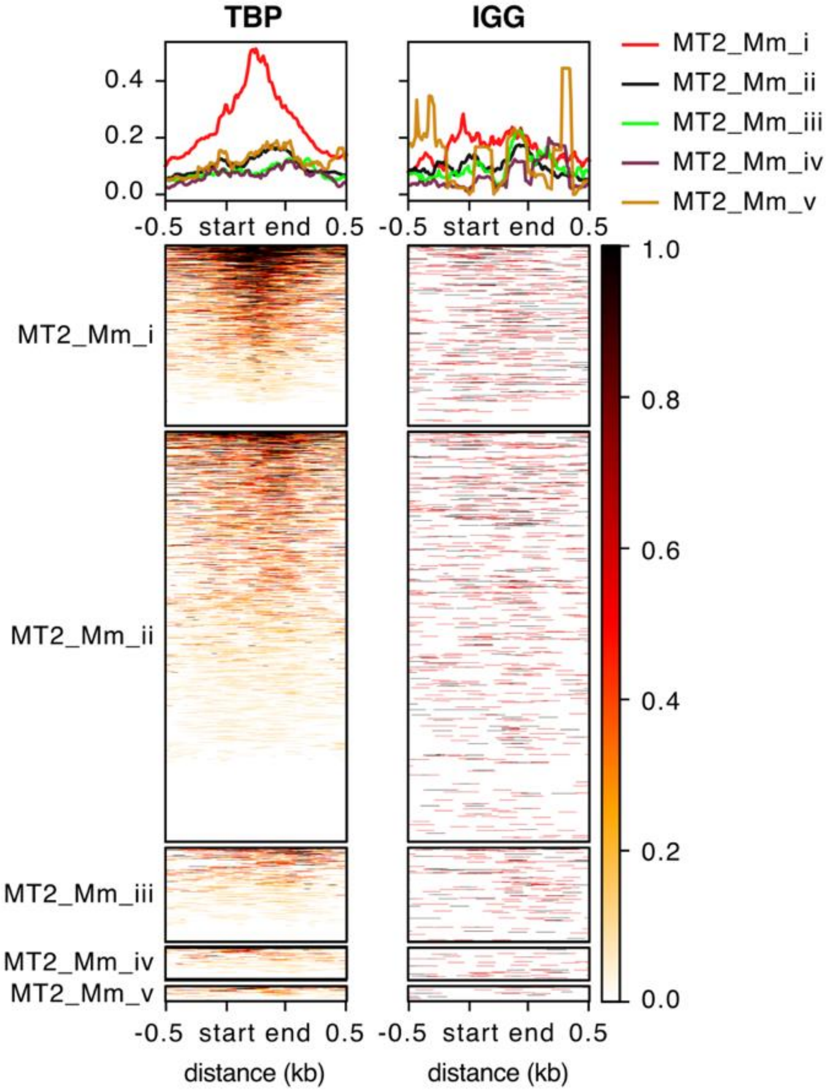


Figure 92. TBP enrichement at MT2_Mm subfamilies in 2-cell stage embryos. Signal aggregate plots and heatmaps of TBP enrichment (left) and IgG control (right) from CUT&Tag reads pooled from the different biological replicates (*N* = 4) over the MT2_Mm subfamilies. Start and end refer to the position of the LTR.

ORR MaLR elements are evolutionarily more ancient than MT2_Mm (Franke et al. 2017). When we performed a similar analysis on ORR1A0 and ORR1A1, examining TBP enrichment across subfamilies defined in the phylogenetic analysis, we found no specific enrichment pattern (**data not shown**). This suggests that TBP regulates MaLR insertions across all subfamilies, irrespective of their phylogenetic lineage.

Collectively, our findings indicate that TBP plays a direct role in regulating ORR MaLR, as well as a subset of evolutionary old MT2_Mm insertions. These results establish TBP as a regulator of specific TE families and subfamilies *in-vivo*, while indicating that its involvement in ZGA is largely independent of MT2_Mm.

5 Discussion

5.1 Overlap in TFBS across evolutionary distinct TEs

The primary objective of this study was to broaden the repertoire of TFs that regulate TE transcription at the onset of mammalian development. Through a footprinting analysis of ATAC-seq data, focusing on six young, rodent-specific TE families, we uncovered 54 novel TFs potentially involved in regulating TE activity in the embryo. The list we have compiled represents TFs whose footprint was detected within at least one of the six TE families and are expressed, at the mRNA level during preimplantation development. While the precise role of most of these TFs in modulating TE expression remains to be further explored, this work provides a foundation with a newly identified set of candidates, offering a starting point for future investigations.

In our footprinting analysis of different TE families from ERVL/MaLR, SINEs, and LINEs superfamilies, we observed a notable overlap of TFs found within their regulatory regions, despite extensive evolutionary distance separating these superfamilies. The overlap between MT2_Mm and MaLR ORR LTRs was expected due to their common ancestry (Franke et al. 2017) and the known tendency of LTRs to accumulate TFBS (Bourque et al. 2008; Feschotte 2008; Ito et al. 2017; Hermant and Torres-Padilla 2021). It was more surprising, though, to observe substantial overlap with LINEs and SINEs as well. This aligns with the hypothesis that strong evolutionary pressure is experienced by TEs to acquire TFBS for TFs expressed during preimplantation development, as it is important for TE fitness to be transcribed (and originally transposed) during these stages (Hermant and Torres-Padilla 2021). It is possible that these TFs regulate multiple classes of TEs, suggesting that these TEs may indeed have evolved similar strategies for amplification and maintenance within the genome.

The effects of overexpressing individual TFs in mESCs were consistent with this concept. Although we cannot exclude indirect effects arising from TF overexpression, we observed that many of the TFs activate transcription of several TE families simultaneously upon overexpression. However, addressing the precise role of these TFs would require further work. While this targeted screen tested multiple TFs concurrently, it did not validate TF function *in-cellulo*. Instead, it provided a basis for selecting candidates for further analyses in the embryo.

5.2 Contrasting TF impacts: *in-cellulo* vs *in-vivo* models

Overexpression of individual TF candidates in mESCs revealed that none match the ability of DUX to activate MERVL. This may reflect the fact that none of these TFs possess pioneer activity, as described for DUX (Choi et al. 2016; Eidahl et al. 2016).

Further, we observed notable differences between the effects of TF overexpression in mESCs and TF removal in embryos. For instance, overexpression of TBP mildly induces MaLR expression in mESCs, while TBP removal in embryos strongly impacts MaLR expression. Embryos exhibit a unique chromatin structure that may facilitate TF binding to TEs and resulting transcriptional activation (Burton and Torres-Padilla 2014). In comparison, mESCs have a more compacted chromatin structure, perhaps limiting TF-mediated activation, particularly when the TF does not have the ability to act as a pioneer TF. This was also observed in the case of YY1, which has been shown to bind to and activate mouse L1 transcription in embryos (Sakamoto and Ishiuchi 2024), but only in mESCs with DNA demethylation disruption and upon HDAC inhibitions (Cusack et al. 2020), conditions that loosen chromatin structure. In contrast, FOXJ3 shows the strongest effect on L1 transcription upon overexpression in mESCs, whereas FOXJ3 removal in embryos has virtually no impact on TE expression. One possible explanation is that in mESCs, FOXJ3 interacts with a factor that is not expressed in the embryo, which could be explored by pulling down overexpressed FOXJ3 in mESCs to identify interactors.

Interestingly, a recent study using a degron system in mESCs revealed that acute TBP depletion has minimal impact on RNAPII transcription but significantly affects tRNA expression through RNAPIII (Kwan et al. 2023). While that study did not assess TE expression, our findings demonstrate that TBP removal in embryos significantly disrupts both MaLR transcription and gene expression. This contrasting effect suggests that although TBP may be dispensable for RNAPII-driven transcription in mESCs, as well as in blastocysts (Martianov et al. 2002), TBP is essential for the widespread gene activation occurring during ZGA in embryos. It is possible that TBP is required for transcription reactivation after mitosis, as suggested by studies showing that TBP depletion impairs gene activation post-mitosis in mESCs (Teves et al. 2018). In contrast, subsequent activation during interphase may proceed independently of TBP (Teves et al. 2018; Kwan et al. 2023). The extensive effect on gene expression that we observe upon TBP depletion in embryos may therefore be linked to the fact that TBP is specifically needed for the activation of transcription. These findings highlight the need for further investigation into the differing impacts of TBP depletion in embryos compared to mESCs.

5.3 Challenging consensus sequences through phylogenetic analysis

The precise annotation and characterization of TE sequence structure are essential for understanding their influence on the genomes they inhabit. TE annotations rely upon consensus sequences representing each family. These sequences are aligned to the genome, and the alignment with highest score is used to determine genomic insertions. Widely used annotation tools, such as RepeatMasker (Smit et al. 2013; Flynn et al. 2020), have facilitated the process of annotating TEs across the genome, often considering families as homogeneous, single clades where all insertions are regarded as identical. However, phylogenetic analyses have begun to challenge this approach, revealing distinct subgroups within families and showing that phylogenetically divergent subfamilies are, in fact, differentially regulated (Carter et al. 2022).

To explore potential diversity among the insertions within our TE families of interest, we conducted a phylogenetic analysis which revealed previously uncharacterized heterogeneity. Phylogenetic analysis of SINEB2 families proved particularly challenging due to their large family size. Including more insertions resulted in long computation times for the phylogenetic tree reconstruction. Attempting to understand the relationship across an entire SINEB2 family would require substantial computational resources. As a results, we limited our analysis to a small subset of sequences from both SINEB2 families, choosing to focus on full-length elements of the same length as used in the footprinting analysis. Despite considerable heterogeneity within B2_Mm1a insertions, which appeared to form a monophyletic subfamily with long genetic distances between mostly individual insertions and small subfamilies, we observed near-complete presence of binding motifs for TBP and FOXJ3. This suggests, though further investigation is required, that part of the sequence may be conserved, preserving the binding motifs for these two TFs, while the remainder of the sequence may be diverging. Neither FOXJ3 nor TBP showed the ability to transcriptionally regulate these elements during preimplantation development. This implies that, at least within this developmental window, these binding motifs may not be important for driving expression of these elements. It could be that these binding motifs are conserved in TEs for other roles than transcriptional regulation.

In contrast, investigating the phylogenies of LTR-containing families is more straightforward due to their smaller number of insertions compared to SINEB2 families. However, ORR1A0 phylogeny is reminiscent to that of B2_Mm1a, featuring a monophyletic structure with limited insertion clustering, yet exhibiting heterogeneity. In contrast, for ORR1A1 and MT2_Mm, we identified two major clades, indicating distinct evolutionary strategies. These findings gain further relevance when considered alongside our TFBS analysis, which explored whether sequence heterogeneity underlies regulatory diversity. We observed that for nearly all TFs in B3A and ORR1A0 families, or for some TFs in B2_Mm1a, ORR1A1 and MT2_Mm, binding sites are either absent from individual insertions, or present in only a few of them. Since these TFBS are originally identified within consensus sequences, this suggests they may be artefacts of the consensus itself and are not present within individual insertions.

Additionally, the binding sites for TFs within the embryo may differ from those in the UniPROBE database. Some TFs, such as DUX, are absent from the database, while others, like OBOX, are known to bind to an extended motif in the embryo, distinct from the motif sequence included in the database (Ji et al. 2023). Even though the OBOX TF family was found to regulate MT2_Mm (Ji et al. 2023), we observed OBOX binding motifs in only a subset of MT2_Mm insertions which were dispersed across subfamilies. Expanding the analysis to include the newly identified extended binding motif could provide more precise insights into the OBOX-mediated regulation of MT2_Mm. Refining the binding motifs for TFs in the embryo, as done for DUX and OBOX (Hendrickson et al. 2017; Ji et al. 2023), and incorporating them into phyloregulatory analyzes such as the ones we performed, may further enhance our understanding of the fine-tuning of TE regulation. For instance, we did not find TBP binding motifs (TATA boxes) in ORR1A0 or ORR1A1 footprints. Nevertheless, our data from embryos clearly show that TBP regulates ORR1A0 and ORR1A1, suggesting that regulation may occur independently of a strong TATA box consensus. In fact, it was suggested in yeast that TBP can bind to promoters in the absence of a TATA box (Kim and Iyer 2004). Due to scarcity of our CUT&Tag data, we were unable to call peaks and generate a de novo TBP binding motif. Improving low-input TF mapping techniques will be crucial to enable such analyses and enhance our understanding of the precise sequence determinants involved in TE regulation in the embryo.

5.4 A complex network of TFs to mediate TE expression in the early embryo

Early studies on L1 regulation and activity have revealed that the youngest elements in both humans and mice are the ones that transpose the most and the most transcriptionally active (DeBerardinis and Kazazian 1999; Boissinot et al. 2000; Goodier et al. 2001; Brouha et al. 2003; Beck et al. 2010). Recurrent changes in the 5’ UTR of L1, a process referred to as the 5’ turnover, has been proposed to be part of an ongoing evolutionary arms race with KRAP-ZFP repressor TFs. In this dynamic, mutations or acquisition of novel a 5’ UTR allow L1 families to evade repression, in turn fueling the expansion of KRAP-ZFP protein families (Thomas and Schneider 2011; Castro-Diaz et al. 2014; Jacobs et al. 2014; Ecco et al. 2017). A compelling illustration of this process is the KRAB-containing KZFP protein ZNF93, whose interaction is restricted to specific L1 families, while notably excluding both oldest and youngest L1 elements (Castro-Diaz et al. 2014; Jacobs et al. 2014). A 129bp deletion within the 5' UTR of the two most recent L1 families completely disrupts the ZNF93 binding motif in L1Hs and L1Pa2 (the youngest human families). Additionally, mutations in the binding sites of ZNF141, ZNF649 and ZNF765 have also shaped the evolution of L1 elements (Imbeault et al. 2017). Beyond escaping host repression, it is proposed that the emergence of new L1 families may be driven by the acquisition of new activating TFBSs (Hermant and Torres-Padilla 2021). This is supported by evidence showing that TF binding is more prevalent in younger L1 promoters (Sun et al. 2018), with a decline in binding and expression correlating with evolutionary age. While the monophyletic origin of L1 allows for a more straightforward linear evolutionary trajectory (Vargiu et al. 2016; Ecco et al. 2017), recent works have begun to explore the evolutionary paths of LTR elements. Notably, work by the lab of Cédric Feschotte uncovered an 8-bp insertion within an LTR7 subfamily, creating a SOX2/3 binding motif that drives expression specifically of that subfamily in pluripotent stem cells, unlike other subfamilies without the motif (Carter et al. 2022). The evolutionary pattern observed suggests that different LTR7 subfamilies expanded simultaneously within the genome, opting for diversification in the tissues and niches they colonize for expression and expansion, rather than following a sequential expansion process over time such as seen in L1 (Carter et al. 2022). Here, the phylogenetic analysis of MT2_Mm showed no correlation between the expression of individual insertions and their evolutionary age, suggesting that recent insertions do not exhibit higher transcriptional activity in the embryo. Notably, we also identified a striking 9bp deletion in the MT2 evolutionary trajectory, preceding the emergence of MT2_Mm, which created the well-known DUX binding site. However, unlike the cases observed with L1 and LTR7, this DUX binding site does not correlate with any change in expression patterns or intensity per se. In fact, MT2C_Mm_vii, which is the youngest subfamily lacking the deletion, was slightly more expressed than MT2C_Mm_viii and _ix at the late 2-cell stage. This underscores the importance of other TFs regulating the expression of MT2C_Mm subfamilies, with particular emphasis on MT2C_Mm_vii. Remarkably, we identified SRF as one such TF, which appeared to specifically control MT2C_Mm_vii expression in mouse embryos.

Our analysis of subfamily expression during preimplantation showed that MT2_Mm subfamilies are all highly transcribed, consistently more so than MT2C_Mm subfamilies, with the exception of MT2C_Mm_v, which exhibited exceptionally high expression, likely due to a single highly expressed insertion, perhaps influenced by its genomic context. These expression differences between MT2_Mm and MT2C_Mm could not be attributed to the acquisition of the DUX binding site, as our luciferase assay confirmed that the deletion event forms a functional DUX binding site, with transcriptional activation by DUX being equally strong on both MT2C_Mm_ix and MT2_Mm_i. Thus, the high transcriptional levels of MT2_Mm subfamilies remain unexplained. It is also unlikely that these differences are due to OBOX binding motifs, which regulate MERVL and MT2_Mm in the embryo (Ji et al. 2023). Deletion of OBOX genes during preimplantation development resulted in equivalent downregulation of both MT2_Mm and MT2C_Mm (Ji et al. 2023), suggesting that OBOX motifs appeared before MT2_Mm expansion.

While our work uncovers an additional regulator of MT2, contributing another piece to the puzzle, the precise factors that make MT2_Mm unique remain unclear. Despite its historical association with DUX binding motif, our findings provide evidence that this motif alone is not responsible for the particularity of MT2_Mm. This implies the involvement of additional TFs in fine-tuning MT2_Mm expression in the embryo. In fact, we also identified that insertions from the oldest subfamily of MT2_Mm, MT2_Mm_i, are bound and regulated by TBP, adding it to the growing list of MT2 regulators.

MT2_Mm expression plays a crucial role for the biology of the host, acting as a central regulator of ZGA and initiating a ZGA program in 2CLC, thereby influencing cell fate decisions. Our findings reveal that both MT2_Mm and MT2C_Mm function as alternative promoters, facilitating chimeric transcript formation and providing TSSs during ZGA to temporally coordinate gene expression. We demonstrate that all MT2C_Mm subfamilies are capable of generating chimeric transcripts, with their expression influenced by the removal of SRF. The expression of the chimeric transcripts derived from insertions belonging to MT2C_Mm_vii were most notably affected by SRF depletion. In fact, we observed a shift in the regulation of these chimeric TE-gene transcripts along the MT2 evolutionary path, initially dependent on SRF and later transitioning to DUX dependence. This shift highlights the evolving nature of transcriptional control, reflecting the complex regulatory dynamics of these elements during development. Importantly, the formation of chimeric transcripts and their role in driving ZGA were not confined to DUX binding motif-containing elements, which have long been linked to this process. We identified SRF as an additional factor involved in regulating chimeric TEs, suggesting the involvement of other, yet-to-be-discovered TFs. Altogether these findings support the hypothesis that the evolution of developmental and transcriptional programs is shaped by the distribution of *cis*-regulatory elements from TEs across the genome, spreading around hubs for TF binding sites that likely work together to ensure the robustness of essential processes in multicellular organisms development (Hermant and Torres-Padilla 2021).

5.5 A remarkable cell-type specificity for TF and TE association

MaLR elements make up a substantial portion of the embryo transcriptome, yet the mechanisms controlling their transcription remain largely uncharacterized. For over two decades, it has been established that these elements serve as alternative promoters in oocytes and mouse embryos (Peaston et al. 2004). 20.4% of oocyte-specific TSSs are associated with TEs, with MaLR elements accounting for about half of these TE-derived TSSs (Veselovska et al. 2015). This regulatory role has now been extended to older MaLR families in embryos across various mammalian species (Oomen et al. 2025). A key finding of the work presented here is the identification of TBP as a regulator of MaLR ORR transcription in the embryo, shedding light on the transcriptional regulation of these elements. Notably, another MaLR family, the MT family, stands out in the oocyte transcriptome, where insertions serve as oocyte-specific promoters, critical for driving the transcriptional network that governs oocyte development (Peaston et al. 2004; Veselovska et al. 2015; Franke et al. 2017). What makes these two MaLR families (ORR and MT) particularly intriguing, is their contrasting, cell-type-specific expression patterns despite belonging to the same superfamily.

TBP expression is transient in oocyte development: it is only present in early stages and reappears upon fertilization. Conversely, its oocyte-specific paralog, TBP2, follows a completely distinct expression pattern, maintaining expression throughout oogenesis to decrease upon ovulation and becoming hardly detectable in nuclei of 2-cell stage blastomeres (Gazdag et al. 2007). Nearly twenty years ago, the authors of these observations hypothesized that the two proteins must have distinct roles in regulating the transcriptional programs of their respective cell types (Gazdag et al. 2007). Indeed, TBP2, in contrast to TBP, has been shown to be essential for ovarian follicle development (Gazdag et al. 2009), and, more importantly, appears to specifically regulate the transcription of MT MaLRs (Yu et al. 2020). Depletion of TBP2 in oocytes results in the downregulation of MTA_Mm, MTB_Mm and MT-int (Yu et al. 2020). MT MaLR dependent on TBP2 are distinct from TBP2-independent insertions based on the presence of high quality TATA boxes, directing sharp transcription initiation and contributing to the oocyte transcriptome (Yu et al. 2020).

We found that TBP regulates the expression of embryo-specific MaLR (ORR) elements, underscoring a striking level of specificity in how two TBP paralogs, TBP and TBP2 regulate distinct cell-type-specific MaLR expression. This interplay between TBP, TBP2 and MaLR elements highlight intricate, dynamic regulation of TE-driven transcription during development. How such general TFs are able to regulate TE with such specificity is both remarkable and surprising. Further elaboration on the mechanisms enabling TBP to specifically recognize these

TEs will be required, especially given the absence of clear TATA box within MaLR ORR. TATA box independent binding of TBP has been reported (Kim and Iyer 2004) and may offer one possible explanation. The genome colonization strategies of MT and ORR MaLRs is reminiscient of those observed for LTR7 (Carter et al. 2022), where distinct developmental niches seem to have been targeted. It remains to be explored whether the amplification of MT and ORR MaLR occurred simultaneously, and what specific sequence determinants enabled the colonization of different developmental niches. TBP appears to regulate ORR insertions regardless of their phylogenetic classification, and is also involved in regulating evolutionary older MT2_Mm insertions (belonging to subfamily MT2_Mm_i). It could be that during evolution of MT2_Mm, the role was eventually taken over by other TFs, such as OBOX, or DUX.

5.6 The early embryo: robustness and resilience

Although TBP depletion results in a significant transcriptional phenotype, with thousands of genes affected, it does not compromise development to the blastocyst stage but merely delays it. About 80% of the genes downregulated in SPT5 LOF (Abe et al. 2022), a global regulator of transcription, were also downregulated in TBP LOF. TBP loss affected twice as many genes as SPT5 loss. Despite this extensive overlap of genes affected, the developmental phenotypes are markedly different. Specifically, SPT5 Trim-Away causes a fully penetrant 2-cell arrest (Abe et al. 2022). While we cannot exclude incomplete TBP depletion with the Trim-Away approach, or transient effects of protein depletion, the transcriptional phenotype observed at the late 2-cell stage does not compromise development. It is possible that the extensive transcriptional changes are not the cause of the 2-cell arrest, but rather that a specific set of genes may be responsible. In fact, it is interesting that in the case of SRF LOF, we observe a 2/4-cell arrest. While non-specific or artefactual effects of the dominant-negative approach cannot be ruled out as contributors to the developmental phenotype, the transcriptional changes are substantially milder, with nearly eight times fewer genes affected compared to TBP LOF. Yet, the developmental phenotype is more pronounced than that of TBP LOF. Although additional methods, such as Trim-Away or genetic approaches, are needed to conclusively determine the effect of SRF removal in early development, overlapping the genes affected in different experiments and correlating them with developmental outcomes would be crucial to pinpoint the essential genes required for the embryo to develop.

Much like the established cooperative and combinatorial action of TFs that modulate gene expression across diverse cell types, analogous mechanisms likely operate within the embryo, fine-tuning gene regulation *in-vivo*. Notably, genes regulated by SRF and DUX show minimal overlap. Moreover, TBP does not control the classic “2C” gene transcription, which are known to be DUX targets. This suggests that, rather than a single master regulator, multiple TFs work in concert, each with distinct yet occasionally overlapping functions, ensuring the robustness in the initial activation of the genome.

In line with this concept, DUX depletion leads to ZGA defects but does not fully abrogate the process. DUX knockout models have shown that DUX is not essential for development, and though at submendelian ratios, *Dux-/-* offsprings are born (Chen and Zhang 2019; De Iaco et al. 2020). However, phenotypic variations across different DUX knock-out studies were observed, initially attributed to differences in genetic background, though they appear to be related to embryo growth conditions. Indeed, DUX-depleted embryos collected pre-ZGA or zygotes microinjected with Cas9 targeting *Dux* showed significant downregulation of MERVL *ex-vivo*, correlating with developmental failure (De Iaco et al. 2017, 2020; Bosnakovski et al. 2021). In contrast, *in-vivo,* MERVL transcription increased from the early to late 2-cell stage in the absence of DUX, correlating with development proceeding normally to the blastocyst stage (Guo et al. 2019). This suggests that, in the absence of DUX, other factors, potentially of reproductive tract origin, may compensate for its loss. Robustness of preimplantation development could be conferred by exogenous factors exclusively present *in-vivo*, which trigger the activation of alternative pathways and may involve other TFs.

Interestingly, we observed a clear difference in SMAD3 subcellular localization between *in-vivo* and *ex-vivo* conditions. SMAD3, a member of the SMAD family of proteins, acts as a mediator of signals initiated by TGF-β superfamily of cytokines. This superfamily encompasses TGF-β isoforms, Activin, NODAL and BMPs (Tarasewicz and Jeruss 2012). TGF-β is the prototypic member of the superfamily, and upon binding to serine/threonine kinase receptors induces complex formation, and phosphorylation cascades activate intracellular SMAD pathway. This involves SMAD3 phosphorylation, interaction with co-SMADs (common SMADs, for example, SMAD4), nuclear entry and gene expression regulation (Attisano and Lee-Hoeflich 2001). TGF-β is expressed from fertilization throughout preimplantation development (Rappolee et al. 1988) and may have redundant roles with Activin, which is maternally inherited and undetectable during cleavage stages (Albano et al. 1993). Single 2-cell stage embryos cultured in microdrops exhibited significantly lower proportions of embryos developing to the blastocyst stage and fewer cells per blastocyst compared to groups of five or ten embryos (Paria and Dey 1990). However, this reduced developmental rate of single embryos can be significantly improved by the addition of exogenous factors such as, for instance, TGF-β (Paria and Dey 1990). These findings suggests that growth factors, from the reproductive tract but also perhaps of embryonic origin play a role in supporting preimplantation development.

It is well established that mouse embryos can develop without exogenous factors, as seen from the simple fact that they progress to the blastocyst stage in chemically defined media, such as the commonly used KSOM. Hence, the differences observed between *ex-vivo* and *in-vivo* grown embryos, such as the absence of SMAD3 nuclear localization in culture embryos, may seem insignificant, given that development still occurs without it. Finding the optimal conditions that support *ex-vivo* development has been historically challenging, mainly due to the “2-cell block” phenomenon. This block, attributed to a combination of factors (reviewed (Biggers 1998)), was eventually overcome with the development of chemically defined media. These media only contain a subset of the compounds and molecules typically found in the embryo natural environment, the female reproductive tract. The absence of these factors creates an imbalance, to which the embryos are forced to adapt if they want to survive (Biggers 1998). In this fragile equilibrium, embryos may be prone to developmental failure if the system is disturbed, such as, for instance, removing a TF, whereas in their natural environment, such disturbances may have little to no impact on developmental outcomes.

5.7 Concluding remarks

Overall, the work achieved during my PhD underscores the complex relationship between TFs and TEs, revealing new insights into the regulatory mechanisms that control ZGA and TE transcriptional activity at the onset of development. Through the identification and characterization of two new TFs regulating TEs, we have successfully expanded the TF repertoire of TEs during preimplantation development. Our findings suggest that we have only reached the tip of the iceberg, as the presence of many other TFs likely orchestrate these processes. The role of additional, yet-to-be-discovered TFs in fine-tuning TE expression and coordinating ZGA remains an important direction for future research.

6 Material and Methods

6.1 Material

6.1.1 Chemicals and reagents

Reagent	Source	Identifier
10X lysis buffer	Clontech	635013
2-mercaptoethanol	Gibco	31350010
Agar	BD DIFCO	214530
Ammonium chloride	Carl Roth	P726.1
Ampicillin	Fisher BioReagents	10193433
AMPure XP beads	Beckman Coulter	A63881
Betaine	Sigma-Aldrich	B03001VL
BSA	Roche	10735078001
$CaCl_2$-$2H_2O$	Thermo Scientific Chemicals	423520250
Cascade blue dextran	Invitrogen	D1976
CHIR99021	Cayman	13122-25
COmplete, EDTA-free Protease Inhibitor Cocktail (PIC)	Sigma-Aldrich	11873580001
D(+)-Glucose	Carl Roth	HN06.1
DMEM	Gibco	41966029
DMEM with GlutaMAX	Gibco	31966047
dNTP mix	Thermo Scientific	R0192
DPBS	Gibco	14190144
EDTA	Carl Roth	8043.2
Ethanol	Merck	1.00983.1000
FCS	PAN-Biotech	P30-3302
Gelatin Solution (0.1% in PBS)	PAN-Biotech	P06-20410
GeneRuler 1kb plus DNA ladder	Thermo Scientific	SM1331
Glycerol	Sigma-Aldrich	G5516-100ML
Glycine	Carl Roth	0079.4
hCG	MSD Animal Health	
HEPES Buffer Solution (1M)	Gibco	15630-056
Hyaluronidase	Sigma-Aldrich	H4272-30MG
Isopropanol	Thermo Scientific Chemicals	327270010
jetPrime	PolyPlus	101000015
Kanamycin	Thermo Fisher	BP906-S
KCl	Sigma-Aldrich	P5405
KH_2PO_4	Sigma-Aldrich	P5655-100G
L-Glutamine	Thermo Scientific Chemicals	A14201
LIF	IGBMC	n/a
Lipofectamine 2000 Transfection Reagent	Invitrogen	11668-019
M2 medium	Sigma-Aldrich	M7167-100ML

MEM Amino Acids Solution	Gibco	11130077
MEM NEAA Solution	Gibco	11140-035
Methanol	Merck	1.06009.2500
MgCl$_2$	Sigma-Aldrich	M1028
MgSO4-7H$_2$O	Thermo Scientific Chemicals	423900250-25G
Na Lactate 60% Syrup	Thermo Scientific Chemicals	250300010
Na Pyruvate	Sigma-Aldrich	P2256-5G
NaCl	Thermo Scientific Chemicals	12314
NaCl (5M), RNase-free	Invitrogen	10609823
NaHCO$_3$	Thermo Scientific Chemicals	424270250
Opti-MEM	Gibco	31985062
PageRuler Prestained Protein ladder	Thermo Scientific	26616
Paraffin Oil	Sigma-Aldrich	18512-1L
PD0325901	Miltenyi	130-106-541
PEG-8000	Sigma-Aldrich	P1458
Penicillin-streptomycin	Gibco	15070063
Penicillin-Streptomycin-Glutamine (100X)	Gibco	10378016
PFA	Sigma-Aldrich	158127
PFA (solution 16%)	Thermo Scientific Chemicals	043368.9M
PMSG	Ceva	
Poly-L-Lysine	Sigma-Aldrich	P4707-100ML
RNA Clean XP	Beckman Coulter	A66514
RNAse inhibitor	Takara	2313A
Spermidine	Sigma-Aldrich	S2501-1G
Sucrose	Sigma-Aldrich	S9378-500G
SuperSignal West Pico PLUS Chemiluminescent Substrate	Thermo Scientific	34580
TAPS	Sigma-Aldrich	T5130
Tet-system approved FBS	Takara	631106
Tris Base	Millipore	1083820100
Triton X-100	Sigma-Aldrich	X100-100ML
Triton X-100 solution (~10% in H2O)	Sigma-Aldrich	93443-100ML
Tween-20	Sigma-Aldrich	P6585
Tyrode's solution, acidic	Sigma-Aldrich	T1788
UltraPure SDS Solution, 10%	Invitrogen	15553027
Vectashield with DAPI	Vector Laboratories	H-1200-10

6.1.2 Kits and enzymes

Kits	Source	Identifier
AccuPrime Pfx DNA polymerase	Invitrogen	10472482

DNA ligation mix "Mighty Mix"	Takara	6023
Dual-Luciferase Reporter Assay Kit	Promega	E1980
GoScript Reverse Transcription System	Promega	A5000
GoTaq qPCR Master Mix	Promega	A6002
HiFi HotStart ReadyMix	KAPA	KM2605
High Sensitivity DNA Kit	Agilent	5067-4626
mMESSAGE mMACHINE T3 Transcription Kit	Ambion	AM1348
mMESSAGE mMACHINE T7 ULTRA Transcription Kit	Ambion	AM1345
NEBNext High-Fidelity 2X PCR Master Mix	NEB	M0541
Nextera XT DNA library Preparation Kit	Illumina	15032354
NucleoBond Xtra MidiKit	MACHEREY-NAGEL	740410.50
NucleoSpin Mini Kit	MACHEREY-NAGEL	740490.250
pA-Tn5 Adaptor Complex	Diagenode	C01090001-30
pGEM-T easy vector systems	Promega	A1360
Q5 Site-directed Mutagenesis Kit	NEB	E0554S
RNase-free DNase Set	Qiagen	79254
RNeasy MinElute Cleanup Kit	Qiagen	74204
RNeasy Mini Kit	Qiagen	74104
SuperScript II reverse transcriptase	ThermoFisher	18064014
Taq DNA polymerase	Thermo Scientific	EP0401
Trypsin-EDTA (0.25%), phenol red	Gibco	25200056
TURBO DNA-free Kit	Invitrogen	AM1907

6.1.3 Plasmids

Only the plasmids used in this study are described here. The insert source (plasmids, cells cDNA or synthesis) is decribed in the methods section.

Name	Experiment	Reference
pCMV-Zscan4c	ESCs overexpression	(Zhang et al. 2019a)
pCMV6-Tbp	ESCs overexpression	This study
pCMV-MYC-Irf9	ESCs overexpression	(Platanitis et al. 2019)
pCMV-MYC-Srf	ESCs overexpression, Luciferase assays	This study
pCMV-MYC-Dux	ESCs overexpression, Luciferase assays	This study
pCMV-MYC-Zbtb7b	ESCs overexpression	This study
pCMV-MYC-Nobox	ESCs overexpression	This study
pCMV-MYC-Ehf	ESCs overexpression	This study
pCMV-MYC-Gabpa	ESCs overexpression, Luciferase assays	This study
pCMV-MYC-Gabpb1	Luciferase assays	This study
pCMV-MYC-Foxj3	ESCs overexpression	This study
pCMV-MYC-Elf3	ESCs overexpression	This study
pCMV-MYC-Rfx7	ESCs overexpression	This study

pCMV-MYC-Smad3	ESCs overexpression	This study
pCMV-MYC-Meis2	ESCs overexpression	This study
pCMV5-Sox8	ESCs overexpression	(Schmidt et al. 2003)
pCMV-SPORT6-Zfp410	ESCs overexpression	Dharmacon
pIRES-NR2F2	ESCs overexpression	This study
pcDNA3-mA-MYB	ESCs overexpression	(Trauth et al. 1994)
pCAGIP-FLAG-MAFK	ESCs overexpression	RIKEN
pCIG-LMX1A	ESCs overexpression	Addgene
pCMV-SPORT6-HMBOX1	ESCs overexpression	Dharmacon
pCMV-GATA3	ESCs overexpression	Addgene
pIRES-DUXBL1	ESCs overexpression	(Tagliaferri et al. 2019)
pCAG-CRX-IRES-GFPd2	ESCs overexpression	Addgene
pcDNA3-ARID5A	ESCs overexpression	(Amano et al. 2011)
pCMV6-ZFP740	ESCs overexpression	OriGene
pCMV6-ZFP281	ESCs overexpression	OriGene
pCMV6-TCF7L2	ESCs overexpression	OriGene
pCMV6-SOX15	ESCs overexpression	OriGene
pCMV6-PKNOX1	ESCs overexpression	OriGene
pCMV6-OBOX1	ESCs overexpression	OriGene
pCMV6-OBOX2	ESCs overexpression	OriGene
pCMV6-OBOX3	ESCs overexpression	OriGene
pCMV6-OBOX5	ESCs overexpression	OriGene
pCMV6-OBOX6	ESCs overexpression	OriGene
pCMV6-LHX8	ESCs overexpression	OriGene
pCMV6-KLF7	ESCs overexpression	OriGene
pCMV6-HPB1	ESCs overexpression	OriGene
pCMV6-FOXK1	ESCs overexpression	OriGene
pCMV6-BBX	ESCs overexpression	OriGene
pCMV6-ATF1	ESCs overexpression	OriGene
pCMV-MYC-SRF-DN	Luciferase assays	This study
pCMV-EMPTY	Luciferase assays	This study
pRN3p-SRF-DN	*In-vitro* transcription	This study
pGEMHE-mCherry-mTrim21	*In-vitro* transcription	Addgene
pCDH-E1Fa-Ren-T2A-mCherry	Luciferase assays	Addgene
pGL2-Scramble	Luciferase assays	This study
pGl2-MT2_Mm_i	Luciferase assays	This study
pGl2-MT2_Mm_ii	Luciferase assays	This study
pGl2-MT2C_Mm_i	Luciferase assays	This study
pGl2-MT2C_Mm_vii	Luciferase assays	This study

pGl2-MT2C_Mm_ix	Luciferase assays	This study
pMT2_Mm_ii_mRuby2	Reporter in embryos	This study
pMT2C_Mm_vii_mRuby2	Reporter in embryos	This study

6.1.3 Oligonucleotides

6.1.3.1 Cloning primers

Target	Sequence 5'-3'	Description	Restriction site
TBP_frd	AAAAGGATCCGCCGCCGCGATCGCCATGGACCAGAACAACAGCCT	Cloning to pCMV6	BamHI
TBP_rev	AAAAGCGGCCGCGTACGCGTAAGGTGGGTTGTGGTCTTCCTGAATCCCTT	Cloning to pCMV6	NotI
SRF_frd	AACAGAATTCTGATGTTACCGAGCCAAGCTG	Cloning to pCMV-Myc	EcoRI
SRF_rev	AACACTCGAGTCATTCACTCTTGGTGCTGTGG	Cloning to pCMV-Myc	XhoI
DUX_frd	AACAGAATTCTGATGGCAGAAGCTGGCAG	Cloning to pCMV-Myc	EcoRI
DUX_rev	AACACTCGAGTCAGAGCATATCTAGAAGAGTCTGAT	Cloning to pCMV-Myc	XhoI
Zbtb7b_frd	AACAGAATTCTGATGGGGAGCCCCGAGGA	Cloning to pCMV-Myc	EcoRI
Zbtb7b_rev	AAAAGGTACCTTAAGAGGACTCCATGGCACCT	Cloning to pCMV-Myc	KpnI
Nobox_frd	AACAGAATTCTGATGGAACCTACGGAGAAGCT	Cloning to pCMV-Myc	EcoRI
Nobox_rev	AAAAGGTACCTCGAGTTACTCTTTAGCTCCAGCGGC	Cloning to pCMV-Myc	KpnI
Ehf_frd	AACAGAATTCTGATGATTCTGGAAGGAAGTGGTGT	Cloning to pCMV-Myc	EcoRI
Ehf_rev	AACACTCGAGTCAGTTCTCATTTTCT	Cloning to pCMV-Myc	XhoI
Gabpa_frd_TA	ATGACTAAGAGAGAAGCAGAAGAG	TA cloning	N/A
Gabpa_rev_TA	AATCTCTTTGTCTGCCTGTAGAG	TA cloning	N/A
Gabpa_frd	AACAGAATTCTGATGACTAAGAGAGAAGCAGAAGAG	Cloning to pCMV-Myc	EcoRI
Gabpa_rev	AAAAGGTACCTCGAGTCAAATCTCTTTGTCTGCCTGTAGAG	Cloning to pCMV-Myc	KpnI

Gabpb1_frd	AACAGAATTCTGATGTCCCTGGTAGATTTGGG	Cloning to pCMV-Myc	EcoRI
Gabpb1_rev	AAAAGGTACCTCGAGCTAAACGGCTTCTTTGTTGG	Cloning to pCMV-Myc	KpnI
Foxj3_frd	AGGCCATGGAGGCCCGAATTATGGGTTTGTATGGACAAG	Cloning to pCMV-Myc	SfiI
Foxj3_rev	AAAACTCGAGCTACACTATTGAATCCCAATCA	Cloning to pCMV-Myc	XhoI
Elf3_frd_TA	ATGGCTGCCACCTGTGAGA	TA cloning	N/A
Elf3_rev_TA	TTAATTCCGACTCTCTCCAACCTCT	TA cloning	N/A
Elf3_frd	AACAGAATTCTGATGGCTGCCACCTG	Cloning to pCMV-Myc	EcoRI
Elf3_rev	AACACTCGAGTTAATTCCGACTCTCTCCAACCTC	Cloning to pCMV-Myc	XhoI
Rfx7_frd	AGGCCATGGAGGCCCGAATTATGGCAGAGGAACAACAAC	Cloning to pCMV-Myc	SfiI
Rfx7_rev	AAAACTCGAGTTATCCCAACATTTCAACAGTAGG	Cloning to pCMV-Myc	XhoI
Smad3_frd_TA	ATGTCGTCCATCCTGCCCTT	TA cloning	N/A
Smad3_rev_TA	CTAAGACACACTGGAACAGCGG	TA cloning	N/A
Smad3_frd	AACAGAATTCTGATGTCGTCCATCCTGCCC	Cloning to pCMV-Myc	EcoRI
Smad3_rev	AACACTCGAGCTAAGACACACTGGAACAGCG	Cloning to pCMV-Myc	XhoI
Meis2_frd	AAAAGAATTCTGATGGCGCAAAGGTACGAT	Cloning to pCMV-Myc	EcoRI
Meis2_rev	AAAACTCGAGTTACATATAGTGCCACTGCCCA	Cloning to pCMV-Myc	XhoI
Srf_muta_frd	TGACTCGAGGTACCGCGG GTTGGTGACTGTGAATGCTGG	Mutagenesis SRF to 1-266	N/A
Srf_muta_rev	GTTGGTGACTGTGAATGCTGG	Mutagenesis SRF to 1-266	N/A
Srf_rm_stop_frd	CTCGAGGTACCGCGGCCG	Mutagenesis 1-266 remove stop	N/A
Srf_rm_stop_rev	GTTGGTGACTGTGAATGCTGGCTTC	Mutagenesis 1-266 remove stop	N/A

GFP_fwd	TCACTCGAGGTGAGCAAGGGCGAGG	Cloning pCMV-Myc-DN-GFP	XhoI
GFP_rev	ACTCTCGAGTTACTTGTACAAGTAGCGTCTTC	Cloning pCMV-Myc-DN-GFP	XhoI
DN_fwd	TCAGGATCCATGTTACCGAGCCAAGCT	Cloning DN to pRN3p	BamHI
DN_rev	ACTGGTAACCTTACTTGTACAAGTAGCGTCTT	Cloning DN to pRN3p	BstEII
Empty_CMV_frd	TGACTCGAGGTACCGCGG	Mutagenesis remove insert in pCMV-Irf9	N/A
Empty_CMV_rev	AATTCGGGCCTCCATGGC	Mutagenesis remove insert in pCMV-Irf9	N/A
Scramble_LTR_frd	TCTGGTACCTGTACGTAGGGAGCAGAAA	Cloning to pGL2	KpnI
Scramble_LTR_rev	CACACTCGAGGCACGCAAGATCCTTATGAA	Cloning to pGL2	XhoI

6.1.3.2 qPCR primers

Target	Sequence 5'-3'	Reference
actin_frd	GCTGTATTCCCCTCCATCGTG	(Cheloufi et al. 2015; Rodriguez-Terrones et al. 2018)
actin_rev	CACGGTTGGCCTTAGGGTTCAG	(Cheloufi et al. 2015; Rodriguez-Terrones et al. 2018)
gapdh_frd	CATGGCCTTCCGTGTTCCTA	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
gapdh_rev	GCCTGCTTCACCACCTTCTT	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
dux_frd	GCCCTGCTATCAACTTTCAAGA	This study
dux_rev	CTGAGACCCCCATTCGCTTG	This study
mervl_int_1_frd	CTCTACCACTTGGACCATATGAC	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
mervl_int_1_rev	GAGGCTCCAAACAGCATCTCTA	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
mervl_int_2_frd	CTCAAGGCCCACCAATAGTTTC	(Zhang et al. 2019a)
mervl_int_2_rev	CCCATGTCAATAAACTCAGCCTG	(Zhang et al. 2019a)
mt2_mm_frd	GGCTACACCTTCTGCTGGAG	(Zhang et al. 2019a)
mt2_mm_rev	TCGCAGCTGTGAATGGAAGT	(Zhang et al. 2019a)
l1_orf1_frd	GGACCAGAAAAGAAATTCCTCCCG	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
l1_orf1_rev	CTCTTCTGGCTTTCATAGTCTCTGG	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
l1_orf2_frd	AGTGCAGAGTTCTATCAGACCTTC	(Fadloun et al. 2013)

l1_orf2_rev	AACCTACTTGGTCAGGATGGATG	(Fadloun et al. 2013)
iapez_frd	AAGCAGCAATCACCCACTTTGG	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
iapez_rev	CAATCATTAGATGCGGCTGCCAAG	(Ishiuchi et al. 2015; Rodriguez-Terrones et al. 2018)
sineb2_frd	GAGCACCTGACTGCTCTTCC	(Fadloun et al. 2013)
sineb2_rev	ACACACCAGAAGAGGGCATC	(Fadloun et al. 2013)
malr_mt_frd	ATGTCTTGGGGAGGACTGTG	(Peaston et al. 2004)
malr_mt_rev	AGCCCCAGCTAACCAGAACT	(Peaston et al. 2004)
malr_orr_int_frd	AAGTATGGCCCCCATAGACTCAT	This study
malr_orr_int_rev	CAACAAGGCCACACCTTCAGA	This study
malr_orr_ltr_frd	CTGGTTGGAATGGGTGTGTC	This study
malr_orr_ltr_rev	GGAGGAGCTGAGAGTTCTATGT	This study

6.1.3.3 Smart-seq+5’ oligonucleotides

Oligonucleotide	Source	Identifier or sequence (5’-3’)
ERCC RNA spike-ins	Ambion	4456653
Oligo-dT30V	Sigma-Aldrich	AAGCAGTGGTATCAACGCAGAGTACT30V
rGrG+G TSO	IDT (ordered as 100nmole RNA oligo; RNAse free HPLC, with modifications)	AAGCAGTGGTATCAACGCAGAGTACATrGrG+G
ISPCR	Sigma-Aldrich	AAGCAGTGGTATCAACGCAGAGT

6.1.3.4 Antisense oligonucleotides

Target	Source	Sequence (5’-3’)
ASO_Dux	IDT	/52MOErG/*/i2MOErA/*/i2MOErT/*/i2MOErT/*/i2MOErC/*C*T*G*C*G*G*T*T*C*T*/i2MOErG/*/i2MOErA/*/i2MOErA/*/i2MOErA/*/32MOErC/
ASO_Scramble	IDT	/52MOErA/*/i2MOErG/*/i2MOErC/*/i2MOErG/*/i2MOErC/*G*G*G*T*A*T*T*G*A*A*/i2MOErC/*/i2MOErC/*/i2MOErA/*/i2MOErG/*/32MOErG/

*52MOEr: 5’ 2-MethoxyEthoxy; 32MOEr: 3’ 2-MethoxyEthoxy; i2MOEr: Internal 2-MethoxyEthoxy; *: Phosphorothioate Bond.*

6.1.4 Antibodies

6.1.4.1 Primary antibodies

Target	Source	Identifier	Applications	Cell type	Dilution
Myc-Tag	Cell Signaling Technology	2276	WB	HEK293	1:10000
H3	Abcam	ab1791	WB	HEK293	1:100000

CRX	Santa Cruz	sc-377138	IF	ESCs and embryos across stages	1:50
GATA3	Cell Signaling	5852	IF	Embryos across stages	1:200
SMAD3	Cell Signaling	9523	IF	Embryos across stages	1:100
ZFP410	Proteintech	14529-1-AP	IF	Embryos across stages	1:200
RFX7	Novus Biologicals	NBP1-71819	IF	Embryos across stages	1:500
LMX1A	Invitrogen	PA5-115517	IF	Embryos across stages	1:500
SRF	Abcam	ab252868	IF	Embryos across stages	1:500
FOXJ3	Affinity Biosciences	af0602	IF, Trim-Away	Embryos across stages and controls Trim-Away	1:500 (dilution for IF)
TBP	(Gazdag et al. 2007)	N/A	IF, Trim-Away	Embryos across stages	1:800 (dilution for IF)
TBP	From Laszlo Tora	N/A	IF	Controls Trim-Away	1:500
Normal Mouse IgG	Merck Millipore	12-371	Trim-Away	N/A	N/A
Normal Rabbit IgG	Cell Signaling	2729S	Trim-Away	N/A	N/A
TBP	Abcam	Ab28175	CUT&Tag	2-cell stage	1:100
Rabbit IgG Isotype Control	Invitrogen	10500C	CUT&Tag	2-cell stage	1:100 from 1mg/ml

6.1.4.2 Secondary antibodies

Target	**Source**	**Identifier**	**Applications**	**Conjugation**	**Dilution**
Rabbit	ThermoFisher	A16110	WB	HRP	1:20000
Mouse	ThermoFisher	A16078	WB	HRP	1:20000
Rabbit	Invitrogen	A21429	IF	Alexa Fluor 555	1:500
Rabbit	Invitrogen	A21245	IF	Alexa Fluor 647	1:500
Mouse	Invitrogen	A11017	IF	Alexa Fluor 488	1:500
Rabbit	Antibodies-Online	ABIN101961	CUT&Tag	None	1:100

6.1.5 Buffers and media

All solutions were prepared with ultrapure H_2O, unless otherwise indicated. All pH values were measured at room temperature and adjusted with NaOH or HCl, unless otherwise stated.

Buffer or media	Ingredients	Final concentrations
SDS Lysis Buffer	SDS Tris-HCl (pH7.5) Glycerol	2% 50mM 10%
TGS 1X	Tris-base Glycine SDS	25mM 192mM 0.1%
Transfer buffer	Tris-base Glycine SDS Ethanol	25mM 192mM 0.03% 20%
TBSt	Tris-base NaCl Tween20	20mM 150mM 0.1%
Blocking buffer (IF on cells)	BSA Triton (in PBS)	3% 0.1%
Washing buffer (IF on cells)	BSA Triton (in PBS)	3% 0.05%
Fixation (IF on embryos)	PFA Triton Tween20 Sucrose (in PBS)	4% 0.04% 0.3% 0.2%
PBSt	Tween20 (in PBS)	0.1%
Ammonium Chloride solution (IF on embryos)	NH_4Cl (in PBS)	2.6mg/ml
Agar (coating plates, IF on embryos)	Agar NaCl	10mg/ml 3.21M
Blocking (IF on embryos)	BSA (in PBSt)	3%
KSMO	NaCl KCl KH_2PO_4 $MgSO_4$-$7H_2O$ Na Lactate D(+)-Glucose $NaHCO_3$ Na Pyruvate $CaCl_2$-$2H_2O$ L-Glutamine EAA NEAA	95mM 2.5mM 0.35mM 0.2mM 22.4mM 0.2mM 25mM 0.2mM 1.71mM 1mM 1X 1X

	BSA EDTA	0.4% 0.01mM
Lysis Buffer (Smart-seq+5')	Clontech 10X ERCC spike-ins	1X 1:581000
NE1 Buffer	HEPES-KOH (pH7.9) KCl Spermidine Triton PIC	20mM 10mM 0.5mM 0.1% 1X
150-wash Buffer	HEPES (pH7.5) NaCl Spermidine PIC	20mM 150mM 0.5mM 1X
Antibody Buffer (CUT&Tag)	EDTA BSA (in 150-wash buffer)	0.002M 1%
300-wash Buffer	HEPES (pH7.5) NaCl Spermidine PIC	20mM 300mM 0.5mM 1X
Tagmentation Buffer	$MgCl_2$ (in 300-wash buffer)	10mM
TAPS Buffer	TAPS (pH8.5) EDTA	10mM 0.0002M
SDS release Buffer	TAPS (pH8.5) SDS	10mM 0.1%
Triton-X quench Buffer	Triton	0.67%

6.1.6 Consumables

Consumable	Source	Identifier
Falcon tubes 15ml	Corning	352196
Falcon tubes 50ml	Corning	C52170
Filter tips 10µl	Sarstedt	70.3010.355
Filter tips 200µl	Sarstedt	70.3031.355
Filter tips 1000µl	Sarstedt	70.3050.355
Tubes & Doomed Caps, Strips of 8	Thermo Scientific	AB0266
Safe-Lock tubes	Eppendorf	0030121023
Petri Dishes	Greiner Bio-One	633180
Pipettes 5ml	Greiner Bio-One	606180
Pipettes 10ml	Greiner Bio-One	607180
Pipettes 25ml	Greiner Bio-One	760180
Pipettes 50ml	Greiner Bio-One	768160
Round Bottom Polypropylene Tubes	Corning	352059
PCR tubes, strips of 8	Merck	BR781332-120EA
6-well culture plates	Corning	353046
100mm culture dishes	Corning	353003
96-well white flat bottom	Corning	353296
4titute FrameStar PCR plates	AZENTA Life Sciences	4ti-0760

Amersham Hybond LFP 0.2 PVDF Western Blotting membrane	Cytiva	10600022
Omnifix-F Duo (syringes)	B.Braun	9161465V
Coverslips 20x20mm	Carl Roth	H873.2
Glass-bottom dishes	MatTek Life Sciences	P35G-0-14C
96-well plates for embryo	Thermo Scientific	3355
Glass pasteur pipettes	Fisher Scientific	ISO7712
SafeSeal Tips Professional 20µl	Biozym	770050
SafeSeal Tips Professional 200µl	Biozym	770200
SafeSeal Tips Professional 1000µl	Biozym	770400
DNA LoBind Tube 1.5ml	Eppendorf	022431021
Amicon Ultra 100K centrifugal filter	Merck Millipore	UFC510024
Capillary glass (for needles)	Harvard Apparatus	GC100TF-10
Capillary glass (for holders)	Harvard Apparatus	GC100-T15
35mm culture dishes (embryos)	Corning	3530001

6.1.7 Equipment

Device	Source
ThermoMixer C	Eppendorf
Mastercycler Nexus	Eppendorf
Mastercycler Nexus gradient	Eppendorf
Centrifuge 5424R	Eppendorf
Centrifuge 5810R	Eppendorf
NanoDrop 2000c	Thermo Scientific
Light Cycler 96 Instrument	Roche
GloMax Discover Microplate Reader	Promega
ChemiDoc Touch Imaging System	Bio-Rad
SP8 Microscope	Leica
Heatplate	Medax
S6E Microscope	Leica
FemtoJet 4i	Eppendorf
DFC 365 FX Microscope	Leica
2100 Bioanalyzer	Agilent
5200 Fragment Analyzer system	Agilent
Flaming/Brown Micropipette puller	Sutter Instrument

6.2 Methods

6.2.1 Footprinting analysis

ATAC sequencing reads were obtained from GEO accession GSE66390 (Wu et al. 2016). Reads were aligned to mm10 with bowtie2 v2.2.9 with the parameter -X 2000. Mitochondrial reads were removed with Samtools v1.9 and duplicate reads removed with Picard MarkDuplicates. Tn5 insertion sites were obtained genome wide using the pyatac ins function in the NucleoATAC package v0.2.1. Tn5 insertion sites were quantified and normalized for library size using SeqMonk v.1.42.1. Meta-repeat analyses were performed in Seqmonk by quantifying Tn5 insertion sites that mapped within intact regulatory elements of each repeats (LTR for MT2_Mm and ORR1A1/0; complete elements for SINEs; monomers for L1) annotated with RepeatMasker. Regulatory elements were considered intact if they corresponded to the length of the consensus sequence in Repbase. In the case of L1MdTf_II, only individual monomers of the 5'UTR were considered. The monomers coordinates were obtained by mapping the monomer consensus sequence using bowtie2 v2.2.9 with parameters -a -x, which were then used to map Tn5 insertion sites as described above. Transcription factor footprints were identified qualitatively, by identification of a local depletion of signal, relative to a higher signal the flanking regions. The sequence underlying each footprint was extracted from the Repbase consensus sequences, and subject to motif analysis using the Tomtom tool from the MEME suite v5.1.1 against the UniProbe Mouse database using Euclidean distance and a significance threshold of E < 30. Using publicly available RNA-seq data (Deng et al. 2014), candidate transcription factors obtained from the footprints were filtered based on expression. The expression matrix from GEO accession GSE45719 (Deng et al. 2014) was downloaded from a GitHub repository ("jhsiao999/singleCellRNASeqMouseDengESC"). The expression matrix was normalized by library size by dividing the counts by the sum of expression across detected genes in each sample. Only TFs with two or more reads in all cells of the mid 2-cell and late 2-cell stages were selected.

6.2.2 Phylogenetic analyses

For sequence selection and filtering, LTRs (MT2_Mm, ORR1A1, ORR1A0 and MT2C_Mm) and internal regions (MERVL-int, ORR1A0-int, ORR1A1-int), complete elements (B3A and B2_Mm1a) coordinates were extracted from the RepeatMasker annotation for the mouse genome (mm10) (RepeatMasker open-4.0.5 - Repeat Library 20140131 (Smit et al. 2013)). L1MdTf_II monomers coordinates were directly recovered from the annotation performed for the footprinting analysis. For LTRs only (excepting MT2C_Mm which does not have any annotated internal sequence) we used OneCodeToFindThemAll.pl (Bailly-Bechet et al. 2014) and rename_mergedLTRelements.pl (Thomas et al. 2018) for MT2_Mm, ORR1A0 and ORR1A1 and their respective internal sequence to identify LTR and internal sequences belonging to the same elements and assign them as 5', 3' or solo LTRs. The LTRs (MT2_Mm, ORR1A0, ORR1A1), full elements (B3A, B2_Mm1a) and monomer (L1MdTf_II) size distributions were visualized by density plot using the ggplot2 (version 3.5.1) library in R (Wickham 2016). All MT2_Mm (solo, 5', 3') with a length between 400-586 bp, ORR1A0 (solo, 5', 3') with a length between 282-410 bp, ORR1A1 (solo, 5', 3') with a length between 282-410 bp, MT2C_Mm with a length between 385-565 bp, B3A with a length between 206-216 bp, B2_Mm1a with a length between 189-195 bp and L1MdTf_II monomers with a length between 210-214bp were used for further analysis. The size-selected sequences were retrieved from mm10 using the getfasta function from the bedtools package (v2.31.1) (Quinlan and Hall 2010) and aligned with MUSCLE (version 3.8.1551) (Sievers et al. 2011) with default parameters. The alignment was trimmed with TrimAl (version 1.4. rev15) (Capella-Gutiérrez et al. 2009) using the option -tg 0.01 to remove columns where more than 99% of the sequences had a gap. A maximum-likelihood phylogeny was generated using IQ-TREE (version 2.1.4-beta) (Minh et al. 2020) with the options --seed 42 -T AUTO -m MFP -B 6000 --ancestral --sup-min 0.95. The consensus tree files (.contree) output from IQ-TREE were visualized using iTOL (Letunic and Bork 2024). Subfamilies for MT2_Mm, MT2C_Mm and B3A were defined as clusters of minimum 30 sequences, supported by a node with UFBootstrap > 0.95, with branch length longer than 0.015. Subfamilies for ORR1A0 were defined as clusters of minimum 20 sequences, supported by a node with UFBootstrap > 0.95, with branch length longer than 0.015. Subfamilies for ORR1A1 were defined as clusters of 50 sequences, supported by a node with UFBootstrap > 0.95, with branch length longer than 0.05. Subfamilies of B2_Mm1a were defined as clusters of 75 sequences, supported by a node with UFBootstrap > 0.95, with branch length longer than 0.015. Subfamilies of L1MdTf_II were defined as clusters of 10 sequences, supported by a node with UFBootstrap > 0.95, with branch length longer than 0.015. These thresholds were defined empirically based on previous work (Carter et al. 2022) and on visual examination of the trees generated by IQ-TREE. Therefore, the criteria are semi arbitrary as there is no standardized manner to perform these analyses across TE families. Insertions that did not qualify as subfamilies because of too few per group, yet separated from the subfamilies with a long enough genetic distance are classified as outgroup and labelled in gray in the figures, except for ORR1A0 where there was an outgroup of three sequences with very long branch length and therefore the tree was pruned in iTOL prior to establishing subfamilies.

6.2.3 Transcription factor binding site mapping

The primary and/or secondary position weight matrices (PWMs) were downloaded from UniPROBE (Newburger and Bulyk 2009; Hume et al. 2015) for the TFs obtained from the footprinting analysis of each TE family, except for DUX, whose PWM was based on the DUX binding site found to bind MT2_Mm in (Hendrickson et al. 2017). A different file containing the PWMs for TFs obtained within each TE family was created. These matrices were used to scan all TE family size-selected single insertions (same as used in the phylogenetic analysis) using the matrix-scan command line from RSAT (version 2020.02.29) (Santana-Garcia et al. 2022) with the following options -pseudo 1 -decimals 1 -2str -origin start -bgfile 2nt_upstream-noorf_Mus_musculus_GRCm38-noov-1str.freq -bg_pseudo 0.01 -return limits -return sites -return pval -return rank -lth score 1 -uth pval 1e-4; except for ORR1A0 and ORR1A1 for which a less stringent p value threshold (5e-4) was used instead of 1e-4. The resulting matrices were converted to binary files and results were displayed using the heatmap function from ggplot2 in R (version 4.2.3). The consensus sequences of MT2_Mm and MT2C_Mm subfamilies were scanned for TFBS presence as described above.

6.2.4 Divergence analysis

Consensus sequences for MT2A, MT2B, MT2C_Mm and MT2_Mm were recovered from Dfam (Storer et al. 2021). These consensus sequences were aligned, trimmed and a phylogenetic tree was reconstructed as described above. The obtained consensus tree file (.contree) was visualized in rectangular mode using iTOL (Letunic and Bork 2024). Based on this phylogenetic tree, the Dfam consensus sequences of the closest ancestor were used to root the phylogenetic trees and establish the genetic distance of each insertion to that root. MT2C_Mm consensus sequence was used to root the MT2_Mm tree. MT2B consensus sequence was used to root the MT2C_Mm tree and a tree containing both MT2_Mm and MT2C_Mm insertions. The consensus sequence was added to the fasta file containing the single insertions, which were aligned, trimmed, and used to reconstruct a phylogenetic tree as described above. The obtained tree files (.treefile) from IQ-TREE were parsed, and the consensus sequence was assigned as root using the phylo package from the biopython module (Cock et al. 2009) (version 1.79). Genetic distances of each insertion to the root were calculated using the distance function available in the phylo package. The distances were visualized using the geom_jitter function from ggplot2 in R.

6.2.5 Consensus sequences and median-joining network analysis

For MT2_Mm and MT2C_Mm new subfamilies, the consensus sequence was established using the majority rule for each nucleotide position using the seqinr package (version 4.2.36) in R (Charif and Lobry 2007). The resulting consensus sequences were aligned with MUSCLE, and alignments were visualized with Jalview (Waterhouse et al. 2009) (version 2.11.3.3). Median-joining network analysis (Bandelt et al. 1999) was reconstructed using POPART (Leigh and Bryant 2015).

6.2.6 Plasmid construction

Standard digestion and ligation techniques were employed for plasmid cloning unless specified otherwise. All restriction enzymes used were from NEB. For ligation, the ligation mix from Takara (6023) was used. Amplification of CDSs from cells (ESCs or MEFs) were performed on cDNA generated like it was done for RT-qPCR in **section 6.2.9**. For some factors, the CDS was first amplified from cDNA and TA cloned using the pGEM-T easy vector systems (Promega, A1360), and then subcloned to the target expression plasmid. For all cloning, competent DH5α cells were used, and DNA was purified using NucleoSpin Plasmid Miniprep kit (Macherey-Nagel, 745088.50). Sanger sequencing was used to verify all plasmids. NucleoBond Xtra Midi kit (Macherey-Nagel, 740410.50) was used to isolate DNA before transfection or *in-vitro* transcription.

6.2.6.1 TF overexpression plasmids

The pCMV-*Zscan4c* plasmid was obtained from Liu Lin (Zhang et al. 2019a). *Tbp* CDS was cloned from pRN3p (Gazdag et al. 2009) to pCMV6. pCMV-MYC-*Irf9* (MYC tag in the N-terminus) was obtained from Thomas Decker (Platanitis et al. 2019). The pCR8/GW/TOPO-SRF (Addgene, 98618) was purchased from Addgene and the cDNA was cloned in the same pCMV-MYC vector as *Irf9*. The *Dux* cDNA was a gift from Didier Trono (De Iaco et al. 2017) and was cloned into the same pCMV-MYC vector. The pMRx-*ThPOK*-IREs-GFP was received from Rémy Bosselut and the *ThPOK* (*Zbtb7b*) CDS was cloned to the same pCMV-MYC vector (Wildt et al. 2007). pCR4-TOPO-*Nobox* was purchased from Dharmacon (MMM1013-211691656) and cloned to the same pCMV-MYC vector. pMXs-IRES-GFP containing flag tagged Mouse *Ehf* cDNA was obtained from Nobuhiro Nakano (Yamazaki et al. 2015) and cloned to the same pCMV-MYC vector. The *Gabpa*, *Gabpb1, Foxj3, Elf3* and *Rfx7* CDS were amplified from mouse ES cells cDNA and cloned to the same pCMV-MYC vector. The *Smad3* and *Meis2* CDS were amplified from MEFs cDNA and cloned to the same pCMV-MYC vector. pCMV5-*Sox8* was gifted from Michael Wegner (Schmidt et al. 2003). pCMV-SPORT6-*Zfp410* was purchased from Dharmacon (MMM1013-202769453) and cloned to pCMV6. pZhC-*Nr2f2*-his-flag was received from Minoru Ko (Nishiyama et al. 2009) and cloned to pIRES. pcDNA3-*mA-MYB* (*Mybl1*) was obtained from Karl-Heinz Klaupner (Trauth et al. 1994). pCAGIP-FLAG-*Mafk* was ordered from RIKEN (RDB15412). pCIG-*Lmx1a* was purchased from Addgene (Kathleen Millen; 45070). pCMV-SPORT6-*Hmbox1* was purchased from Dharmacon (MMM1013-202702395). pCMV-*Gata3* was purchased from Addgene (Douglas Engel, 83818). pIRES-*Duxbl1* was obtained from Geppino Falco (Tagliaferri et al. 2019). CAG-*Crx*-IRES-GFPd2 was purchased from Addgene (Connie Cepko; 73997) (Wang et al. 2014c). pcDNA3-*Arid5a* was received from Riko Nishimura (Amano et al. 2011). The pCMV6-*Zfp740* (MR200524), pCMV6-*Zfp281* (MR215383), pCMV6-*Tcf7l2* (MR224182), pCMV6-*Sox15* (MR225149), pCMV6-*Pknox1* (MR222826), pCMV6-*Obox1* (MR223867), pCMV6-*Obox2* (MR219600), pCMV6-*Obox3* (MR224473), pCMV6-*Obox5* (MR202205), pCMV6-*Obox6* (MR215428), pCMV6-*Lhx8* (MR226908), pCMV6-*Klf7* (MR204201), pCMV6-*Hbp1* (MR208277), pCMV6-*Foxk1* (MR222304), pCMV6-*Bbx* (MR224547), pCMV6-*Atf1* (MR223254) were all purchased from OriGene.

6.2.6.2 Reporter assays plasmids

For renilla luciferase we used the pCDH-E1Fa-Ren-T2A-mCherry vector (Addgene, 104833). The scrambled LTR was designed using the Random DNA sequence generator (https://users-birc.au.dk/palle/php/fabox/random_sequence_generator.php) with the following criteria: similar size as an LTR (500bp), similar GC content as the mouse genome (42%) and minimal TFBS compared to MT2_Mm, which was controlled using RSAT (Santana-Garcia et al. 2022). The sequence was synthesized by Eurofins and amplified using primers introducing KpnI and XhoI restriction sites on the 5’ and 3’ ends, respectively. The consensus sequences (for MT2_Mm_i, MT2_Mm_ii, MT2C_Mm_i, MT2C_Mm_vii and MT2C_Mm_ix) were synthesized by Eurofins in pEX-A128 with KpnI and XhoI restriction sites in the 5’ and 3’ ends, respectively. All sequences were subsequently cloned to the firefly luciferase-containing vector (which does not contain a promoter): pGL2-basic (referred to in the results as pGL2-empty). The mRuby plasmid without promoter was previously described in (Oomen et al. 2025). The consensus sequences for MT2_Mm_ii and MT2C_Mm_vii were cloned upstream of the mRuby.

6.2.6.3 Mutagenesis

All mutagenesis were performed using Q5 Site-directed Mutagenesis (NEB, E0554S). The dominant negative mutant of SRF was generated in the pCVM-MYC. sfGFP was cloned in frame into the 3’ end of the SRF dominant negative, creating the final insert referred to as “SRF-DN”. pCMV-MYC-*Irf9* was mutated to remove the insert (*Irf9* CDS) and generate pCMV-emtpy.

6.2.6.4 *In-vitro* transcribed plasmids

The mutant SRF-DN was cloned to pRN3p-HA construct. The mCherry-Trim21 plasmid was obtained from Addgene (105522) (Clift et al. 2017).

6.2.7 Cell culture

6.2.7.1 mESCs

Mouse E14 ESC lines carrying the 2C:tbGFP-PEST (Nakatani et al. 2022) were cultured in Dulbecco's modified Eagle's medium (DMEM) with GlutaMAX (Gibco, 31966047) containing 15% fetal calf serum (FCS) (PAN-Biotech, P30-3302), 0.1mM NEAA (Gibco, 11140-035), 2x leukemia inhibitory factor (LIF) (IGBMC), penicillin-streptomycin (Gibco, 15070063), 0.1 mM 2-mercaptoethanol (Gibco, 31350010), 3 µM CHIR99021 (Cayman, 13122-25) and 1 µM PD0325901 (Miltenyi, 130-106-541) on gelatin-coated plates (PAN-Biotech, P06-20410), at 37°C in 5% CO_2. Medium was changed every day, and cells were passaged every other day. To passage the cells, cells were first washed twice with Dulbecco's PBS (DPBS) (Gibco, 14190144) and treated for 5 min at 37°C with 0.25% Trypsin-EDTA (Gibco, 25200056) diluted to 0.1% in DBPS. Trypsin was quenched with culture medium, cells were then centrifuged and 1:10 of the cells were plated for maintenance.

6.2.7.2 HEK293

HEK293 Tet-on cells (Clontech, 631182) were cultured in Dulbecco's modified Eagles's medium (DMEM) (Gibco, 41966-029), complemented with 10% Tet-system Approved FBS (Takara 631106) and 1% Penicillin-Streptomycin-Glutamine (Gibco, 10378016), at 37°C in 5% CO_2. Cells were passaged twice a week. To passage the cells, cells were treated for 7 min at 37°C with 1ml of 0.25% Trypsin-EDTA (Gibco, 25200056). Trypsin was quenched with 9 ml the volume of culture medium, and 1:10 of the cells were plated for maintenance.

6.2.8 Plasmid transfections in mESCs

2.5 µg of plasmid encoding for the CDS of each of the 40 TFs were transfected in duplicates, alongside an empty CMV plasmid and no plasmid control. Briefly, the 2.5 µg of plasmid were incubated with 250 µl of Opti-MEM (Gibco, 31985062). At the same time, 5 µl of Lipofectamine 2000 (Invitrogen, 11668-019) were incubated with 250 µl of Opti-MEM. After five minutes of incubation, the two were combined, and incubated for 20 min at RT. The 500 µl of Opti-MEM with DNA and lipofectamine solution were added to 500 µl of medium containing the 250 000 cells. The cells were transfected in suspension, and after 5 min at RT, were plated in 6-well gelatin-coated plates. 48h after transfection, cells were harvested and ½ of the cell pellets were kept at -80 °C, the other half was resuspended in SDS lysis buffer (2% SDS, 50mM Tris-HCl ph7.5, 10% glycerol) and kept at -20 °C.

6.2.9 RNA extraction, reverse transcription and qPCR

RNA was extracted using RNeasy Mini Kit (Qiagen, 74104) following manufacturer’s instructions. RNAs were treated with DNase in two steps with two different enzymes, first the RNase-free DNase I (Qiagen, 79254) for 30’ on RNA extraction column. 5 µg of RNA were used for 1H TURBO DNase treatment with the TURBO DNA-free kit (Invitrogen, AM1907) following manufacturer’s instructions. The RNA was subsequently purified again using the RNeasy MinElute Cleanup Kit (Qiagen, 74204). For each sample, 1 µg of RNA was used to generate cDNA using the GoScript Reverse Transcription System (Promega, A5000) with random hexamers. For each reverse transcription reaction, identical reaction was conducted without RT enzyme (no reverse transcription control). Real-time PCR was done using the GoTaq qPCR Master Mix (Promega, A6002) and a Roche Applied Science Lightcycler (LightCycler96). Cycling conditions were as follows: 600s at 95 °C, 10s at 95°C, 10s at 55°C and 10s at 72°C, for a total of 45 cycles. The sequences for the primers used are provided in the materials section. Primer sequences were either obtained from previous publications or PCR products were verified by sanger sequencing. PCR were performed in triplicates, melting curve analysis was performed for each sample to control for proper amplification, and the mean Ct was calculated. In addition, no RT control PCR were conducted for each sample in triplicate and analyzed to control for DNA contamination. Ct values were normalized using Actin primer set, and fold change over no vector control was calculated. Empty CMV control vector fold change over no vector control was also calculated and always analyzed aside every sample. The mean fold change over the 2 biological replicates for all TFs were plotted using ggplot2.

6.2.10 Luciferase reporter assay

24h before transfection $2x10^5$ cells were seeded in 6-well plates. Cells were transfected with the firefly luciferase plasmid (1.5µg), the renilla luciferase plasmid (20ng) and the pCMV-TF plasmid for the corresponding transcription factor using the amount indicated in the figures (from 5 to 500ng). The pCMV-TF plasmid but with no insert (pCMV-empty) was used to adjust the levels of DNA transfected to 2 µg in all conditions except for GABPA and GABPB1 where it was adjusted to 2.5 µg. Transfections were performed using jetPrime (PolyPlus 101000015) using 1:1 ratio (DNA:jetPrime). The medium was replaced on the day following the transfections and cells were lysed 48h after transfection. Luciferase activity was measured using the Dual-Luciferase Reporter Assay kit (Promega, E1980) following the manufacturer’s instructions. The ratios of luciferase to renilla were computed for all experiments and fold change over control for each replicate was calculated. The fold change values were log2 transformed and plotted in R using ggplot2 (version 3.5.1). For statistical analysis a linear model was fitted to the data excluding the intercept to evaluate group differences (R version 4.1.2). All MT2 luciferase performed were used together to fit the model, for GABPA, SRF and DUX separately. All controls luciferase performed were used together to fit the model, for SRF and DUX separately. Pre-selected hypotheses were tested and corrected for multiple comparisons using the glht function from the multcomp package (version 1.4-25). To determine significance, an adjusted p value threshold set to 0.05 was used.

6.2.11 Western blot analysis

HEK293 (Human Embryonic Kidney) Tet-on cells were cultured and transfected as above, except with 500ng of pCMV-TF vector. Cell lysates containing proteins were recovered as performed for luciferase assay (using Dual-Luciferase Reporter Assay kit lysis buffer). Proteins were separated on 12% polyacrylamide gel in TGS 1X buffer (25mM Tris, 192 mM glycine and 0.1% SDS in water) which was subsequently transferred in transfer buffer (25 mM Tris, 192 mM glycine, 0.03% SDS and 20% Ethanol) to 0.2 μm PVDF membrane (Cytiva 10600022) previously activated in methanol. The membranes were blocked in 3% BSA in TBSt for 1H at RT. Membranes were then incubated with primary antibodies in blocking solution overnight at 4°C on a nutator. Antibodies used were anti-Myc-Tag (Cell Signaling 2276, 1/10000 dilution) and anti-histone H3 (Abcam ab1791, dilution 1/100000). 3x5min washes with TBSt were followed by 1h incubation with HRP-conjugated secondary antibodies in blocking solution for 1H at RT. The secondary antibodies used were: anti-rabbit (ThermoFisher A16110, dilution 1/20000), anti-mouse (ThermoFisher A16078, dilution 1/20000). After 3x5min washes in TBSt, membranes were visualized using SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo Scientific, 34580) with ChemiDoc Touch Imaging System (Bio-Rad).

6.2.12 Embryo collection

All animal experiments were in compliance of the legislation from the Government Upper Bavaria. For immunostainings, CD1 females (6-10 weeks old) were mated with CD1 males (2-8 months old). Zygotes, early 2-cell, late 2-cell, 4-cell and 8-cell were collected at ~16h, ~32-33h, ~41-42h, ~48h, ~56-57h post coitum, respectively. For microinjections and CUT&Tag, F1 (C57BL/6J × CBA/H) females (<10 weeks old) were mated with F1 males (3-6 months old). Ovulation was induced by injection of pregnant mare serum gonadrotropin (PMSG, Ceva) followed by human chorionic gonadotropin (hCG, MSD Animal Health) 48h later.

6.2.13 Immunostainings

The list of primary and secondary antibodies used is provided in the materials section.

6.2.13.1 Immunostaining in mESCs

For anti-CRX antibody positive control in ESCs, cells were transfected as described above and seeded directly onto coverslips. 24h later, cells were washed twice with PBS and fixed for 15min in 3% PFA in PBS. Cells were subsequently washed three times with PBS and permeabilized 5min in ice-cold 0.5% triton in PBS. Cells were then washed once and incubated for 15min in blocking buffer (3% BSA, 0.1% triton in PBS). Primary antibody incubation was performed in washing buffer (1% BSA, 0.05% triton in PBS) for 45 min with primary antibody. Cells were washed three times in washing buffer, and incubated for 30-40 min in washing buffer with secondary antibody. Cells were finally washed three times in washing buffer and mounted in vectashield containing DAPI (Vector Laboratories, H-2000). Confocal microscopy was done using 63x oil objective of SP8 microscope (Leica).

6.2.13.2 Immunostaining in embryos

Unless performed on injected embryos or otherwise stated, all immunostainings were performed using embryos coming from CD1 mice. Immunostainings were performed as described previously (Torres-Padilla et al. 2006). The zona pellucida was removed with acid Tyrode (Sigma-Aldrich). After PBS wash, the embryos were fixed in 4% PFA, 0.04% Triton, 0.3% Tween-20, 0.2% sucrose on a glass-bottom dish at 37°C for 20min. Embryos were washed 3X in PBS and permeabilized with 0.5% Triton X-100 for 20min at RT. After three washes in PBSt (0.1% Tween-20 in PBS), embryos were incubated in 2.6mg/ml ammonium chloride (in PBS) solution, washed again twice in PBSt and blocked for 4-5h in 3% BSA in PBSt at 4°C. Overnight incubation with primary antibody in 3% BSA was performed. The dilution used for each antibody is provided in the materials section. The day after, embryos were washed three times in PBSt, briefly blocked (20min in 3% BSA), and incubated with the secondary antibody for 3-4h at RT in 3% BSA. The dilution used for each secondary antibody is provided in the materials section. Finally, embryos were washed twice in PBSt, once in PBS for 20min and mounted on coverslips coated with poly-L-lysine in vectashield containing DAPI (Vector Laboratories, H-2000). Confocal microscopy was done using 63x oil objective of SP8 microscope (Leica).

6.2.14 In-vitro transcription and antibody purification

mRNAs for SRF-DN, dsRed and GFP (all in pRN3p) were transcribed *in-vitro* using the T3 mMESSAGE mMACHINE transcription kit (Ambion AM1348). The mCherry-Trim21 mRNA was transcribed *in-vitro* from the T7 promoter using the mMESSAGE mMACHINE T7 Ultra kit (Ambion AM1345). 50 µg of each antibody used for Trim-Away were purified using Amicon Ultra 100K Centrifugal Filter (Merck Millipore UFC510024), following manufacturer's instruction without addition of NP-40.

6.2.15 Embryo manipulation and culture

For all embryo manipulation experiments, upon collection of zygotes from the females oviducts, embryos were briefly incubated with M2 containing hyaluronidase (Sigma-Aldrich) to remove cumulus cells. After manipulations, embryos were cultured in K-modified simplex optimized (KSOM) microdrops covered with paraffin oil (Sigma-Aldrich) at 37°C and 5% CO2 until collection time, as indicated for each experimental design.

6.2.15.1 Dominant-negative and ASO-mediated LOF

For RNA-seq, zygotes were collected between 24-25h post-hCG and microinjected with 150ng/µl dsRed mRNA with in addition: 20µM of either scramble ASO or anti-Dux ASO (Guo et al. 2024) and with or without SRF-DN mRNA (500ng/µl). Embryos were cultured until collection at 48h post-hCG for single embryo RNA-seq. For SRF LOF, only 2-cell stage embryos with nuclear GFP signal were collected. 9 CONTROL, 9 DUX LOF, 7 SRF LOF and 10 DOUBLE LOF 2-cell stage embryos from at least 3 independent experiments were collected for Smart-seq+5' and 1 embryo from the DUX LOF group was removed after quality control analysis. For development experiments, SRF LOF and CONTROL embryos were collected between 24-25h post-hCG and injected with SRF-DN mRNA (500ng/µl) and GFP mRNA (250ng/µl), respectively, then cultured and scored on day 1 (48h phCG), 2 (72h phCG), 3 (96h phCG) and 4 (116h phCG) for developmental progression. Developmental progression was plotted with excel.

6.2.15.2 Reporter assay

For reporter assay in embryos, zygotes were collected between 18-19h post-hCG and injected with 40 ng/µl of either MT2_Mm_ii, MT2C_vii or no promoter reporter plasmids. The embryos were cultured until 48h post-hCG when the ruby signal was observed with an epifluorescent microscope and the number of positive embryos were counted.

6.2.15.3 Trim-Away LOF

Zygotes were collected between 17-18h post-hCG and microinjected with 0.33% dextran cascade blue (Invitrogen, D1976) and 500ng/µl of purified antibody in PBS. Antibodies used were the following: the mouse monoclonal 3G3 anti-TBP (Gazdag et al. 2007; Brou et al. 1993) and the normal mouse IgG (Merck Millipore, 12-371) for the TBP loss of function. For the FOXJ3 loss of function, the rabbit polyclonal anti-FOXJ3 (Affinity Biosciences, af0602) and the normal rabbit IgG (Cell Signaling, 2729S) were used. Zygotes were cultured for 4h, then injected a second time with 200ng/µl Trim21-mCherry mRNA. Embryos were cultured until collection at 48h post-hCG for single embryo RNA-seq. 10 TBP LOF with 9 CONTROL (mIgG) and 17 FOXJ3 LOF with 13 CONTROL (rIgG) embryos from at least 3 independent experiments were collected for Smart-seq+5'. 1 TBP LOF CONTROL embryo as well as 1 FOXJ3 LOF embryo were removed after quality control analysis. For every collection, a few embryos were kept for immunostainings to control for proper protein depletion, performed as described above. The primary and secondary antibodies used are provided in the materials section. For TBP LOF development experiments, TBP LOF and CONTROL embryos were injected as indicated above, and development was monitored at 48h, 66h, 72h, 96h, and 116h post-hCG. Developmental progression was plotted with excel.

6.2.16 RNAseq (Smart-seq+5')

SMART-seq+5' was modified from the Smart-seq2 protocol (Picelli et al. 2013, 2014) as previously described (Oomen et al. 2025). For SRF LOF, DUX LOF, DOUBLE LOF and their CONTROL embryos, and independently for TBP LOF and CONTROL embryos, all samples were collected in the same lysis buffer, which was stored at -80°C until use (Clontech 10X lysis buffer (635015) diluted to 1X in H2O supplemented with ERCC RNA spike-ins (diluted to 1:581,000, Ambion 4456653) and aliquoted in PCR tubes (5.8μl/tube)). For FOXJ3 LOF, a first batch of library preparation was performed with all samples collected in the same buffer aliquot (10 FOXJ3 LOF and 7 CONTROL embryos), and processed together. A second batch of library preparation was done from embryos collected using the same buffer, but from a different aliquot (7 FOXJ3 LOF and 6 CONTROL embryos) and thus processed as a different batch. At the time of collection (48h post-hCG), embryos were washed 3X in PBS, transferred to tubes containing lysis buffer, snap-frozen in liquid nitrogen and kept at -80°C until further processing. RNA was extracted with RNAClean XP beads (Beckman Coulter, A66514) and resuspended in 1µL of dNTP mix (Thermo Scientific, R0192), 1µL of oligo-dT30V (10μM, Sigma-Aldrich, 5'-AAGCAGTGGTATCAACGCAGAGTACT30V-3') and 1µl of nuclease-free H2O containing 5% RNAse inhibitor (Takara, 2313A). The samples were first incubated for 3 min at 72°C and kept on ice until further processing. In the meantime, the reverse transcription solution was prepared, for one sample: 2µl of Superscript II 5x RT buffer (ThermoFisher, 18064014), 1,6µl PEG-8000 40% (Sigma-Aldrich, P1458), 0,5µl DTT, 0,25µl RNAse inhibitors (Takara, 2313A), 0,1µl 100µM TSO (IDT, 5'-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3'), 0.06µl MgCl2 1M (Sigma-Aldrich, M1028), 2µl Betaine 5M (Sigma-Aldrich, B0300-1VL) and 0.5µl Superscript II RT (ThermoFisher, 18064014). 7µl of the reverse transcription mix was added to the 3µl of annealed RNA mix and incubated for 90min at 42°C followed by 15min at 70°C. Preamplification of the obtained cDNA was performed using KAPA HiFi ReadyMix (KM2605) for 14 cycles with ISPCR primers (10µM, Sigma-Aldrich, 5′-AAGCAGTGGTATCAACGCAGAGT-3′), and purified using Ampure XP beads (Beckman Coulter, A63881). 2.5µl of 120 pg/µl cDNA for each sample was used for tagmentation, which was performed with the Nextera XT kit (Illumina, 15032354). The preamplified cDNA was mixed with 5µl of Tagment DNA buffer and 2.5µl of Amplicon Tagment Mix and incubated at 55°C for 5min. The tagmentation reaction was stopped with 2.5µl NT buffer, and samples were incubated at RT for 5min. Tagmented DNA was then amplified for 12 cycles using the two standard i5 and i7 Nextera Unique Double Indexes together with a tailed i7 index containing an overhang enabling the capture of the 5' of the transcripts. The libraries were verified using the 2100 Bioanalyzer with the High Sensitivity DNA kit (Agilent) or using the 5200 Fragment Analyzer System (Agilent). 150 bp paired-end sequencing protocol was used on Illumina NovaSeq 6000 platform.

6.2.17 RNA-seq analysis

6.2.17.1 ES cells RNA-seq analysis

Expression analysis of TFs in ESCs was done from ArrayExpress accession E-MTAB-2684 (Ishiuchi et al. 2015). Reads were aligned to mm10 with STAR using parameter –quantMode GeneCounts to count reads. The expression matrix was normalized by number of exonic kb and library size (RPKM). The normalized RNA levels obtained for each TFs were plotted using ggplot2 with hierarchical clustering applied to rows.

6.2.17.2 Smart-seq2 analysis

The expression matrix from GEO accession GSE45719 (Deng et al. 2014) was downloaded from a GitHub repository ("jhsiao999/singleCellRNASeqMouseDengESC"). The expression matrix was normalized by library size by dividing the counts by the sum of expression across detected genes in each sample. The mean normalized relative RNA levels for TFs were plotted using ggplot2.

6.2.17.3 Mapping and processing of Smart-seq+5'

Smart-seq+5' libraries were processed as previously described (Oomen et al. 2025) as indicated in https://github.com/meoomen/Smartseq5. In brief, sequence quality was verified using FASTQC. Trimmomatic (Bolger et al. 2014) configured in paired-end (PE) mode was used to remove adaptor and low-quality sequences. A custom Python script was used to sort between 5' transcript ends and internal transcript fragments according to their adaptor sequence. Unless otherwise stated, all analyzes were performed with the 5' reads. For Trim-Away samples only, the reference genome was modified to contain the construct "pGEMHE-mCherry-mTrim21" fasta file, and the endogenous Trim-21 locus was masked using bedtools (version 2.31.0). The modified reference fasta file was used to prepare STAR index files. The reads were mapped to GRCm38 using STAR (v2.7.11a). TEcount and TElocal from TEtranscript (Jin et al. 2015) were used to count TE and gene reads. The RepeatMasker annotation file for mm10 (mm10_rmsk.gtf) was modified to two new annotations. First, L1MdTf_II elements were removed to avoid double counting of reads after addition of the L1MdTf_II monomer coordinates (described in the footprinting section) to the annotation file. The latter mentioned annotation, containing L1MdTf_II monomer coordinates, was further modified, to include the information of the new identified phylogenetic subfamilies. Insertion coordinates corresponding to specific subfamilies were identified using bedfiles generated during the phylogenetic analysis. These coordinates were then used to modify the annotation file by modifying the "gene_id" field for TEcount analysis and the "transcript_id" field for TElocal analysis. Not all insertions were matched from the phylogenetic analysis leading to slight insertion number differences for each subfamily. For all analysis of RNA-seq datasets generated within this study, the second modified annotation was used to count TE reads. When indicated, the subfamily information was taken into consideration. TEcount was also used to count reads coming from genes, using the gencode M20 gene annotation. For TElocal, we created an index file based on our modified TE annotation with the TElocal_indexer script available at https://labshare.cshl.edu/shares/mhammelllab/www-data/private/TEindexer/TElocal_indexer (Jin et al. 2015). Reads from individual insertions were counted using this index file.

Expression analysis across development stages was performed directly using the BAM files containing the 5' reads from (Oomen et al. 2025). TEcount was used to count gene and TE reads using, for expression patterns at the family level, the first modified RepeatMasker annotation containing the L1MdTf_II monomers, while for subfamily expression the modified annotation containing phylogenetically-established subfamilies was used. Plots for both genes and TEs were made using ggplot2 (version 3.5.1) in R. To analyze single insertion expression from non-manipulated embryos, the fastq files of the 2-cell stage embryos from (Oomen et al. 2025) were processed as described above and TElocal was used to count reads from single TE insertions, using the same index file as described above. The data was visualized using jitter dot plots from ggplot2 package in R. Single insertion TE local values were combined in a table with genetic distances obtained in the divergence analysis and expression against age was plotted using ggplot2 in R.

6.2.17.4 Differential expression analysis

The differential expression analysis was performed independently for each experiment: SRF LOF, DUX LOF, DOUBLE LOF and their CONTROL embryos were considered one experiment, which we will refer to as SRF/DUX experiment. TBP LOF and their control embryos were considered as one experiment, referred to as TBP experiment. Finally, FOXJ3 and their control embryos were considered as one experiment, referred to as FOXJ3 experiment. For each experiment, the sample count tables generated by TEcount from the 5' reads or internal reads data were merged into a single table using a bash script and loaded in R. Reads per million (rpm) were calculated for each sample. Quality control thresholds were set as follows: a minimum of $5x10^5$ reads, a minimum of 1000 detected genes, the maximum percentage of reads assigned to mitochondrial DNA was set to 10 percent, and the maximum percentage of reads assigned to ERCC spike ins was set to 15 percent. From these criteria, one embryo from TBP experiment, as well as one embryo from FOXJ3 experiment, were removed from further analysis. Expression levels of the TFs SRF and DUX were also assessed in SRF/DUX experiment, and one DUX LOF sample was removed due to higher *Dux* expression (compared to controls) indicating that down-regulation of *Dux* had not worked in this embryo. Only genes and TEs with at least one read in at least 25% of the total samples were considered expressed and included in the differential expression analysis. Differential expression analysis was performed using DESeq2 (v1.38.3) (Love et al. 2014) with read counts per gene and TEs calculated by TEcount. For SRF/DUX experiment, we performed two DESeq objects, the first one, which we refer to as General DESeq object in our pipeline, by using the TE read counts at the family level, for which we compiled all the subfamilies that we identified in our phylogeny as one single family and comprise all the full length insertions used for the phylogeny. In this annotation, the label “TE_family:TE_superfamily:TE_class:OTHERS” (for example: “MT2_Mm:ERVL:LTR:OTHERS”) comprises all the (fragmented) sequences that were not included in the size selection for phylogenetic analysis. In the second DESeq object we used TE read counts for individual subfamilies as determined in our phylogenetic analyses (Subfamily DESeq object). For TBP and FOXJ3 experiments independently, we performed one DESeq object in which we used TEs read counts at the family level and compiled all the insertions from each family regardless of the size selection and the subfamilies defined in the phylogenetic analysis pipeline. For FOXJ3 experiment, batch was added as co-variate to the DESeq2 model and batch effect correction was applied. In the case of TBP, as the effect of TBP removal on the transcriptome was extensive, ERCCs were used as size vector for normalization in DESeq2. MA plots were used to present the differential expression analysis, showing the log2 fold change between the LOF experiment and its respective control. For gene MA plots, scattermore package (v1.2) was used for plotting. Comparison of the DUX and SRF DE genes were visualized using scatterplots (using ggplot2). Adjusted p value (padj) threshold was set to 0.05 for significance. Genes significantly up or downregulated in SRF LOF were assigned to maternal RNAs, minor ZGA, major ZGA or other using the Database of Transcriptome in Mouse Early Embryo (DBTMEE) (Park et al. 2015). Changes between individual embryos were visualized in heatmaps displaying the log2 of normalized counts centered on the row mean using pheatmap (v1.0.12) in R with hierarchical clustering applied to rows for each group of genes. Venn diagrams were created using the VennDiagram package in R (v1.7.3), using gene list for major ZGA from DBTMEE (Park et al. 2015), and for CDK9 inhibition and SPT5 Trim-away from (Abe et al. 2022). For embryonic PCA analysis, embryos from the different embryonic stages from (Oomen et al. 2025) alongside on the one hand SRF LOF embryos, DUX LOF embryos and their CONTROL embryos and on the other hand TBP LOF embryos and their CONTROL embryos, counts were log2-transformed to generate the PCA. For single insertion differential expression analysis, the sample count tables from TElocal for TBP experiment were merged into a single table using the same bash script as described above and loaded in R. Representation, filtering and significance were performed as described above.

6.2.18 Identification of TE-initiated gene transcripts

Chimeric TE-gene interactions were identified with chimeraTE (Oliveira et al. 2023) using mode-1 and only keeping TE-initiated chimeric transcripts. First, mapped quality-passed 5' fragments read pairs were converted back to fastq format using samtools fastq. Converted fastq files were used as input for ChimeraTE with parameters –strand rf-stranded and --window 150000. For quantification, we used the modified TE annotation containing the newly established subfamilies and the GRCm38 gene annotation. Only TE-initiated chimeric transcripts present in 2 or more replicates per experimental condition were used in downstream analysis and visualization. Using chimeric transcripts present in control embryos, we plotted the combined relative (centered on mean across conditions) log2 counts of all chimeric reads within each family and for each condition as boxplots using ggplot2 in R. Next, we identified transcript isoforms de novo with bambu (Chen et al. 2023) using the junction files outputted by chimeraTE as input. Lastly, we visualized promoter usage of canonical or chimeric promoters using proactive (Demircioğlu et al. 2019).

6.2.19 CUT&Tag

50-70 and 85-100 late 2-cell stage embryos were collected from the females between 46-48h post-hCG for TBP and IgG control replicates, respectively. The embryos were washed in M2, in PBS, in ice-cold NE1 buffer (HEPES-KOH (pH7.9) 20mM, KCl 10mM, Spermidine 0.5mM (Sigma-Aldrich, S2501-1g), Triton 0.1%, PIC 1x (Sigma-Aldrich, 11873580001)) and transferred to another ice-cold NE1 buffer drop for 10min incubation on ice. Embryos were subsequently washed in PBS, and in 150-Wash buffer (HEPES-NaOH (pH7.5) 20mM, NaCl 150mM, Spermidine 0.5mM (Sigma-Aldrich, S2501-1g), 1X PIC (Sigma-Aldrich, 11873580001)), followed by overnight incubation in antibody buffer (0.002M EDTA, 1% BSA in Wash buffer 150) with anti-TBP antibody (Abcam ab28175, 1/100 dilution) or rabbit IgG Isotype Control (Invitrogen 10500C, 1/100 dilution from 1mg/ml). The embryos were transferred to a secondary antibody solution with guinea pig anti-rabbit antibody (Antibodies-Online ABIN101961, 1/100 dilution) in 150-Wash buffer for 30min at RT. Embryos were further washed in 150-Wash buffer, and transferred to pA-Tn5 adaptor complex (Diagenode C01090001-30, 1/200 dilution) solution in 300-Wash buffer (HEPES-NaOH (pH7.5) 20mM, NaCl 300mM, Spermidine 0.5mM (Sigma-Aldrich, S2501-1g), 1X PIC (Sigma-Aldrich, 11873580001)) for 1H at RT. After washes in 300-Wash buffer, tagmentation was performed in freshly made tagmentation buffer (10mM MgCl2 in 300-Wash buffer) for 1H in 37°C incubator. Embryos were subsequently washed in freshly made TAPS buffer (10mM TAPS (pH8.5), 0.0002M EDTA in H2O), and transferred to a PCR strip containing 5µL SDS release buffer (10mM TAPS (pH8.5), 0.1% SDS in H2O). Release was performed for 1H at 58°C, and quenched with addition of 15µL of Triton-X quench buffer (0.67% Triton in H2O). 2µL of 10µM barcoded i5 and i7 primer solutions were added to each sample as well as 25µL NEBNext HiFi 2x PCR master mix (NEB, M0541) and amplified for 18 cycles or 22 cycles, for TBP and IgG embryos, respectively. The libraries were purified using Ampure XP beads (Beckman Coulter, A63881) and verified using 5200 Fragment Analyser System (Agilent).

6.2.20 CUT&Tag analysis

CUT&Tag sequence reads were trimmed for sequence adaptors and low quality ends using Trimmomatic (Bolger et al. 2014). Paired-end reads were then mapped to mm10 using bowtie2 (Langmead and Salzberg 2012) maintaining a maximum insert length of 2000bp (-x 2000). Using samtools, only paired, unique mapping alignments with a minimum alignment length of at least 30bp were kept for downstream analysis. Optical and PCR duplicates were removed using picard (http://broadinstitute.github.io/picard). Mitochondrial reads were removed. Libraries were normalized to rpm values and visualized as bigwigs using bedtools genomecov and bedGraphToBigWig. Replicates were merged on a bigwig level using mean values. CUT&Tag libraries were not selected for insert length due to low sequence depth/complexity, however this was taken into account in our QC assessment of all libraries. IgG or TBP signal was visualized as a pile-up on genes or TEs using deeptools (Ramírez et al. 2016). For signal visualization on genes, the TSS annotation for mm10 (gencode M20 gene annotation) was used using deeptools “reference-point”. For TEs, the modified RepeatMasker annotation including subfamilies used for the RNA-seq analysis was used using deeptools “scale-regions” where the average TE length was used as region length. Both genes and TEs were categorized as significant up/down or non-significant based on the differential expression analysis detailed previously.

6.2.21 Data deposition

The RNA-seq data from SRF/DUX experiments from this study are available from the Gene Expression Omnibus (GEO) database, accession number GSE271983. The RNA-seq data from TBP and FOXJ3 experiments from this study are available from the Gene Expression Omnibus (GEO) database, accession number GSE288110 (Reviewer accession token: ifyfuwaybxebfyj). The CUT&Tag data from this study are available from the Gene Expression Omnibus (GEO) database, accession number GSE288111 (Reviewer accession token: kvojkkkaldyjjop). Previously published smart-seq+5' RNA-seq datasets across development re-analyzed are available under accession code GSE225056.

7 References

Abe K, Schauer T, Torres-Padilla M-E. 2022. Distinct patterns of RNA polymerase II and transcriptional elongation characterize mammalian genome activation. *Cell Rep* **41**: 111865.

Albano RM, Groome N, Smith JC. 1993. Activins are expressed in preimplantation mouse embryos and in ES and EC cells and are regulated on their differentiation. *Dev Camb Engl* **117**: 711–723.

Allen TA, Von Kaenel S, Goodrich JA, Kugel JF. 2004. The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. *Nat Struct Mol Biol* **11**: 816–821.

Amano K, Hata K, Muramatsu S, Wakabayashi M, Takigawa Y, Ono K, Nakanishi M, Takashima R, Kogo M, Matsuda A, et al. 2011. Arid5a cooperates with Sox9 to stimulate chondrocyte-specific transcription. *Mol Biol Cell* **22**: 1300.

Aoki F, Worrad DM, Schultz RM. 1997. Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. *Dev Biol* **181**: 296–307.

Arsenian S, Weinhold B, Oelgeschläger M, Rüther U, Nordheim A. 1998. Serum response factor is essential for mesoderm formation during mouse embryogenesis. *EMBO J* **17**: 6289–6299.

Attisano L, Lee-Hoeflich ST. 2001. The Smads. *Genome Biol* **2**: REVIEWS3010.

Bachvarova R. 1988. Small B2 RNAs in mouse oocytes, embryos, and somatic tissues. *Dev Biol* **130**: 513–523.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Res* **43**: W39–W49.

Bailly-Bechet M, Haudry A, Lerat E. 2014. “One code to find them all”: a perl tool to conveniently parse RepeatMasker output files. *Mob DNA* **5**: 13.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Belaguli NS, Zhou W, Trinh T-HT, Majesky MW, Schwartz RJ. 1999. Dominant Negative Murine Serum Response Factor: Alternative Splicing within the Activation Domain Inhibits Transactivation of Serum Response Factor Binding Targets. *Mol Cell Biol* **19**: 4582–4591.

Bénit L, Lallemand JB, Casella JF, Philippe H, Heidmann T. 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J Virol* **73**: 3301–3308.

Biggers JD. 1998. Reflections on the culture of the preimplantation embryo. *Int J Dev Biol* **42**: 879–884.

Boeke JD, Corces VG. 1989. Transcription and reverse transcription of retrotransposons. *Annu Rev Microbiol* **43**: 403–434.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma Oxf Engl* **30**: 2114–2120.

Bošković A, Eid A, Pontabry J, Ishiuchi T, Spiegelhalter C, Raghu Ram EVS, Meshorer E, Torres-Padilla M-E. 2014. Higher chromatin mobility supports totipotency and precedes pluripotency in vivo. *Genes Dev* **28**: 1042–1047.

Bosnakovski D, Gearhart MD, Ho Choi S, Kyba M. 2021. Dux facilitates post-implantation development, but is not essential for zygotic genome activation†. *Biol Reprod* **104**: 83–93.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**: 1752–1762.

Brou C, Wu J, Ali S, Scheer E, Lang C, Davidson I, Chambon P, Tora L. 1993. Different TBP-associated factors are required for mediating the stimulation of transcription in vitro by the acidic transactivator GAL-VP16 and the two nonacidic activation functions of the estrogen receptor. *Nucleic Acids Res* **21**: 5.

Buratowski S, Hahn S, Guarente L, Sharp PA. 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**: 549–561.

Buratowski S, Sopta M, Greenblatt J, Sharp PA. 1991. RNA polymerase II-associated proteins are required for a DNA conformation change in the transcription initiation complex. *Proc Natl Acad Sci U S A* **88**: 7509–7513.

Burns KH. 2017. Transposable elements in cancer. *Nat Rev Cancer* **17**: 415–424.

Burton A, Brochard V, Galan C, Ruiz-Morales ER, Rovira Q, Rodriguez-Terrones D, Kruse K, Le Gras S, Udayakumar VS, Chin HG, et al. 2020. Heterochromatin establishment during early mammalian development is regulated by pericentromeric RNA and characterized by non-repressive H3K9me3. *Nat Cell Biol* **22**: 767–778.

Burton A, Torres-Padilla M-E. 2014. Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. *Nat Rev Mol Cell Biol* **15**: 723–735.

Burton A, Torres-Padilla M-E. 2010. Epigenetic reprogramming and development: a unique heterochromatin organization in the preimplantation mouse embryo. *Brief Funct Genomics* **9**: 444–454.

Butler JEF, Kadonaga JT. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**: 2583–2592.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma Oxf Engl* **25**: 1972–1973.

Carter TA, Singh M, Dumbović G, Chobirko JD, Rinn JL, Feschotte C. 2022. Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. *eLife* **11**: e76257.

Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, Duc J, Jang SM, Turelli P, Trono D. 2014. Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev* **28**: 1397–1409.

Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. 2015. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biol Evol* **7**: 567–580.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (eds. U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo), pp. 207–232, Springer, Berlin. doi:10.1007/978-3-540-35306-5_10.

Cheloufi S, Elling U, Hopfgartner B, Jung YL, Murn J, Ninova M, Hubmann M, Badeaux AI, Euong Ang C, Tenen D, et al. 2015. The histone chaperone CAF-1 safeguards somatic cell identity. *Nature* **528**: 218–224.

Chen C, Ara T, Gautheret D. 2009. Using Alu elements as polyadenylation sites: A case of retroposon exaptation. *Mol Biol Evol* **26**: 327–334.

Chen FX, Smith ER, Shilatifard A. 2018. Born to run: control of transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **19**: 464–478.

Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J. 2023. Context-aware transcript quantification from long-read RNA-seq data with Bambu. *Nat Methods* **20**: 1187–1195.

Chen Z, Zhang Y. 2019. Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nat Genet* **51**: 947–951.

Choi SH, Gearhart MD, Cui Z, Bosnakovski D, Kim M, Schennum N, Kyba M. 2016. DUX4 recruits p300/CBP through its C-terminus and induces global H3K27 acetylation changes. *Nucleic Acids Res* **44**: 5161–5173.

Choi YJ, Lin C-P, Risso D, Chen S, Kim TA, Tan MH, Li JB, Wu Y, Chen C, Xuan Z, et al. 2017. Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells. *Science* **355**: eaag1927.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**: 71–86.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**: 1083–1087.

Clift D, McEwan WA, Labzin LI, Konieczny V, Mogessie B, James LC, Schuh M. 2017. A Method for the Acute and Rapid Degradation of Endogenous Proteins. *Cell* **171**: 1692-1706.e18.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**: 105–114.

Costas J. 2003. Molecular characterization of the recent intragenomic spread of the murine endogenous retrovirus MuERV-L. *J Mol Evol* **56**: 181–186.

Cusack M, King HW, Spingardi P, Kessler BM, Klose RJ, Kriaucionis S. 2020. Distinct contributions of DNA methylation and histone acetylation to the genomic occupancy of transcription factors. *Genome Res* **30**: 1393–1406.

De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D. 2017. DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet* **49**: 941–945.

De Iaco A, Verp S, Offner S, Grun D, Trono D. 2020. DUX is a non-essential synchronizer of zygotic genome activation. *Development* **147**: dev177725.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**: e1002384.

DeBerardinis RJ, Kazazian HH. 1999. Analysis of the promoter from an expanding mouse retrotransposon subfamily. *Genomics* **56**: 317–323.

Demircioğlu D, Cukuroglu E, Kindermans M, Nandi T, Calabrese C, Fonseca NA, Kahles A, Lehmann K-V, Stegle O, Brazma A, et al. 2019. A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell* **178**: 1465-1477.e17.

Deng Q, Ramsköld D, Reinius B, Sandberg R. 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**: 193–196.

Deshpande A, Shetty PMV, Frey N, Rangrez AY. 2022. SRF: a seriously responsible factor in cardiac development and disease. *J Biomed Sci* **29**: 38.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**: 41–48.

Dewannieux M, Heidmann T. 2005. L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *J Mol Biol* **349**: 241–247.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.

Du AY, Chobirko JD, Zhuo X, Feschotte C, Wang T. 2024. Regulatory transposable elements in the encyclopedia of DNA elements. *Nat Commun* **15**: 7594.

Dynlacht BD, Hoey T, Tjian R. 1991. Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. *Cell* **66**: 563–576.

Ecco G, Imbeault M, Trono D. 2017. KRAB zinc finger proteins. *Dev Camb Engl* **144**: 2719–2729.

Eckersley-Maslin MA, Alda-Catalinas C, Reik W. 2018. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nat Rev Mol Cell Biol* **19**: 436–450.

Eickbush TH, Malik HS. 2007. Origins and Evolution of Retrotransposons. In *Mobile DNA II*, pp. 1111–1144. (eds N.L. Craig, R. Craigie, M. Gellert and A.M. Lambowitz).

Eidahl JO, Giesige CR, Domire JS, Wallace LM, Fowler AM, Guckes SM, Garwick-Coppens SE, Labhart P, Harper SQ. 2016. Mouse Dux is myotoxic and shares partial functional homology with its human paralog DUX4. *Hum Mol Genet* **25**: 4577–4589.

Evsikov AV, de Vries WN, Peaston AE, Radford EE, Fancher KS, Chen FH, Blake JA, Bult CJ, Latham KE, Solter D, et al. 2004. Systems biology of the 2-cell mouse embryo. *Cytogenet Genome Res* **105**: 240–250.

Fadloun A, Le Gras S, Jost B, Ziegler-Birling C, Takahashi H, Gorab E, Carninci P, Torres-Padilla M-E. 2013. Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat Struct Mol Biol* **20**: 332–338.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**: 563–571.

Ferrigno O, Virolle T, Djabari Z, Ortonne JP, White RJ, Aberdam D. 2001. Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat Genet* **28**: 77–81.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397–405.

Festuccia N, Vandormael-Pournin S, Chervova A, Geiselmann A, Langa-Vives F, Coux R-X, Gonzalez I, Collet GG, Cohen-Tannoudji M, Navarro P. 2024. Nr5a2 is dispensable for zygotic genome activation but essential for morula development. *Science* **386**: eadg7325.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet TIG* **5**: 103–107.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**: 9451–9457.

Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, et al. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet* **46**: 558–566.

Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, Kuzman M, Fulka H, Cernohorska M, Urbanova J, et al. 2017. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res* gr.216150.116.

Furano AV. 2000. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Prog Nucleic Acid Res Mol Biol* **64**: 255–294.

Gassler J, Kobayashi W, Gáspár I, Ruangroengkulrith S, Mohanan A, Gómez Hernández L, Kravchenko P, Kümmecke M, Lalic A, Rifel N, et al. 2022. Zygotic genome activation by the totipotency pioneer factor Nr5a2. *Science* **378**: 1305–1315.

Gazdag E, Rajkovic A, Torres-Padilla ME, Tora L. 2007. Analysis of TATA-binding protein 2 (TBP2) and TBP expression suggests different roles for the two proteins in regulation of gene expression during oogenesis and early mouse development. *Reprod Camb Engl* **134**: 51–62.

Gazdag E, Santenard A, Ziegler-Birling C, Altobelli G, Poch O, Tora L, Torres-Padilla M-E. 2009. TBP2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte. *Genes Dev* **23**: 2210–2223.

Genet M, Torres-Padilla M-E. 2020. The molecular and cellular features of 2-cell-like cells: a reference guide. *Development* **147**: dev189688.

Geng LN, Yao Z, Snider L, Fong AP, Cech JN, Young JM, van der Maarel SM, Ruzzo WL, Gentleman RC, Tawil R, et al. 2012. DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev Cell* **22**: 38–51.

Gifford WD, Pfaff SL, Macfarlan TS. 2013. Transposable elements as genetic regulatory substrates in early development. *Trends Cell Biol* **23**: 218–226.

Göke J, Lu X, Chan Y-S, Ng H-H, Ly L-H, Sachs F, Szczerbinska I. 2015. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**: 135–141.

Goodier JL. 2016. Restricting retrotransposons: a review. *Mob DNA* **7**: 16.

Goodier JL, Kazazian HH. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**: 23–35.

Goodier JL, Ostertag EM, Du K, Kazazian HH. 2001. A Novel Active L1 Retrotransposon Subfamily in the Mouse. *Genome Res* **11**: 1677–1685.

Greenberg ME, Siegfried Z, Ziff EB. 1987. Mutation of the c-fos Gene Dyad Symmetry Element Inhibits Serum Inducibility of Transcription In Vivo and the Nuclear Regulatory Factor Binding In Vitro. *Mol Cell Biol* **7**: 1217–1225.

Greenberg ME, Ziff EB. 1984. Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene. *Nature* **311**: 433–438.

Greenberg MVC, Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**: 590–607.

Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**: 221–225.

Guo F, Li X, Liang D, Li T, Zhu P, Guo H, Wu X, Wen L, Gu T-P, Hu B, et al. 2014. Active and Passive Demethylation of Male and Female Pronuclear DNA in the Mammalian Zygote. *Cell Stem Cell* **15**: 447–459.

Guo M, Zhang Y, Zhou J, Bi Y, Xu J, Xu C, Kou X, Zhao Y, Li Y, Tu Z, et al. 2019. Precise temporal regulation of Dux is important for embryo development. *Cell Res* **29**: 956–959.

Guo Y, Kitano T, Inoue K, Murano K, Hirose M, Li TD, Sakashita A, Ishizu H, Ogonuki N, Matoba S, et al. 2024. Obox4 promotes zygotic genome activation upon loss of Dux ed. F. Lu. *eLife* **13**: e95856.

Hamatani T, Carter MG, Sharov AA, Ko MSH. 2004. Dynamics of global gene expression changes during mouse preimplantation development. *Dev Cell* **6**: 117–131.

Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**: 9.

Hendrickson PG, Doráis JA, Grow EJ, Whiddon JL, Lim J-W, Wike CL, Weaver BD, Pflueger C, Emery BR, Wilcox AL, et al. 2017. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat Genet* **49**: 925–934.

Hermant C, Torres-Padilla M-E. 2021. TFs for TEs: the transcription factor repertoire of mammalian transposable elements. *Genes Dev* **35**: 22–39.

Home P, Ray S, Dutta D, Bronshteyn I, Larson M, Paul S. 2009. GATA3 is selectively expressed in the trophectoderm of peri-implantation embryo and directly regulates Cdx2 gene expression. *J Biol Chem* **284**: 28729–28737.

Honda S, Hatamura M, Kunimoto Y, Ikeda S, Minami N. 2024. Chimeric PRMT6 protein produced by an endogenous retrovirus promoter regulates cell fate decision in mouse preimplantation embryos†. *Biol Reprod* **110**: 698–710.

Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. 2015. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **43**: D117-122.

Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**: 550–554.

Ishiuchi T, Enriquez-Gasca R, Mizutani E, Bošković A, Ziegler-Birling C, Rodriguez-Terrones D, Wakayama T, Vaquerizas JM, Torres-Padilla M-E. 2015. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat Struct Mol Biol* **22**: 662–671.

Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, Inoue I. 2017. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet* **13**: e1006883.

Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. 2017. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet* **49**: 1502–1510.

Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**: 242–245.

Jacques P-É, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* **9**: e1003504.

Jansz N. 2019. DNA methylation dynamics at transposable elements in mammals. *Essays Biochem* **63**: 677–689.

Ji S, Chen F, Stein P, Wang J, Zhou Z, Wang L, Zhao Q, Lin Z, Liu B, Xu K, et al. 2023. OBOX regulates mouse zygotic genome activation and early development. *Nature* **620**: 1047–1053.

Jiang J-C, Upton KR. 2019. Human transposons are an abundant supply of transcription factor binding sites and promoter activities in breast cancer cell lines. *Mob DNA* **10**: 16.

Jin Y, Tam OH, Paniagua E, Hammell M. 2015. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinforma Oxf Engl* **31**: 3593–3599.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet TIG* **19**: 68–72.

Jukam D, Shariati SAM, Skotheim JM. 2017. Zygotic Genome Activation in Vertebrates. *Dev Cell* **42**: 316–332.

Kazazian HH, Moran JV. 2017. Mobile DNA in Health and Disease. *N Engl J Med* **377**: 361–370.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13**: R107.

Kigami D, Minami N, Takayama H, Imai H. 2003. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biol Reprod* **68**: 651–654.

Kim J, Iyer VR. 2004. Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. *Mol Cell Biol* **24**: 8104–8112.

Kinene T, Wainaina J, Maina S, Boykin LM. 2016. Rooting Trees, Methods for. *Encycl Evol Biol* 489–493.

Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. *Heredity* **107**: 487–495.

Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**: 631–634.

Kwak H, Lis JT. 2013. Control of Transcriptional Elongation. *Annu Rev Genet* **47**: 483–508.

Kwan JZ, Nguyen TF, Uzozie AC, Budzynski MA, Cui J, Lee JM, Van Petegem F, Lange PF, Teves SS. 2023. RNA Polymerase II transcription independent of TBP in murine embryonic stem cells eds. I. Davidson, K. Struhl, and I. Davidson. *eLife* **12**: e83810.

Lai F, Li L, Hu X, Liu B, Zhu Z, Liu L, Fan Q, Tian H, Xu K, Lu X, et al. 2023. NR5A2 connects zygotic genome activation to the first lineage segregation in totipotent embryos. *Cell Res* **33**: 952–966.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction. *Methods Ecol Evol* **6**: 1110–1116.

Letunic I, Bork P. 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* gkae268.

Liu H, Kim J-M, Aoki F. 2004. Regulation of histone H3 lysine 9 methylation in oocytes and early pre-implantation embryos. *Development* **131**: 2269–2280.

Loeb DD, Padgett RW, Hardies SC, Shehee WR, Comer MB, Edgell MH, Hutchison CA. 1986. The sequence of a large L1Md element reveals a tandemly repeated 5’ end and several features found in retrotransposons. *Mol Cell Biol* **6**: 168–182.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**: 57–63.

Mager DL, Stoye JP. 2015. Mammalian Endogenous Retroviruses. *Microbiol Spectr* **3**: MDNA3-0009–2014.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**: 793–805.

Martianov I, Viville S, Davidson I. 2002. RNA polymerase II transcription in murine cells lacking the TATA binding protein. *Science* **298**: 1036–1039.

Mcclintock B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21**: 197–216.

McClintock B. 1950. The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci U S A* **36**: 344–355.

Miano JM. 2010. Role of serum response factor in the pathogenesis of disease. *Lab Invest* **90**: 1274–1284.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**: 1530–1534.

Mintz B. 1964. SYNTHETIC PROCESSES AND EARLY DEVELOPMENT IN THE MAMMALIAN EGG. *J Exp Zool* **157**: 85–100.

Modzelewski AJ, Shao W, Chen J, Lee A, Qi X, Noon M, Tjokro K, Sales G, Biton A, Anand A, et al. 2021. A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform essential for development. *Cell* **184**: 5541-5558.e22.

Naas TP, DeBerardinis RJ, Moran JV, Ostertag EM, Kingsmore SF, Seldin MF, Hayashizaki Y, Martin SL, Kazazian HH. 1998. An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO J* **17**: 590–597.

Nakatani T, Lin J, Ji F, Ettinger A, Pontabry J, Tokoro M, Altamirano-Pacheco L, Fiorentino J, Mahammadov E, Hatano Y, et al. 2022. DNA replication fork speed underlies cell fate changes and promotes reprogramming. *Nat Genet* **54**: 318–327.

Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **37**: D77-82.

Nishiyama A, Xin L, Sharov AA, Thomas M, Mowrer G, Meyers E, Piao Y, Mehta S, Yee S, Nakatake Y, et al. 2009. Uncovering early response of gene regulatory networks in ES cells by systematic induction of transcription factors. *Cell Stem Cell* **5**: 420.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature* **485**: 381–385.

Norman C, Runswick M, Pollock R, Treisman R. 1988. Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. *Cell* **55**: 989–1003.

Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, Nakamura M, Tokunaga Y, Nakamura M, Watanabe A, et al. 2014. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc Natl Acad Sci* **111**: 12426–12431.

Oliveira DS, Fablet M, Larue A, Vallier A, Carareto CMA, Rebollo R, Vieira C. 2023. ChimeraTE: a pipeline to detect chimeric transcripts derived from genes and transposable elements. *Nucleic Acids Res* **51**: 9764–9784.

Oomen ME, Rodriguez-Terrones D, Kurome M, Zakhartchenko V, Mottes L, Simmet K, Noll C, Nakatani T, Mourra-Diaz CM, Aksoy I, et al. 2025. An atlas of transcription initiation reveals regulatory principles of gene and transposable element expression in early mammalian development. *Cell* S0092-8674(24)01426–0.

Oomen ME, Torres-Padilla M-E. 2024. Jump-starting life: balancing transposable element co-option and genome integrity in the developing mammalian embryo. *EMBO Rep* **25**: 1721–1733.

Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, Moreno-Santillán DD, Crookshanks C, Roberts J, Garcia C, et al. 2023. Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science* **380**: eabn1430.

Padeken J, Methot SP, Gasser SM. 2022. Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. *Nat Rev Mol Cell Biol* **23**: 623–640.

Padgett RW, Hutchison CA, Edgell MH. 1988. The F-type 5' motif of mouse L1 elements: a major class of L1 termini similar to the A-type in organization but unrelated in sequence. *Nucleic Acids Res* **16**: 739–749.

Paria BC, Dey SK. 1990. Preimplantation embryo development in vitro: cooperative interactions among embryos and role of growth factors. *Proc Natl Acad Sci U S A* **87**: 4756–4760.

Park S-J, Shirahige K, Ohsugi M, Nakai K. 2015. DBTMEE: a database of transcriptome in mouse early embryos. *Nucleic Acids Res* **43**: D771-776.

Passmore S, Elble R, Tye BK. 1989. A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes. *Genes Dev* **3**: 921–935.

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* **7**: 597–606.

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**: 1096–1098.

Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**: 171–181.

Platanitis E, Demiroz D, Schneller A, Fischer K, Capelle C, Hartl M, Gossenreiter T, Müller M, Novatchkova M, Decker T. 2019. A molecular switch from STAT2-IRF9 to ISGF3 underlies interferon-induced gene transcription. *Nat Commun* **10**: 2921.

Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. 2019. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**: 724-735.e5.

Poon D, Weil PA. 1993. Immunopurification of yeast TATA-binding protein and associated factors. Presence of transcription factor IIIB transcriptional activity. *J Biol Chem* **268**: 15325–15328.

Puschendorf M, Terranova R, Boutsma E, Mao X, Isono K, Brykczynska U, Kolb C, Otte AP, Koseki H, Orkin SH, et al. 2008. PRC1 and Suv39h specify parental asymmetry at constitutive heterochromatin in early mouse embryos. *Nat Genet* **40**: 411–420.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma Oxf Engl* **26**: 841–842.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160-165.

Rappolee DA, Brenner CA, Schultz R, Mark D, Werb Z. 1988. Developmental expression of PDGF, TGF-alpha, and TGF-beta genes in preimplantation mouse embryos. *Science* **241**: 1823–1825.

Reinberg D, Horikoshi M, Roeder RG. 1987. Factors involved in specific transcription in mammalian RNA polymerase II. Functional analysis of initiation factors IIA and IID and identification of a new factor operating at sequences downstream of the initiation site. *J Biol Chem* **262**: 3322–3330.

Rodriguez-Terrones D, Gaume X, Ishiuchi T, Weiss A, Kopp A, Kruse K, Penning A, Vaquerizas JM, Brino L, Torres-Padilla M-E. 2018. A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat Genet* **50**: 106–119.

Rodriguez-Terrones D, Torres-Padilla M-E. 2018. Nimble and Ready to Mingle: Transposon Outbursts of Early Development. *Trends Genet* **34**: 806–820.

Rosmarin AG, Resendes KK, Yang Z, McMillan JN, Fleming SL. 2004. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein–protein interactions. *Blood Cells Mol Dis* **32**: 143–154.

Rosspopoff O, Trono D. 2023. Take a walk on the KRAB side. *Trends Genet* **39**: 844–857.

Rowe HM, Trono D. 2011. Dynamic control of endogenous retroviruses during development. *Virology* **411**: 273–287.

Royall AH, Maeso I, Dunwell TL, Holland PWH. 2018. Mouse Obox and Crxos modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes. *EvoDevo* **9**: 2.

Roy-Engel AM, El-Sawy M, Farooq L, Odom GL, Perepelitsa-Belancio V, Bruch H, Oyeniran OO, Deininger PL. 2005. Human retroelements may introduce intragenic polyadenylation signals. *Cytogenet Genome Res* **110**: 365–371.

Sakamoto M, Ishiuchi T. 2024. YY1-dependent transcriptional regulation manifests at the morula stage. *MicroPublication Biol* **2024**.

Sakashita A, Kitano T, Ishizu H, Guo Y, Masuda H, Ariura M, Murano K, Siomi H. 2023. Transcription of MERVL retrotransposons is required for preimplantation embryo development. *Nat Genet* **55**: 484–495.

Santana-Garcia W, Castro-Mondragon JA, Padilla-Gálvez M, Nguyen NTT, Elizondo-Salas A, Ksouri N, Gerbes F, Thieffry D, Vincens P, Contreras-Moreira B, et al. 2022. RSAT 2022: regulatory sequence analysis tools. *Nucleic Acids Res* **50**: W670–W676.

Santoni FA, Guerra J, Luban J. 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**: 111.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348.

Schmidt K, Glaser G, Wernig A, Wegner M, Rosorius O. 2003. Sox8 Is a Specific Marker for Muscle Satellite Cells and Inhibits Myogenesis *. *J Biol Chem* **278**: 29769–29775.

Schramm L, Hernandez N. 2002. Recruitment of RNA polymerase III to its target promoters. *Genes Dev* **16**: 2593–2620.

Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H. 1990. Genetic Control of Flower Development by Homeotic Genes in Antirrhinum majus. *Science* **250**: 931–936.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-Dimensional Folding and Functional Organization Principles of the *Drosophila* Genome. *Cell* **148**: 458–472.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.

Smit, AFA, Hubley R, Green P. 2013. RepeatMasker Open-4.0. <http://www.repeatmasker.org> (Accessed June 28, 2024).

Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, Eggan K, Meissner A. 2014. DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**: 611–615.

Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, Meissner A. 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**: 339–344.

Sommer H, Beltrán JP, Huijser P, Pape H, Lönnig WE, Saedler H, Schwarz-Sommer Z. 1990. Deficiens, a homeotic gene involved in the control of flower morphogenesis in Antirrhinum majus: the protein shows homology to transcription factors. *EMBO J* **9**: 605–613.

Steel M. 2010. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing edited by LEMEY, P., SALEMI, M., and VANDAMME, A.-M. *Biometrics* **66**: 324–325.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**: 2.

Sun X, Wang X, Tang Z, Grivainis M, Kahler D, Yun C, Mita P, Fenyö D, Boeke JD. 2018. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc Natl Acad Sci U S A* **115**: E5526–E5535.

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1963–1976.

Svoboda P, Stein P, Anger M, Bernstein E, Hannon GJ, Schultz RM. 2004. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Dev Biol* **269**: 276–285.

Tagliaferri D, Mazzone P, Noviello TMR, Addeo M, Angrisano T, Del Vecchio L, Visconte F, Ruggieri V, Russi S, Caivano A, et al. 2019. Retinoic Acid Induces Embryonic Stem Cells (ESCs) Transition to 2 Cell-Like State Through a Coordinated Expression of Dux and Duxbl1. *Front Cell Dev Biol* **7**: 385.

Tarasewicz E, Jeruss JS. 2012. Phospho-specific Smad3 signaling: impact on breast oncogenesis. *Cell Cycle Georget Tex* **11**: 2443–2451.

Taylor KD, Pikó L. 1987. Patterns of mRNA prevalence and expression of B1 and B2 transcripts in early mouse embryos. *Development* **101**: 877–892.

Teves SS, An L, Bhargava-Shah A, Xie L, Darzacq X, Tjian R. 2018. A stable mode of bookmarking by TBP recruits RNA polymerase II to mitotic chromosomes ed. J.M. Espinosa. *eLife* **7**: e35621.

Thomas J, Perron H, Feschotte C. 2018. Variation in proviral content among human genomes mediated by LTR recombination. *Mob DNA* **9**: 36.

Thomas JH, Schneider S. 2011. Coevolution of retroelements and tandem zinc finger genes. *Genome Res* **21**: 1800–1812.

Torres-Padilla M-E. 2020. On transposons and totipotency. *Philos Trans R Soc B Biol Sci* **375**: 20190339.

Torres-Padilla M-E, Bannister AJ, Hurd PJ, Kouzarides T, Zernicka-Goetz M. 2006. Dynamic distribution of the replacement histone variant H3.3 in the mouse oocyte and preimplantation embryos. *Int J Dev Biol* **50**: 455–461.

Trauth K, Mutschler B, Jenkins NA, Gilbert DJ, Copeland NG, Klempnauer KH. 1994. Mouse A-myb encodes a trans-activator and is expressed in mitotically active cells of the developing central nervous system, adult testis and B lymphocytes. *EMBO J* **13**: 5994–6005.

Treisman R. 1987. Identification and purification of a polypeptide that binds to the c-fos serum response element. *EMBO J* **6**: 2711–2717.

Treisman R. 1986. Identification of a protein-binding site that mediates transcriptional response of the c-fos gene to serum factors. *Cell* **46**: 567–574.

van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet TIG* **19**: 530–536.

Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J. 2016. Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **13**: 7.

Veselovska L, Smallwood SA, Saadeh H, Stewart KR, Krueger F, Maupetit-Méhouas S, Arnaud P, Tomizawa S, Andrews S, Kelsey G. 2015. Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome Biol* **16**: 209.

Vogt VM. 1997. Retroviral virions and genomes. In: Retroviruses (ed. Coffin JM et al.), p 27–69. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. https://www.ncbi.nlm.nih.gov/books/NBK19454

Voliva CF, Jahn CL, Comer MB, Hutchison CA, Edgell MH. 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. *Nucleic Acids Res* **11**: 8847–8859.

Wang C, Liu X, Gao Y, Yang L, Li C, Liu W, Chen C, Kou X, Zhao Y, Chen J, et al. 2018. Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20**: 620–631.

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al. 2014a. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**: 405–409.

Wang L, Zhang J, Duan J, Gao X, Zhu W, Lu X, Yang L, Zhang J, Li G, Ci W, et al. 2014b. Programming and Inheritance of Parental DNA Methylomes in Mammals. *Cell* **157**: 979–991.

Wang S, Sengel C, Emerson MM, Cepko CL. 2014c. A gene regulatory network controls the binary fate decision of rod and bipolar cells in the vertebrate retina. *Dev Cell* **30**: 513–527.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* **104**: 18613–18618.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinforma Oxf Engl* **25**: 1189–1191.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Whiddon JL, Langford AT, Wong C-J, Zhong JW, Tapscott SJ. 2017. Conservation and innovation in the DUX4-family gene network. *Nat Genet* **49**: 935–940.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973–982.

Wickham H. 2016. *ggplot2*. Springer International Publishing, Cham http://link.springer.com/10.1007/978-3-319-24277-4 (Accessed June 28, 2024).

Wildt KF, Sun G, Grueter B, Fischer M, Zamisch M, Ehlers M, Bosselut R. 2007. The Transcription Factor Zbtb7b Promotes CD4 Expression by Antagonizing Runx-Mediated Activation of the CD4 Silencer1. *J Immunol* **179**: 4405–4414.

Wilming LG, Boychenko V, Harrow JL. 2015. Comprehensive comparative homeobox gene annotation in human and mouse. *Database J Biol Databases Curation* **2015**: bav091.

Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, et al. 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**: 652–657.

Yamazaki S, Nakano N, Honjo A, Hara M, Maeda K, Nishiyama C, Kitaura J, Ohtsuka Y, Okumura K, Ogawa H, et al. 2015. The Transcription Factor Ehf Is Involved in TGF-β-Induced Suppression of FcεRI and c-Kit Expression and FcεRI-Mediated Activation in Mast Cells. *J Immunol Baltim Md 1950* **195**: 3427–3435.

Yang J, Cook L, Chen Z. 2024. Systematic evaluation of retroviral LTRs as cis-regulatory elements in mouse embryos. *Cell Rep* **43**: 113775.

Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM. 1990. The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. *Nature* **346**: 35–39.

Young JM, Whiddon JL, Yao Z, Kasinathan B, Snider L, Geng LN, Balog J, Tawil R, van der Maarel SM, Tapscott SJ. 2013. DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. *PLoS Genet* **9**: e1003947.

Yu C, Cvetesic N, Hisler V, Gupta K, Ye T, Gazdag E, Negroni L, Hajkova P, Berger I, Lenhard B, et al. 2020. TBPL2/TFIIA complex establishes the maternal transcriptome through oocyte-specific promoter usage. *Nat Commun* **11**: 6439.

Zhang W, Chen F, Chen R, Xie D, Yang J, Zhao X, Guo R, Zhang Y, Shen Y, Göke J, et al. 2019a. Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. *Nucleic Acids Res* **47**: 8485–8501.

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al. 2019b. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**: 1380–1388.

Zhou M, Smith AD. 2019. Subtype classification and functional annotation of L1Md retrotransposon promoters. *Mob DNA* **10**: 14.

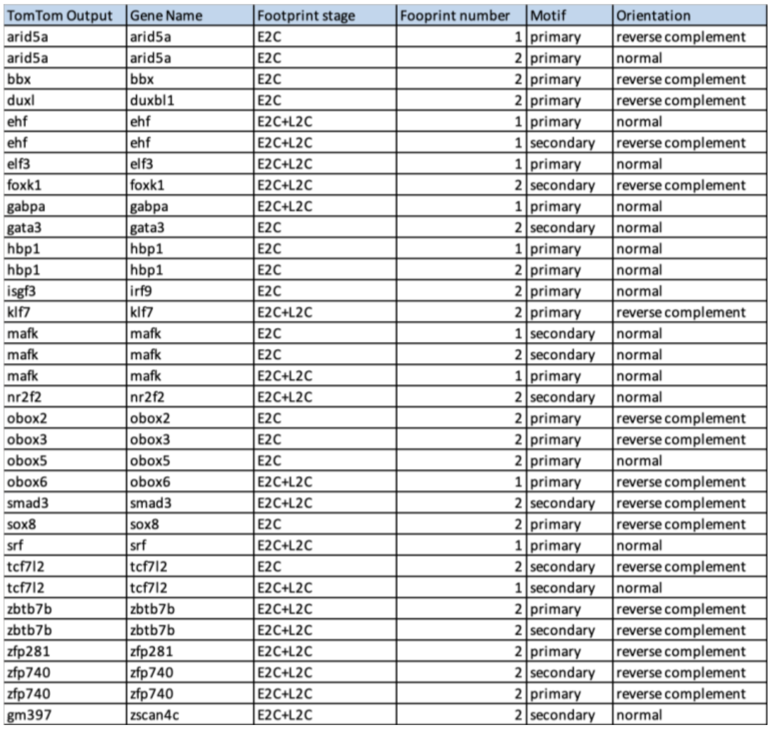
Zou Z, Zhang C, Wang Q, Hou Z, Xiong Z, Kong F, Wang Q, Song J, Liu B, Liu B, et al. 2022. Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation. *Science* **378**: abo7923.

8 List of Abbreviations

Abbreviation	Definition	Abbreviation	Definition
2CLC	2-cell like cells	**MII**	Metaphase II
ASO	Antisense oligonucleotide	**mRNA**	Messenger RNA
ATAC-seq	Assay for tranposase accessibility chromatin using sequencing	**MSA**	Multiple sequence alignment
Bp	Basepairs	**MYA**	Million years ago
C-ter	C-terminal	**MZT**	Maternal to Zygotic transition
CDS	Coding sequence	**N-ter**	N-terminal
CTD	C-terminal domain	**Orf**	Open reading frame
DBTMEE	Database of transcriptome in mouse early embryo	**Padj**	Adjusted p-value
DN	Dominant-negative	**PCA**	Principal component analysis
DNA	Deoxyribonucleic Acid	**PGC**	Primordial Germ Cell
ERV	Endogenous retrovirus	**qPCR**	Quantitative real-time PCR
FSHD	Facioscapulohumeral muscular dystrophy	**RNA**	Ribonucleic Acid
GOF	Gain-of-function	**RNAPII**	RNA Polymerase II
HAT	Histone acetyltransferase	**RNAPIII**	RNA Polymerase III
hCG	Human chorionic gonadotropin	**Rpm**	Read per million
hESCs	Human embryonic stem cells	**RT**	Reverse Transcriptase
ICM	Inner Cell Mass	**SINE**	Short Interspersed Nuclear Element
IF	Immunofluorescence	**siRNA**	Small interfering RNA
iPSCs	Induced pluripotent stem cells	**SNP**	Single nucleotide polymorphism
KSOM	K-modified simplex optimized	**TAD**	Topologically associating domain
KZFP	Krab-zinc finger proteins	**TE**	Transposable Elements
LINE	Long Interspersed Nuclear Element	**TF**	Transcription Factor
LOF	Loss-of-function	**TFBS**	Transcription factor binding site
LTR	Long Terminal Repeat	**tRNA**	Transfer RNA
MaLR	Mammalian Apparent LTR retrotransposons	**TSS**	Transcription start site
mESCs	Mouse embryonic stem cells		

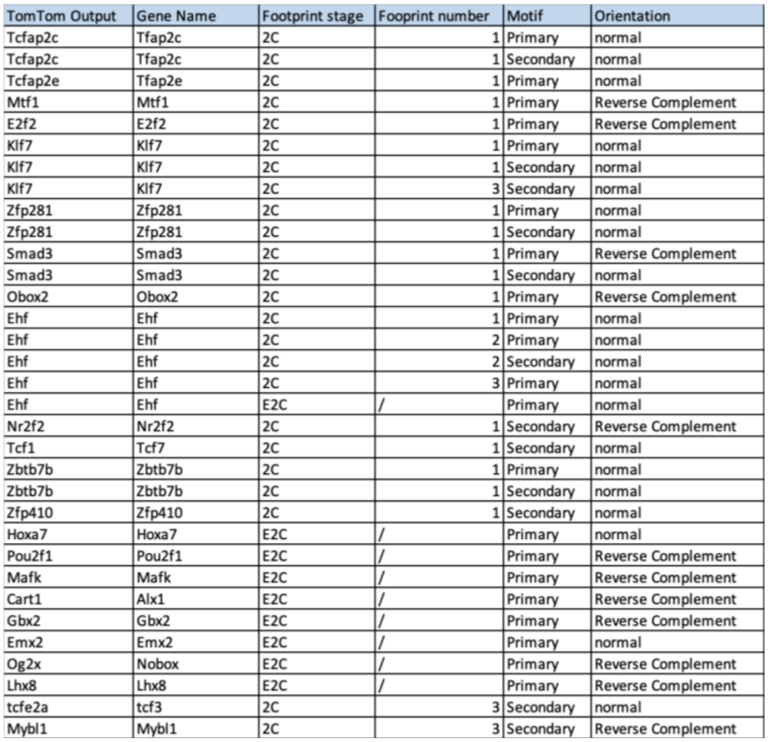
9 Appendices

Appendix 1. Complete lists of TF motifs found per TE family. (A-F) List of TF motifs found in (A) MT2_Mm, (B) L1MdTf_II, (C) B2_Mm1a, (D) B3A, (E) ORR1A1 or (F) ORR1A0, with indicated Tomtom Output, related Gene Name, Stage of footprint identified, Footprint number (if several footprints at the same stage), Type of motif (primary or secondary) and orientation of motif (Normal or Reverse Complement).



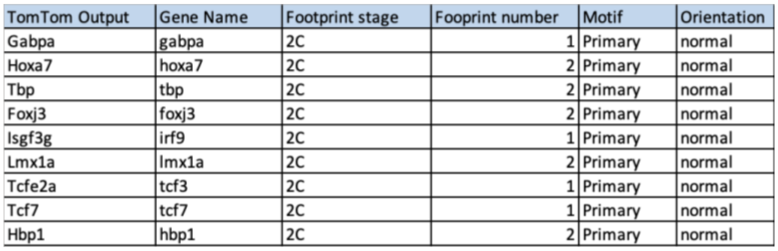
TomTom Output	Gene Name	Footprint stage	Fooprint number	Motif	Orientation
arid5a	arid5a	E2C	1	primary	reverse complement
arid5a	arid5a	E2C	2	primary	normal
bbx	bbx	E2C	2	primary	reverse complement
duxl	duxbl1	E2C	2	primary	reverse complement
ehf	ehf	E2C+L2C	1	primary	normal
ehf	ehf	E2C+L2C	1	secondary	reverse complement
elf3	elf3	E2C+L2C	1	primary	normal
foxk1	foxk1	E2C+L2C	2	secondary	reverse complement
gabpa	gabpa	E2C+L2C	1	primary	normal
gata3	gata3	E2C	2	secondary	normal
hbp1	hbp1	E2C	1	primary	normal
hbp1	hbp1	E2C	2	primary	normal
isgf3	irf9	E2C	2	primary	normal
klf7	klf7	E2C+L2C	2	primary	reverse complement
mafk	mafk	E2C	1	secondary	normal
mafk	mafk	E2C	2	secondary	normal
mafk	mafk	E2C+L2C	1	primary	normal
nr2f2	nr2f2	E2C+L2C	2	secondary	normal
obox2	obox2	E2C	2	primary	reverse complement
obox3	obox3	E2C	2	primary	reverse complement
obox5	obox5	E2C	2	primary	normal
obox6	obox6	E2C+L2C	1	primary	reverse complement
smad3	smad3	E2C+L2C	2	secondary	reverse complement
sox8	sox8	E2C	2	primary	reverse complement
srf	srf	E2C+L2C	1	primary	normal
tcf7l2	tcf7l2	E2C	2	secondary	reverse complement
tcf7l2	tcf7l2	E2C+L2C	1	secondary	normal
zbtb7b	zbtb7b	E2C+L2C	2	primary	reverse complement
zbtb7b	zbtb7b	E2C+L2C	2	secondary	reverse complement
zfp281	zfp281	E2C+L2C	2	primary	reverse complement
zfp740	zfp740	E2C+L2C	2	secondary	reverse complement
zfp740	zfp740	E2C+L2C	2	primary	reverse complement
gm397	zscan4c	E2C+L2C	2	secondary	normal

(A) MT2_Mm



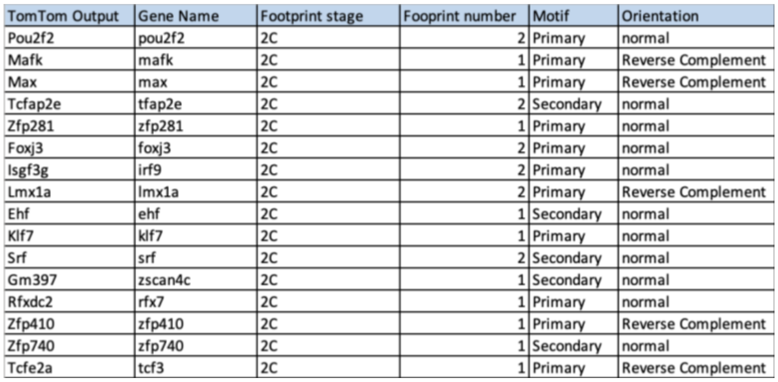
TomTom Output	Gene Name	Footprint stage	Fooprint number	Motif	Orientation
Tcfap2c	Tfap2c	2C	1	Primary	normal
Tcfap2c	Tfap2c	2C	1	Secondary	normal
Tcfap2e	Tfap2e	2C	1	Primary	normal
Mtf1	Mtf1	2C	1	Primary	Reverse Complement
E2f2	E2f2	2C	1	Primary	Reverse Complement
Klf7	Klf7	2C	1	Primary	normal
Klf7	Klf7	2C	1	Secondary	normal
Klf7	Klf7	2C	3	Secondary	normal
Zfp281	Zfp281	2C	1	Primary	normal
Zfp281	Zfp281	2C	1	Secondary	normal
Smad3	Smad3	2C	1	Primary	Reverse Complement
Smad3	Smad3	2C	1	Secondary	normal
Obox2	Obox2	2C	1	Primary	Reverse Complement
Ehf	Ehf	2C	1	Primary	normal
Ehf	Ehf	2C	2	Primary	normal
Ehf	Ehf	2C	2	Secondary	normal
Ehf	Ehf	2C	3	Primary	normal
Ehf	Ehf	E2C	/	Primary	normal
Nr2f2	Nr2f2	2C	1	Secondary	Reverse Complement
Tcf1	Tcf7	2C	1	Secondary	normal
Zbtb7b	Zbtb7b	2C	1	Primary	normal
Zbtb7b	Zbtb7b	2C	1	Secondary	normal
Zfp410	Zfp410	2C	1	Secondary	normal
Hoxa7	Hoxa7	E2C	/	Primary	normal
Pou2f1	Pou2f1	E2C	/	Primary	Reverse Complement
Mafk	Mafk	E2C	/	Primary	Reverse Complement
Cart1	Alx1	E2C	/	Primary	Reverse Complement
Gbx2	Gbx2	E2C	/	Primary	Reverse Complement
Emx2	Emx2	E2C	/	Primary	normal
Og2x	Nobox	E2C	/	Primary	Reverse Complement
Lhx8	Lhx8	E2C	/	Primary	Reverse Complement
tcfe2a	tcf3	2C	3	Secondary	normal
Mybl1	Mybl1	2C	3	Secondary	Reverse Complement

(B) L1MdTf_II



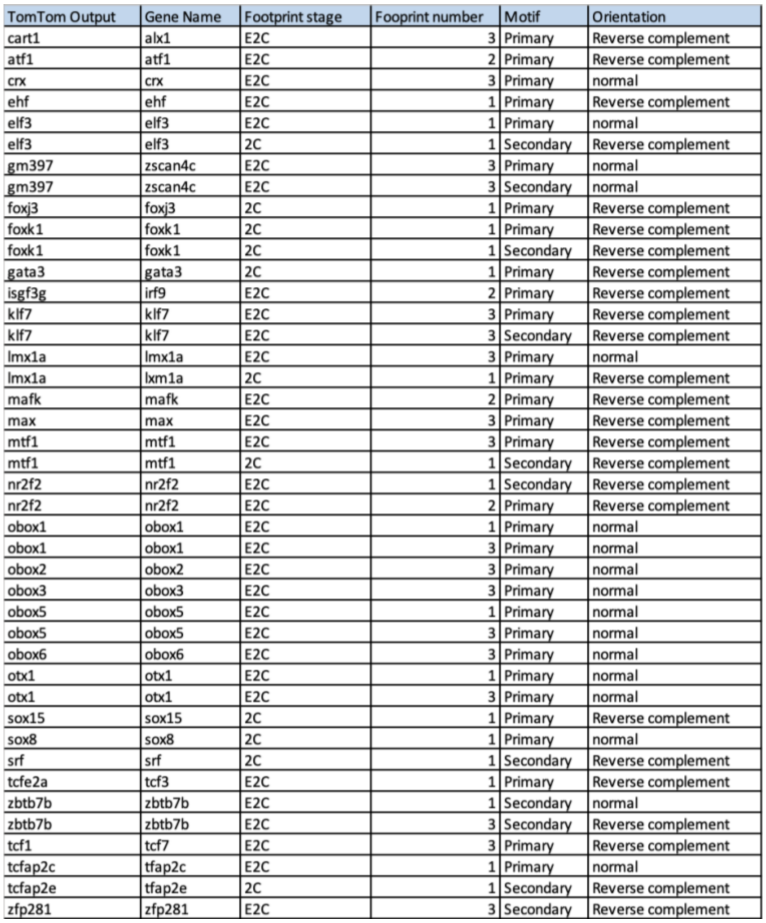
TomTom Output	Gene Name	Footprint stage	Fooprint number	Motif	Orientation
Gabpa	gabpa	2C	1	Primary	normal
Hoxa7	hoxa7	2C	2	Primary	normal
Tbp	tbp	2C	2	Primary	normal
Foxj3	foxj3	2C	2	Primary	normal
Isgf3g	irf9	2C	1	Primary	normal
Lmx1a	lmx1a	2C	2	Primary	normal
Tcfe2a	tcf3	2C	1	Primary	normal
Tcf7	tcf7	2C	1	Primary	normal
Hbp1	hbp1	2C	2	Primary	normal

(C) B2_Mm1a



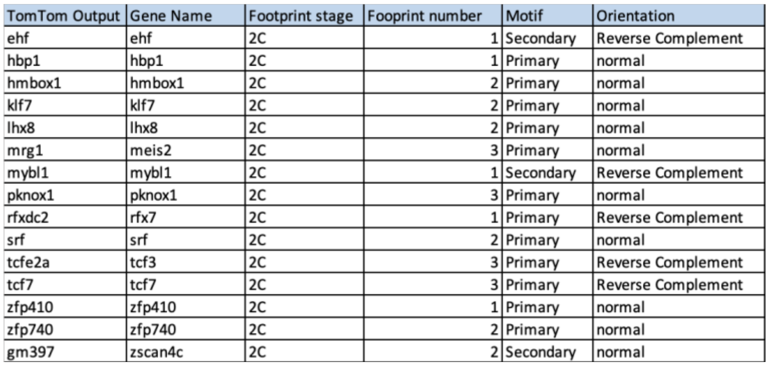
TomTom Output	Gene Name	Footprint stage	Fooprint number	Motif	Orientation
Pou2f2	pou2f2	2C	2	Primary	normal
Mafk	mafk	2C	1	Primary	Reverse Complement
Max	max	2C	1	Primary	Reverse Complement
Tcfap2e	tfap2e	2C	2	Secondary	normal
Zfp281	zfp281	2C	1	Primary	normal
Foxj3	foxj3	2C	2	Primary	normal
Isgf3g	irf9	2C	2	Primary	normal
Lmx1a	lmx1a	2C	2	Primary	Reverse Complement
Ehf	ehf	2C	1	Secondary	normal
Klf7	klf7	2C	1	Primary	normal
Srf	srf	2C	2	Secondary	normal
Gm397	zscan4c	2C	1	Secondary	normal
Rfxdc2	rfx7	2C	1	Primary	normal
Zfp410	zfp410	2C	1	Primary	Reverse Complement
Zfp740	zfp740	2C	1	Secondary	normal
Tcfe2a	tcf3	2C	1	Primary	Reverse Complement

(D) B3A



TomTom Output	Gene Name	Footprint stage	Fooprint number	Motif	Orientation
cart1	alx1	E2C	3	Primary	Reverse complement
atf1	atf1	E2C	2	Primary	Reverse complement
crx	crx	E2C	3	Primary	normal
ehf	ehf	E2C	1	Primary	Reverse complement
elf3	elf3	E2C	1	Primary	normal
elf3	elf3	2C	1	Secondary	Reverse complement
gm397	zscan4c	E2C	3	Primary	normal
gm397	zscan4c	E2C	3	Secondary	normal
foxj3	foxj3	2C	1	Primary	Reverse complement
foxk1	foxk1	2C	1	Primary	Reverse complement
foxk1	foxk1	2C	1	Secondary	Reverse complement
gata3	gata3	2C	1	Primary	Reverse complement
isgf3g	irf9	E2C	2	Primary	Reverse complement
klf7	klf7	E2C	3	Primary	Reverse complement
klf7	klf7	E2C	3	Secondary	Reverse complement
lmx1a	lmx1a	E2C	3	Primary	normal
lmx1a	lxm1a	2C	1	Primary	Reverse complement
mafk	mafk	E2C	2	Primary	Reverse complement
max	max	E2C	3	Primary	Reverse complement
mtf1	mtf1	E2C	3	Primary	Reverse complement
mtf1	mtf1	2C	1	Secondary	Reverse complement
nr2f2	nr2f2	E2C	1	Secondary	Reverse complement
nr2f2	nr2f2	E2C	2	Primary	Reverse complement
obox1	obox1	E2C	1	Primary	normal
obox1	obox1	E2C	3	Primary	normal
obox2	obox2	E2C	3	Primary	normal
obox3	obox3	E2C	3	Primary	normal
obox5	obox5	E2C	1	Primary	normal
obox5	obox5	E2C	3	Primary	normal
obox6	obox6	E2C	3	Primary	normal
otx1	otx1	E2C	1	Primary	normal
otx1	otx1	E2C	3	Primary	normal
sox15	sox15	2C	1	Primary	Reverse complement
sox8	sox8	2C	1	Primary	normal
srf	srf	2C	1	Secondary	Reverse complement
tcfe2a	tcf3	E2C	1	Primary	Reverse complement
zbtb7b	zbtb7b	E2C	1	Secondary	normal
zbtb7b	zbtb7b	E2C	3	Secondary	Reverse complement
tcf1	tcf7	E2C	3	Primary	Reverse complement
tcfap2c	tfap2c	E2C	1	Primary	normal
tcfap2e	tfap2e	2C	1	Secondary	Reverse complement
zfp281	zfp281	E2C	3	Secondary	Reverse complement

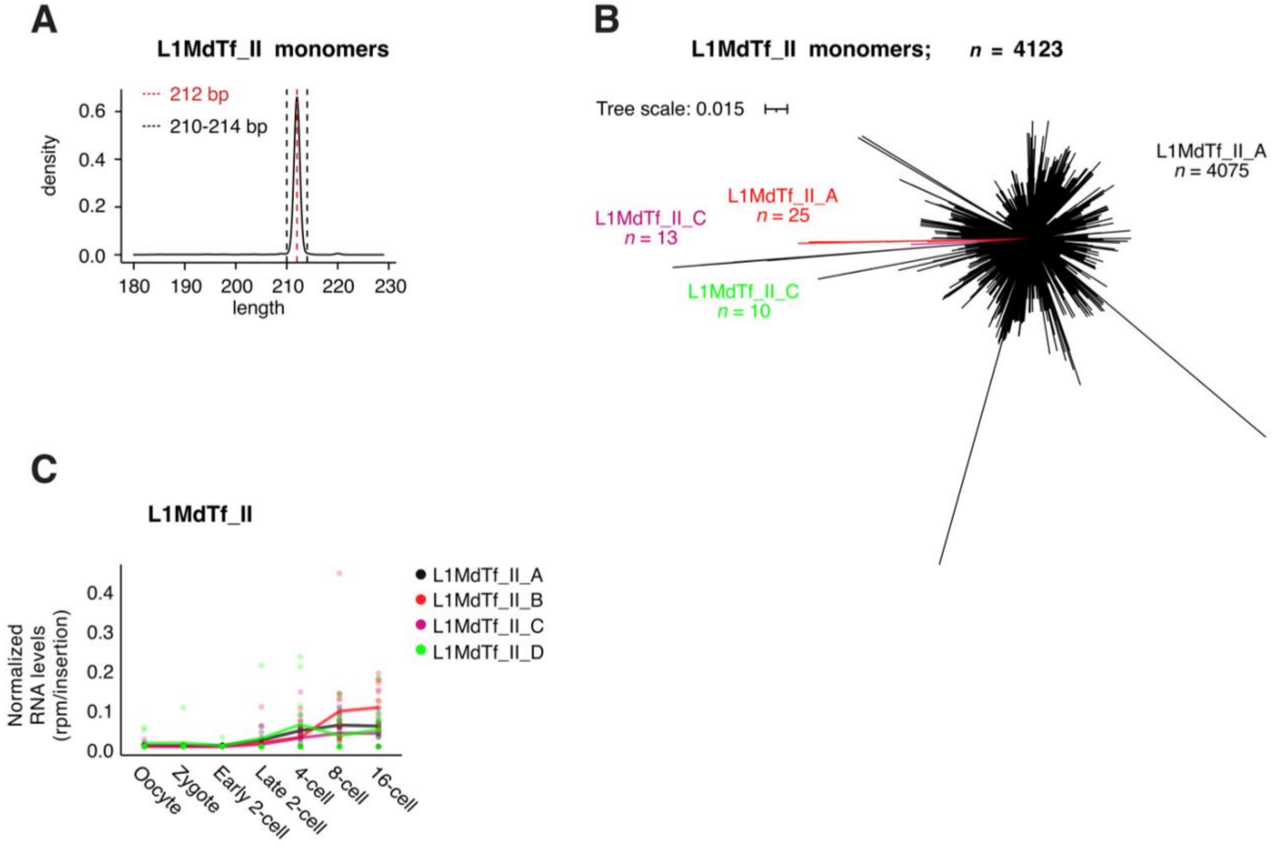
(E) ORR1A1



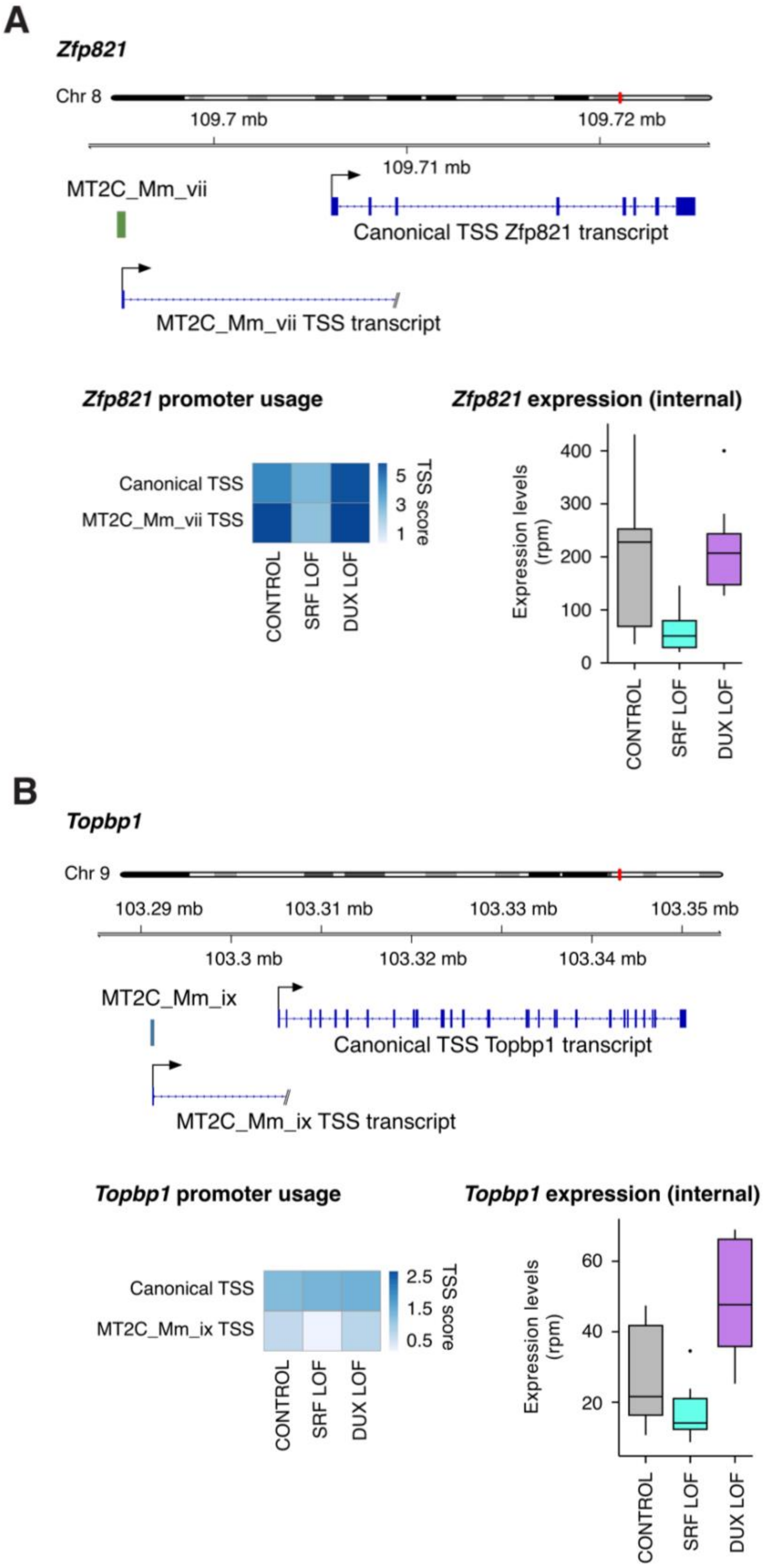
TomTom Output	Gene Name	Footprint stage	Fooprint number	Motif	Orientation
ehf	ehf	2C	1	Secondary	Reverse Complement
hbp1	hbp1	2C	1	Primary	normal
hmbox1	hmbox1	2C	2	Primary	normal
klf7	klf7	2C	2	Primary	normal
lhx8	lhx8	2C	2	Primary	normal
mrg1	meis2	2C	3	Primary	normal
mybl1	mybl1	2C	1	Secondary	Reverse Complement
pknox1	pknox1	2C	3	Primary	normal
rfxdc2	rfx7	2C	1	Primary	Reverse Complement
srf	srf	2C	2	Primary	normal
tcfe2a	tcf3	2C	3	Primary	Reverse Complement
tcf7	tcf7	2C	3	Primary	Reverse Complement
zfp410	zfp410	2C	1	Primary	normal
zfp740	zfp740	2C	2	Primary	normal
gm397	zscan4c	2C	2	Secondary	normal

(F) ORR1A0

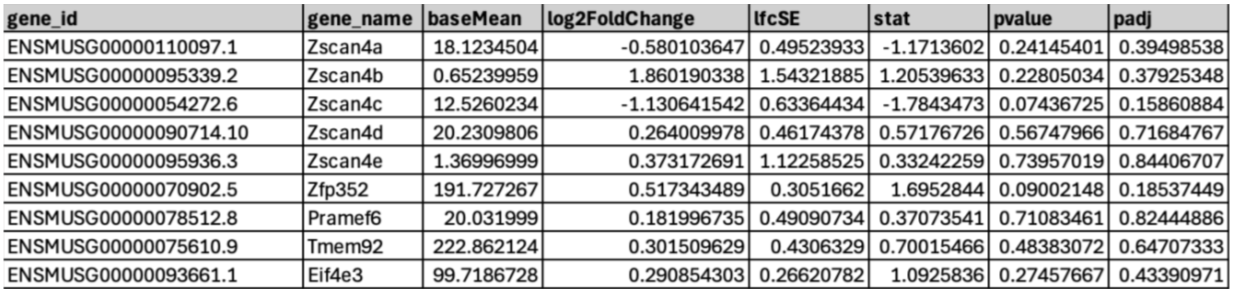
Appendix 2. Phylogenetic analysis of L1MdTf_II. (A) Size density distribution plot of L1MdTf_II monomers. The consensus length is displayed in a red dashed line. The selected size range is indicated in black dashed lines. (B) Unrooted phylogenetic tree of L1MdTf_II insertions ranging between 210 and 214bp in length. *n* is the number of sequences per group indicated. The tree scale is indicated. (C) Expression of all L1MdTf_II subfamilies across the different stages of preimplantation development. Data was reanalyzed from (Oomen et al. 2025). Each dot is a single embryo at the indicated stage. The trend line connects the mean values across embryos of individual stages.



Appendix 3. Additional examples of expression and promoter usage of chimeric-TE genes involving MT2C_Mm at the late 2-cell stage. (A,B) Snapshot of the genomic regions containing *Zfp821* (A) and *Topbp1* (B) canonical and MT2C_Mm-derived transcript. Heatmap displaying TSS score of both canonical and TE-derived TSSs in CONTROL, SRF LOF and DUX LOF embryos. The boxplot shows the expression of either *Zfp821* or *Topbp1*, as the median rpm and interquartile range of transcripts from internal sequencing reads. Whiskers display the highest and lowest value within 1.5 times the IQR.

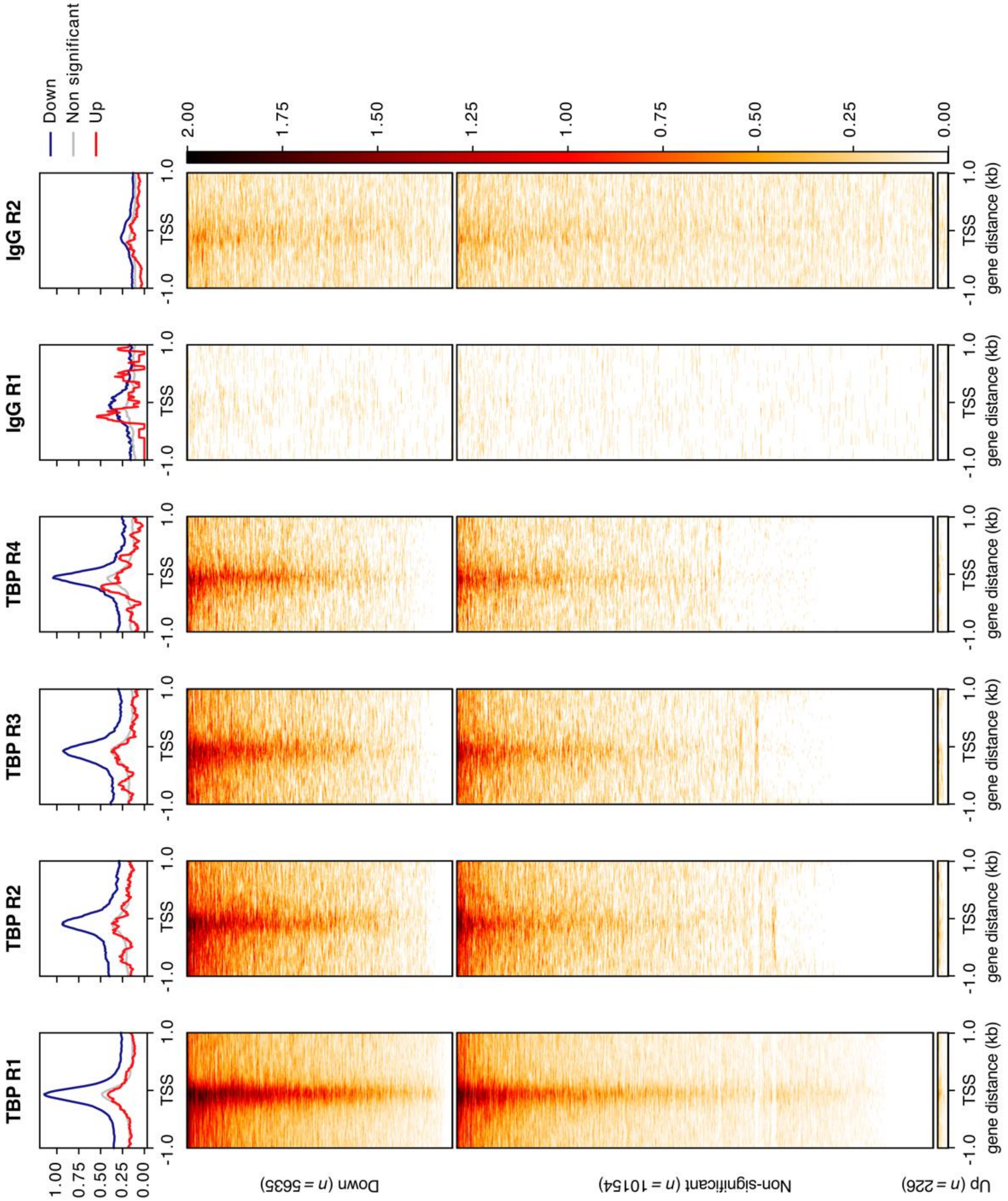


Appendix 4. Results of differential gene expression analysis of TBP LOF for notable Dux targets. Significance is given by padj column.



gene_id	gene_name	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSMUSG00000110097.1	Zscan4a	18.1234504	-0.580103647	0.49523933	-1.1713602	0.24145401	0.39498538
ENSMUSG00000095339.2	Zscan4b	0.65239959	1.860190338	1.54321885	1.20539633	0.22805034	0.37925348
ENSMUSG00000054272.6	Zscan4c	12.5260234	-1.130641542	0.63364434	-1.7843473	0.07436725	0.15860884
ENSMUSG00000090714.10	Zscan4d	20.2309806	0.264009978	0.46174378	0.57176726	0.56747966	0.71684767
ENSMUSG00000095936.3	Zscan4e	1.36996999	0.373172691	1.12258525	0.33242259	0.73957019	0.84406707
ENSMUSG00000070902.5	Zfp352	191.727267	0.517343489	0.3051662	1.6952844	0.09002148	0.18537449
ENSMUSG00000078512.8	Pramef6	20.031999	0.181996735	0.49090734	0.37073541	0.71083461	0.82444886
ENSMUSG00000075610.9	Tmem92	222.862124	0.301509629	0.4306329	0.70015466	0.48383072	0.64707333
ENSMUSG00000093661.1	Eif4e3	99.7186728	0.290854303	0.26620782	1.0925836	0.27457667	0.43390971

Appendix 5. Comparisons of the four TBP replicates on genes. Signal aggregate plots and heatmaps of TBP enrichment from the four individual CUT&Tag replicates and two IgG control over the TSS of down-regulated (padj < 0.05), non-significant and up-regulated (padj < 0.05) genes upon TBP LOF relative to CONTROL embryos. *n* is the number of genes per indicated category.



Appendix 6. List of significantly downregulated individual MT2_Mm insertions upon TBP LOF.

```
transcript_id "MT2_Mm_dup2391_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup1455_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1196_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2085_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup541_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2361_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup2180"; family_id
transcript_id "MT2_Mm_dup1714_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2130_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup86_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup2274"; family_id
transcript_id "MT2_Mm_dup701_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup376_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup11_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2612_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup318_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1688_MT2-Mm_iv"; family_id
transcript_id "MT2_Mm_dup463"; family_id
transcript_id "MT2_Mm_dup2360_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1221_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup517_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1816_MT2-Mm_iii"; family_id
transcript_id "MT2_Mm_dup1698_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup579_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup388_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1741_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1389_MT2-Mm_iii"; family_id
transcript_id "MT2_Mm_dup605_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup259_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1819_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup2179"; family_id
transcript_id "MT2_Mm_dup251_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2502_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1197_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup136_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2114_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1200_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup952_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1724_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2366_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup174"; family_id
transcript_id "MT2_Mm_dup218_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup395_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup1815"; family_id
transcript_id "MT2_Mm_dup2375_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1865_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2449_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1352_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup1719_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup173_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2469_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup268_MT2-Mm_iii"; family_id
transcript_id "MT2_Mm_dup1720_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup580_MT2-Mm_i"; family_id
transcript_id "MT2_Mm_dup2027_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup582_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup653_MT2-Mm_ii"; family_id
transcript_id "MT2_Mm_dup685_MT2-Mm_i"; family_id
```

10 Acknowledgements

I feel lucky and proud to have been part of the Torres-Padilla lab. I could hardly have been in a better place to do my PhD, thanks to all the people that I have been surrounded by, whom I wish to thank here. Above all, I want to thank Maria Elena. You have been an incredible supervisor, and I am deeply grateful to have had the opportunity to be part of your lab. I first joined the lab for my master's thesis and over the past six years your guidance influenced not only my scientific growth but also the person I have become. Thank you for believing in me, sometimes more than I believed in myself. You have given me incredible amount of support for which I am not sure I would be able to thank you enough.

I would like to thank also my TAC members, John Parsch and Cedric Feschotte, for their advice, constructive criticism throughout my PhD. My project has evolved and been shaped through discussions and interactions with you, thank you for that.

I want to thank Mich without whom an important part of this work would not have been possible. I am very proud of the work that we achieved together, and I hope our paths cross again in the future. I also want to thank Ana, the first and only student I have directly supervised, I really enjoyed working with you and know you are now a brilliant master's student.

Thomas and Laura, thank you for always helping me even when I come with silly admin questions because I don't understand anything, or to cope with the fact that I always paid for my coffee three months late. Pilar and Melissa, I feel like we did not overlap long enough to actually know each other, but thank you for the lunch breaks and the kitchen chats. Thank you Antje for always being so nice to me. Marga, thank you for doing the hard job that is taking care of the lab when I know it is a very demanding job. Andreas, thank you for your constant help when it comes to computers or microscopes, and for the chats at the bench when you are back to experiments. Tamas, thank you for all the help that you gave me with analysis and statistics, and all the help you gave Mich for the analysis of my data. A small nod to Jose, we officially did not overlap, and in practice only crossed paths for three weeks, which is a shame. I have a feeling we would have gotten along well. Yuki and Jiezhen, I thank you together because you are the new queens of the small lab, it was fun to see the small lab changing with you two arriving, with your energy and your contagious laughs! Adam, thank you for always willing to help, for the scientific discussions, and for maintaining the Torres-Padilla lab together. Tsune, I am glad that I did not have to experience the Torres-Padilla lab without you because it must feel so…weird! I learned a lot from you, I thank you for that, but also for being always full of surprises and stories! Yicong, thanks for your energy (so much energy!), thanks for your help with the phylogenetic analysis and the discussions about our common loved interest, prime TE club member! Iliya, thanks for being such a great member of the lab. Thanks for the jokes, thanks for ski trips, thanks for being my Torres-Padilla lab IRTG buddy, I feel like that has gotten us closer and it was always a lot of fun. Fede, thank you for all the fun and the laughter in the small lab. The IES sports day excitement, the singing, the motivational quotes... Thank you also for the deep conversations that we sometimes ended up having. I laughed to the point it was hurting my cheeks, but also cried in the small lab. I miss it already, to sit in the same lab as you. Mrinmoy, thank you so much for the past 6 years, we started together and finished almost exactly at the same time. Thank you for being always so helpful, and always there for me. Thanks also for all the cooking, for the beer outings, and all the fun we had outside of the lab. Antoine, I miss not having you coming to my desk every day at around 10am to either tell me about your last scientific reading, your last exciting experiment results, the fun movie you watched the day before or talk about how you got upset listening to the interview in the “matinale de France Inter”. I think we can now officially agree on calling each other friends and not colleagues anymore, and I know that wherever you end up you will be part of my life. Marlies, thank you for being my anchor and constant through the second half of my PhD. I always knew I could count on you. I genuinely don’t know how I would have navigated those years without you. I have always admired you, both as a scientist and as a person, and that admiration has only grown over time.

Looking further back, I would also like to acknowledge those who left the lab before me and still had meaningful impacts on my PhD journey. Juana, thank you for being simply so nice to everyone. Thanks also to Marie-Sophie and Petra for help and support. Ane, Ken, Hiromi, thank you for the scientific and non-scientific discussions in the lab. Ksenia, it was a short overlap but it was so nice to have you around. Luis, it was great to have you in the lab for two years. Thanks for helping me do the first “digging” in phylogenetic analyses, which started with an unreadable nine-sheet-long rectangular phylogenetic tree and ended with a beautiful colorful unrooted tree that I am really proud of. Amelie, thanks for joining my coffee breaks outside when you were at the IES, which lead us to become friends. I hope you and Maya will come visit me in France one day. Yung Li, we sat back-to-back for about 4 years and a half. Thanks for being the molecular cloning master and helping me so much with my molecular biology issues. You are a talented scientist, and I am sure you will do an amazing postdoc. Marion, thanks for becoming my friend in the lab, and remaining my friend after you left. I'll see you soon in Paris. Camille, also known as the queen of the lab, thank you for all the help you gave me when I joined. You were truly incredible and was making everyone's life easier. I am very happy that we stayed in touch after you left and that we are still part of each other's lives. Natasha, it feels like you left the lab a lifetime ago. In the end, we did not overlap for that long, but it somehow feels like we did. I feel connected to you even though you live about as far away as possible. Thank you for being the post-doc I looked up to when I first joined, and for becoming my friend along the way. Manuel, thank you, for basically everthing. For being such a close friend, both in and out of the lab. For the breaks at work, the beers, the skiing, and the countless evenings at Frida. Getting used to life in Munich without you was though, because we were always together.

My time in the lab was made all the more enjoyable thanks to the supportive environment at the IES. I'm especially grateful to Stephan, Antonio, Eva, Nico and, towards the end, Boyan. Thank you for contributing to making the positive atmosphere, for showing interest in my work, giving me suggestions, and helping me to grow scientifically. I also want to thank everyone in your labs for the engaging seminars, scientific discussions, jokes in the corridor, and all the moments that made the IES such a lively place, from the occasional parties to the IES sports day. I want to mention Elisabeth and Maxime, thank you for the laughter, the jokes, the dinners. I'm glad I got to know both you both. A little nod here to Marco and Kim, whom I met far too late, but I hope the Scialdone lab will enjoy playing with the little ball I left behind. Ana, thank you for all our lunch conversations and for the political reflections we shared, they meant a lot. Meghana, thank you for your ever-present positive energy, for all the singing and dancing, and for being such a beautiful soul. Elmir, thank you for becoming one of my closest friends. I have missed your contagious laugh around the institute. And Matthias, thank you for being part of my life. You have been around when things were hard, thank you for that.

I also want to thank Elizabeth, for the energy you put in the IRTG program and for the commitment to the well-being and career development of the IRTG students. I am grateful that I could be part of the program during my PhD. I have learned a lot and met many incredible people through the chromatin community.

Outside of the lab, many people made my time in Munich, and this PhD, full of memories. It all started with my first flatshare in Veilchenstr., Goku, Shriya, Daniel and Haya. We only lived

together for six months, yet six years later, you are still among the most important people in my life in Munich. A special mention to Goku: we ended up being the longest together in Munich, and later with Surudhi. I always knew I could come over when I was not okay, and that meant everything during harder times. Thank you for always picking up the phone and being there. To the extended Veilchenstr. family, Joop and Maja, thank you for all the lovely times together.
A big thank you to the "Frenchies". Carole, I would not have had a flat, a phone, skis, a tennis racket, or half of the things I needed without you. Covid times would have been so much harder without your presence. And of course, thank you to the rest of the "French Republic of Maxvorstadt", Baptiste and Guillaume (and later Vlada), for all the time we spent together in that flat.
Fast forward a few years, Julia and Tom, the last flatmates I had in Munich. You were the nicest, funniest, sweetest flatmates one could dream of. One of the hardest things about leaving Munich was leaving that flat, and not living with you anymore. But I know you will stay in my life, wherever I go.
It may have only been the last year of my stay it Munich, but I feel like it ended up being such an important part, thank you to TTT. It all starts with Pavi, the group's guru and cult initiator, and now a very dear friend. The group has grown so much I could not name everyone, but I have to mention Anandi, who seems to be following me everywhere and honestly, it is always fun to have you around, just like the good old days. Thank you Moe, Omar, Brandon, Paul, Taher and everyone else who was part of that group. And of course, Janet. You hold a special place in my heart. Thank you for being such a big part of this final year of PhD.

Finally, to the people outside of Munich, thank you to my friends in France, in Belgium or in Germany, whom I met from school, scouts, university, erasmus, all of you have been present and important during my PhD. Thank you also to all my family in Belgium who was always asking me how were "my little cells" doing when I'd come back for Christmas.

Fanny et Paola, même si la distance et le temps ont parfois pu nous éloigner, on sait qu'on est toujours là les unes pour les autres. Qu'importe les temps de silence, le lien est fort. Merci d'être vous, et merci de contribuer à façonner celle que je suis.

Maman, Papa, sans vous, je ne serais pas là, dans tous les sens du terme. Merci pour votre amour, votre soutien et votre présence inconditionelle, tout au long de ces six années, et bien avant cela.