

The role of MERVLs as alternative promoters in ZGA

Srishti Arya

Supervisor: Professor Juan M. Vaquerizas

Developmental Epigenomics Group

Medical Research Council Laboratory of Medical Sciences

Institute of Clinical Sciences

Thesis presented for the degree of Master of Philosophy

Imperial College London

June 2024

Statement of originality

I declare that this MPhil thesis has been written by me and all the work is my own unless otherwise stated. This work has not been submitted for any other degree.

Srishti Arya

June 2024

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non-Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

The most impactful realisation I had while completing this project is:

“It takes a village.”

I’m so grateful to my village, and I feel so lucky to have lived in it.

First, I want to thank my supervisor, Juanma. For obvious reasons, this project wouldn’t have happened without you, but even so, I’m still so incredibly grateful for your time, enthusiasm, and excellent ideas during the past couple of years. Liz, you thought about my work with me so consistently, even when you had so much else to focus on. Maria, your knowledge, logic, and kindness were invaluable to me. I’m indebted to you all for your mentorship and boundless patience. I’m also very grateful to my assessors, Michelle and Boris. Thank you for the thorough discussions about the project.

The Developmental Epigenomics lab has a wonderful, collegial, and kind environment. To my lab mates and past lab mates, I’m so thankful for your guidance, support, and advice, and for how lovely you are as people. Noura, Fabian, Sara, Jahnavi, Irina, Mel, Virginie, Quirze, Ben, and Christos, you’ve been there for me in amazing ways. You’ve become dear friends. Also, thanks for putting up with my terrible jokes. I think you’re all pretty ***MERVL-ous***.

I’m deeply appreciative of the support and warmth of the LMS community, and the PhD committee in particular. Julia, Bonnie, Lucas, Ferran, Darren, Paul, Bryony, and *especially Ilinca and Korak*, you guys have been my rock. I’m so glad to know you, and I’m so excited to see what you all do next!

There are only so many ways to say the phrase ‘thank you’, and none of them cover it to explain how much my friend deserve thanks. To my friends (and let’s be

real, my family) from KCL and UCL, Alex, Alice, Ash, Lily, Arturo, and Zoe, your love and care means the world to me. You've shaped who I am. And I'm stronger because of you. Finally, to Flat 17 too. You have become such close friends. Thanks a million, Riley, Torben, Melker, and Emma!

मेरे मम्मा, पापा, सृजन, प्रकृति, और चाक के लिए | आप के प्यार कि वजह से यह थीसिस हो पाई | जब मुश्किलें बढीं, आप ने हमेशा मेरा साथ दिया और मुश्किलें पार करवा दीं | मैं आप सब से बहुत प्यार करती हूँ और दिल से आप को धन्यवाद कहना चाहती हूँ |

Abstract

Careful tuning of transcription and translation during embryogenesis, specifically at the initial awakening of the embryonic genome (zygotic genome activation, or ZGA) is crucial for correct embryogenesis. Mouse ZGA, at the 2-cell stage, is hallmarked by the de-repression of murine endogenous retroviruses with leucine primers (MERVLs). Co-option of the MERVLs' LTRs as alternative promoters has been observed for embryogenesis-related genes. However, no transcriptome-wide studies have been completed specifically for ZGA genes with MERVL alternative promoters yet.

Here, I present an analysis of published long-read RNA-seq data from 2-cell embryos to examine the repertoire of chimeric transcripts promoted by MERVLs, supplemented by short-read RNA-seq data from 2C embryos and 2-cell-like cells (2CLCs). 2CLCs provide an effective *in vitro* system to study early mouse embryogenesis, due to their emulation of 2-cell stage embryos' epigenetic, metabolic, and transcriptomic state, with characteristic MERVL expression. I collate a list of 549 high-confidence transcripts promoted by MERVLs, from 167 genes, and assess their expression across preimplantation development. Subsequently, I examine the coding potential of these transcripts, identifying 15 genes with 2C specific transcripts. Of these, 6 genes have MERVL-promoted isoforms with altered protein domains, and therefore, possibly altered functions. This study opens doors to examination of changes in protein domains or signal peptides for MERVL-promoted transcripts, as well as further examination of tertiary protein structure and binding partners using computational and experimental techniques.

Table of Contents

<i>Statement of originality</i>	<i>2</i>
<i>Copyright declaration</i>	<i>2</i>
<i>Acknowledgements.....</i>	<i>3</i>
<i>Abstract.....</i>	<i>5</i>
<i>List of Figures and Tables.....</i>	<i>10</i>
<i>1. Introduction.....</i>	<i>16</i>
1.1 Preimplantation development	16
1.1.1 Oogenesis and Fertilisation	16
1.1.2 Zygotic Genome Activation	17
1.1.3 Further preimplantation development.....	18
1.1.4 Key Players in ZGA	21
1.1.5 Two-cell like cells	21
1.2 Transposons in early embryogenesis.....	27
1.2.1 Transposable element phylogeny.....	27
1.2.2 Transposon profiles in preimplantation development	32
1.2.3 Co-option.....	35
1.3 MERVL elements and transcriptional control	39
1.3.1 Murine Endogenous retroviruses - L.....	39
1.3.2 MERVL LTR promoters	40
1.3.3 MERVLs in preimplantation development	43
1.4 Sequencing technologies	45
1.4.1 Short- and Long-read sequencing.....	45
1.4.2 CAGE-seq	46
1.5 Aims	50

Aim 1. To define and characterise a repertoire of high-confidence MERVL-promoted alternative transcripts isoforms in 2CLCs.....	50
Aim 2. To understand the functional impact of selected MERVL-promoted alternative protein isoforms in 2CLC function.	51
2. Materials and Methods	52
Code availability.	52
Data availability.	52
MT2-Mm MERVL locus definition.....	55
CAGE-seq processing	55
Short-read RNA-seq processing	56
Long-read RNA-seq processing.....	56
Isoform assembly and definition of MERVL-promoted isoforms.	57
Expression analysis of novel isoforms.....	58
Gene ontology analysis.....	58
Signal visualisation at MERVL loci	59
Transcript isoform visualisation.....	59
Coding potential examination, ORF prediction, and protein domain homology search	60
Chapter 3: Results	61
3.1 MERVL elements are independently expressed at the 2-cell stage embryo	61
3.1.1 Curation of a list of MERVL elements	62
3.1.2 MERVLs show transcriptional signals in vivo and in vitro at the 2C stage.....	68
3.1.3 CAGE-seq data shows MERVL expression initiates in LTRs	72
3.1.4 Transcriptional signal initiating in LTRs is biased to the sense strand.....	78
3.2 Examining the impact of MERVL promoters on the 2C transcriptome. ...	82
3.2.1 Identifying MERVL-promoted transcripts.....	83
3.2.2 Constructing a list of high-confidence ‘parent’ genes.	90
3.2.3 Comparison of parent genes retrieved in this study to prior works.	91
3.2.4 Characterising MERVL-promoted transcripts in the context of the 2C transcriptome.	94

3.3 Cellular dynamics of MERVL-promoted isoforms.....	100
3.3.1 Expression patterns of MERVL-promoted transcripts using long-read data.	101
3.3.2. MERVL-promoted transcripts are statistically less likely to be coding compared to reference transcripts.	118
3.3.3. Protein domain homology in MERVL-promoted transcripts.	127
4. Discussion and future perspectives.	146
4.1 Contextualising results.	147
4.1.1 Defining MERVL elements.....	147
4.1.2 Transcriptional signal at MERVL loci.	148
4.1.3 Transcriptional initiation in MERVL LTRs.....	150
4.1.4 Definition of a high-confidence list of ZGA-specific MERVL-promoted isoforms.	152
4.1.5 Expression and translational capacity of transcripts.....	153
4.1.6 Specific cases	154
4.2 Technical considerations.....	156
4.2.1 Transcript Structure Resolution.	157
4.2.2 Batch effects.	158
4.2.3 CAGE-seq read length and mappability.	160
4.3 Future perspectives.	160
4.3.1 How robust are genes to alternative promoters?	160
4.3.2 What is the impact of MERVL-promoted protein isoforms?	162
4.3.3 How does the MERVL LTR promoter function at the molecular level?	165
4.3.4 A note on MERVL LTRs as enhancers.....	167
5. References	168
<i>Appendix 1: RNA-seq QC</i>	<i>191</i>
<i>Appendix 2: Troubleshooting parent genes</i>	<i>194</i>
<i>Appendix 3: Transcript isoform structures.....</i>	<i>198</i>

Appendix 4: ‘best’ ORF per gene (MERVL-promoted versus reference)

.....**222**

List of Figures and Tables

- Figure 1.1: Preimplantation development.
- Figure 1.2: 2CLC induction. Source of mES cell cultures and proportion of endogenous 2CLCs with different induction techniques.
- Figure 1.3: Proportion of different classes and families of TEs in the mm10 genome assembly
- Figure 1.4: Phylogeny of TE classes, families, and subfamilies, highlighting the categorisations of MERVL elements.
- Figure 1.5: Profiles of TE expression across murine embryogenesis.
- Figure 1.6: Structure of MERVL elements and regulatory elements within the LTR.
- Figure 1.7: Simplified workflow (library preparation, sequencing, and read alignment) for sequencing technologies used in this study.
- Figure 1.8: Simplified workflow for library prep, sequencing, and read alignment using SLIC-CAGE.
- Figure 3.1: IGV tracks of MERVL elements conserved between C57Bl6J and 129P2/Ola genomes.
- Figure 3.2: IGV tracks of MERVL elements conserved between C57Bl6J and 129P2/Ola genomes.
- Figure 3.3. Distribution of lengths of MERVL elements conserved between C57Bl6J/129P2/Ola strains.
- Figure 3.4: MERVL expression in 2CLCs, mESCs, and 2C embryos using short-read data.
- Figure 3.5: MERVL expression in 2C embryos using long-read data.

- Figure 3.6. Transcription initiates in MERVL LTRs across late oocyte maturation and early embryogenesis.
- Figure 3.7 Transcription initiates in MERVL LTRs across late oocyte maturation and early embryogenesis.
- Figure 3.8: Strand-based examination of transcription initiation in full MERVL LTRs across late oocyte maturation and early embryogenesis.
- Figure 3.9: Strand-based examination of transcription initiation in solo MERVL LTRs across late oocyte maturation and early embryogenesis.
- Figure 3.10. Schematic outlining definition of MERVL-promoted transcripts.
- Figure 3.11. Breakdown of transcript type assembled per dataset.
- Figure 3.12. Number genes (parent genes) with MERVL-promoted transcripts per dataset unique to each dataset and shared between datasets.
- Figure 3.13: Comparison of high-confidence parent genes in this study to other similar works.
- Figure 3.14: Working model of MERVL-promotion of transcripts.
- Figure 3.15: MERVL elements can promote transcription from introns, for confirmed and unconfirmed genes.
- Figure 3.16: Reference transcripts outnumber MERVL-promoted transcripts.
- Figure 3.17: Exon lengths are shorter than intron lengths.
- Figure 3.18: PCA plot for SCAN-seq (Fan et al., 2020) long-read RNA-seq dataset.
- Figure 3.19: Expression analysis of MERVL-promoted transcripts across preimplantation development.
- Figure 3.20. No MERVL-promoted transcripts are shared between clusters.
- Figure 3.21: Few parent genes are shared between clusters.

- Figure 3.22: Comparison of MERVL-promoted transcript numbers from parent genes shared between the oocyte-specific cluster and the 2C specific cluster.
- Figure 3.23: Summarised expression analysis of parent genes of MERVL-promoted transcripts across preimplantation development.
- Figure 3.24: Expression analysis of transcripts assembled in this study belonging to parent genes of MERVL-promoted transcripts defined by Modzelewski et al., (2021).
- Figure 3.25: Expression patterns across preimplantation development of 2C-specific MERVL-promoted transcripts in context of their parent genes, for known genes.
- Figure 3.26: Figure 3.26: Expression patterns across preimplantation development of 2C-specific MERVL-promoted transcripts in context of their parent genes, for unconfirmed genes.
- Figure 3.27: Counts of coding and non-coding ORFs derived from all MERVL-promoted and reference transcripts.
- Figure 3.28: Examination coding and non-coding ORFs derived from MERVL-promoted and reference transcripts.
- Figure 3.29: Venn diagram comparing whether a parent gene has coding or non-coding ORFs.
- Figure 3.30: Comparison of coding potential, mRNA transcript length, and ORF length.
- Figure 3.31: Expression patterns across preimplantation development of 2C-specific MERVL-promoted transcripts in context of their parent genes, for genes with coding transcripts.

- 3.32: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Abcb5*.
- 3.33: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Csrnp3*.
- 3.34: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Fundc1*.
- 3.35: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Gnpnat1*.
- 3.36: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Sgk3*.
- 3.37: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Zfp560*.
- Appendix 1, Figure 1: Quality control for short-read RNA-seq data.
- Appendix 1, Figure 2: Quality control for long-read RNA-seq data. Both datasets are single-cell modifications to cDNA-PCR Oxford Nanopore sequencing.
- Appendix 2, Figure 1: schematic defining how parent genes are selected.
- Appendix 2, Figure 2: IGV view of example gene *Apol7b*, where one dataset has a MERVL-promoted transcript (Berrens et al., (2021), CELLO-seq) and

one does not (Fan et al., (2020), SCAN-seq), highlighted with a light blue rectangle.

- Appendix 2, Figure 3: IGV view of *Igk* gene locus, removed from analysis due to complexity.
- Appendix 3, Figure 1: Sashimi plot of *Ddit4l* locus (Chr3:137325270-137336094).
- Appendix 3, Figure 2: Sashimi plot of *Rplpo* locus (Chr5:115694567-115702078).
- Appendix 3, Figure 3: Sashimi plot of *Ppig* locus (Chr2:69551152-69586356).
- Appendix 3, Figure 4: Sashimi plot of *Pemt* locus (Chr11:59859446-59939394).
- Appendix 3, Figure 5: Sashimi plot of *Abcb5* locus (Chr12:118819122-118932169).
- Appendix 3, Figure 6: Sashimi plot of *Csrnp3* locus (Chr2:65673628-65907061).
- Appendix 3, Figure 7: Sashimi plot of *Fundc1* locus (ChrX:17383598-17460387).
- Appendix 3, Figure 8: Sashimi plot of *Gnpnat1* locus (Chr14:45610944-45643676).
- Appendix 3, Figure 9: Sashimi plot of *P4ha2* locus (Chr11:53987410-54024752).
- Appendix 3, Figure 10: Sashimi plot of *Plaat3* locus (Chr19:7526655-7567910).
- Appendix 3, Figure 11: Sashimi plot of *Sgk3* locus (Chr1:9866355-9975296).

- Appendix 3, Figure 12: Sashimi plot of *Zfp560* locus (Chr9:20255936-20296964).
- Table 1. Positive and negative regulators of 2CLC induction.
- Table 2. Examples of transposon co-option.
- Table 3: datasets analysed in this study.
- Table 4: breakdown of GFFcompare class codes assigned to all assembled transcripts relative to mm39 NCBI RefSeq genome assembly.
- Table 5: Net changes in protein domains between the reference ORF and MERVL-promoted ORF with the highest coding potentials per 2C-peaking gene.
- Appendix 4, table: 'best' ORF per gene (MERVL-promoted versus reference)

1. Introduction

1.1 Preimplantation development

1.1.1 Oogenesis and Fertilisation

Mammalian embryonic development initiates at the fertilisation of a haploid oocyte by a haploid spermatozoon. The process of fertilisation requires two highly specialised cell types to fuse and combine their genetic material; thus, they must undergo intricate developmental processes before fertilisation and embryogenesis can even begin. Both spermatozoa and oocytes arise from primordial germ cells, specifying into spermatogonia or oogonia prior to the entry to meiosis (summarised by Margolin et al., 2014, Schultz et al., 2018, and Subrini and Turner, 2021).

It has been argued that embryogenesis begins pre-fertilisation, during oogenesis. Oogenesis occurs in the follicles of the adult ovary; in mice, this can happen as early as 2.5 weeks postnatally (Schultz et al., 2018). After entry to meiosis, immature oocytes undergo waves of transcriptional and translational activation and inactivation with the phases of meiosis (Schultz et al., 1979). The first transcriptional inactivation occurs in prophase I, followed by a re-activation after a change in nucleolar structure (from a non-surrounded nucleolus to a surrounded nucleolus) (Bonnet-Garnier et al., 2012; Eleftheriou et al., 2022). These changes are required to allow meiotic competence of the oocyte (the ability to progress through meiosis). Finally, the oocyte arrests at metaphase II, producing a mature oocyte capable of fertilisation and subsequent embryonic development (developmentally competent) (Schultz et al., 2018).

The mature MII oocyte contains the complete repertoire of mRNA and protein required for zygotic development until the activation of the zygotic genome (zygotic genome activation, or ZGA, sometimes referred to as embryonic genome activation).

1.1.2 Zygotic Genome Activation

At fertilisation, both the oocyte and sperm are transcriptionally dormant, resulting in an equally transcriptionally dormant 1-cell zygote. To kick-start development post-fertilisation, mRNAs and proteins pre-deposited in the mature oocyte begin driving the initial phases of the first cell cycle before the newly formed zygotic genome is functionally activated (reviewed in Svoboda, 2018; Svoboda et al., 2015; Vastenhouw et al., 2019). ZGA onset is concurrent with the degradation of these maternally inherited transcripts and proteins (Figure 1.1), and is characterised by several hallmarks. 92% of DNA methylation marks are removed from the genome, mainly through active DNA demethylation mechanisms from the paternally inherited genome and mostly through passive DNA demethylation mechanisms from the maternally inherited genome. After pronuclear syngamy (the fusing of the maternal and paternal pronuclei at the 1-cell stage), post-translational modifications to histones are rewritten *en masse*.

Interrelated with these is the opening of heterochromatin regions, broadly in minor ZGA and more specifically in major ZGA, and the emergence of topologically associated domain (TAD) structures. While the causal relationship between 3D genome structure and transcription is contested, one hypothesis suggests that the precise regulation of transcriptional activation is a result of the combination of these epigenetic hallmarks (reviewed by Eckersley-Maslin et al., 2018 and Ing-Simmons et al., 2022).

This project focuses on ZGA, in mice. This occurs in 2 bursts, the minor and major waves of ZGA. The minor wave occurs in time between the S phase of the 1-cell stage and the G1 phase of the 2-cell stage. During minor ZGA, promiscuous transcription resulting in minimally spliced and incompletely processed transcripts has been reported (Abe et al., 2015; Asami et al., 2023; Bouniol et al., 1995; Perry et al., 2023), and inhibition of RNA Polymerase II at this time results in a failure to continue past the 1-cell stage (Abe et al., 2018). The mid-to-late 2-cell stage is the timing for the major ZGA wave. At this time point, not only is the transcriptional burst much more targeted to specific ZGA genes, but the resulting transcripts are processed more completely, with a larger repertoire of alternatively spliced transcripts making up the transcriptome, with modulation of mRNA poly-adenylation tail length (Abe et al., 2022; Lee et al., 2024; Svoboda et al., 2015; Tian et al., 2020; Wyatt et al., 2022). A shift is also noted in the promoter structure or exact position at ZGA; while maternal promoters active in the development of the oocyte tend to be narrow, i.e., have sharper, more closely clustered transcriptional start sites (TSS) related to DNA sequence motifs, once ZGA is initiated, the promoter choice in the genome becomes broader with more spaced-out TSSs regulated by nucleosome positioning (Cvetesic et al., 2020a; Haberle et al., 2014). Exit from the 2C stage requires maturation of the nucleolus and consistent transcription of rRNA via RNA Polymerase I (Xie et al., 2022). While several definitions for totipotency exist (summarised by Baker and Pera in 2018), mouse blastomeres of the 2-cell stage are the latest cell stage considered to be totipotent.

1.1.3 Further preimplantation development

The process of ZGA in mice ends at the second cell division. Blastomeres of the 4-cell stage are no longer considered totipotent, and exhibit differential gene

expression, protein content, and histone modifications, in order to establish an embryonic axis (Goolam et al., 2016; Rosa et al., 2024; Torres-Padilla et al., 2007). After the 3rd cell division, the blastomeres of the 8-cell stage undergo compaction, whereby the cells adhere more strongly, and the embryo flattens slightly. This is attributable to adhesion proteins and is necessary for correct cell adhesion, cell-cell communication, and axis specification (Cockburn and Rossant, 2010; Fierro-González et al., 2013). Post-compaction, the embryo forms a morula, and small cavities begin to form at the 32-cell stage, and eventually will fuse to create the fluid-filled blastocoel (Fujimori, 2010). At this stage, the outermost cells have committed to forming the trophectoderm, and cells collect at the base of the embryo, forming the inner cell mass (Dyce et al., 1987). The trophectoderm is the source of extraembryonic tissues, while the inner cell mass will further specify into the epiblast and primitive endoderm lineages (summarised by Cockburn and Rossant, 2010). The late blastocyst then will implant into the uterine wall, and organogenesis will follow (summarised by Kojima et al., 2014).

While the transcriptome varies widely during preimplantation development in mice, the proteome remains far more stable and does not mimic such fluctuating patterns (Zhang et al., 2023).

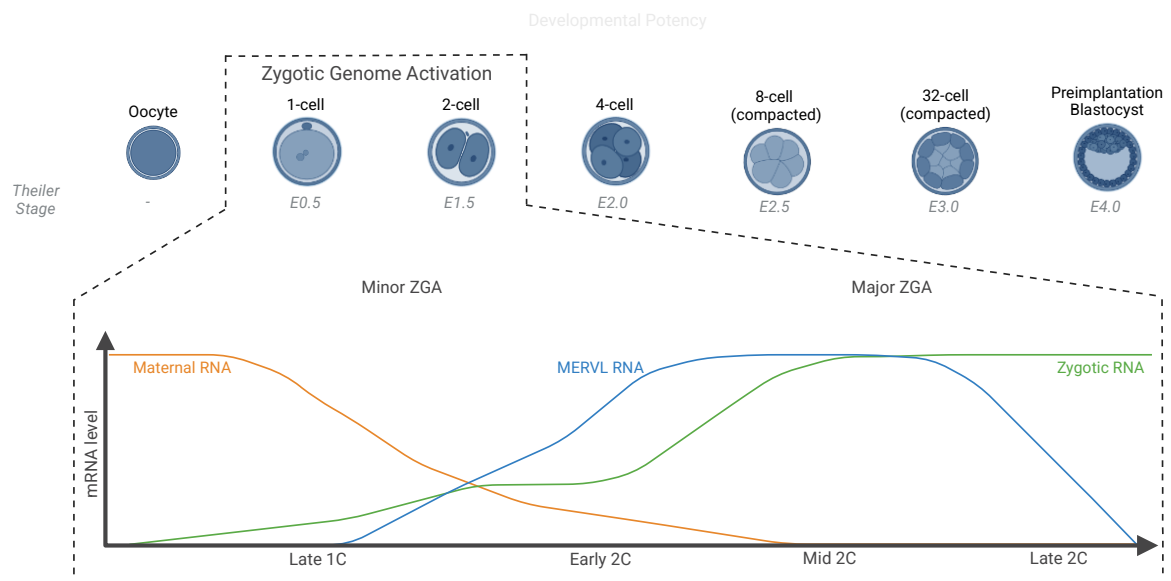


Figure 1.1: Preimplantation development. Cell stages across murine embryogenesis, changes in stem cell potency, and focusing on key transcriptional changes in minor and major ZGA

1.1.4 Key Players in ZGA

Numerous transcription factors, receptors, and metabolites are implicated in the initiation of ZGA and the exit to the 4C stage, to different extents. These include the orphan nuclear receptor *Nr5a2* (Gassler et al., 2022), *Zscan4* (Srinivasan et al., 2020; Zhang et al., 2019), the homologs *DPPA2* and *DPPA4* (De Iaco et al., 2019), *Stella* (Huang et al., 2017), and *Yap1* (Yu et al., 2016). Recently, the function of *Obox* (oocyte-specific homeobox) factors has been confirmed in ZGA onset, potentially as a ZGA initiation factor (Ji et al., 2023). The double homeobox transcription factor *Dux* is a key pioneer factor during ZGA, one of the earliest active in ZGA (the mouse homolog for the human DUX4 with the same function, (De Iaco et al., 2017; Vuoristo et al., 2022)) though its depletion is not embryonic lethal (Hendrickson et al., 2017; De Iaco et al., 2020).

1.1.5 Two-cell like cells

Given that in mice major ZGA occurs when the embryo is comprised of only 2 cells, the study of its specific mechanisms remains difficult. Therefore, *in vitro* options for studying mechanisms active at the 2-cell stage in mice remain important. In 2012, Macfarlan et al., discovered that 0.5-1% of cultured mouse embryonic stem cells (mESCs) spontaneously reverted to a state transcriptionally emulating 2-cell embryos, termed 2-cell-like cells (2CLCs, Figure 1.2). Since this discovery, a multitude of molecular hallmarks such as differentially expressed transcription factors, histone modifiers, metabolites, and RNA binding proteins have been documented as differentially functional in 2CLCs as compared to mESCs (summarised by Fu et al., 2020; Iturbide and Torres-Padilla, 2020; Rodriguez-Terrones et al., 2017; Torres-Padilla, 2020; Zhou and Dean, 2015). Manipulation of these hallmarks (e.g., *Dux* overexpression using a doxycycline-inducible system, or knockdown of certain

subunits of *Caf-1*) allows for induction of 2CLCs in an mESC culture, increasing yield of 2CLCs from 0.5-1% to up to 70%, in the case of *Dux* overexpression (Hendrickson et al., 2017; De Iaco et al., 2020; Ishiuchi et al., 2015). Finally, these hallmarks have been used to great effect in understanding key features of 2-cell embryos themselves, using conversion rates to 2CLCs from mESCs as a metric (Table 1). Recently, a study examined all 2CLC induction methods to date and their similarity in mRNA and protein content to 2C embryos and endogenous 2CLCs (Rosa et al., 2024). In this study, I employ 2CLCs induced by *Dux* overexpression and the knockdown of the p150 and p60 subunits of the *Caf-1* histone chaperone (Hendrickson et al., 2017; Ishiuchi et al., 2015).

Factor/element	Type	Change	Reference
Dux	Transcription Factor	+	(Hendrickson et al., 2017; De Iaco et al., 2017)
Caf-1	Histone chaperone	-	(Ishiuchi et al., 2015)
MERVL	Transposon	+	(Macfarlan et al., 2012; Yang et al., 2020)
Zscan4	Transcription Factor	+	(Eckersley-Maslin et al., 2016; Rodriguez-Terrones et al., 2018)
Retinoic acid	Metabolite	+	(Iturbide et al., 2021; Wang et al., 2021a)
CTCF	Transcription Factor	-	(Olbrich et al., 2021)
DPPA2/4	Transcription Factor	+	(Eckersley-Maslin et al., 2019; De Iaco et al., 2019)
NELFA	Transcription Elongation Factor	+	(Hu et al., 2020; Wang et al., 2021a)
G9a/GLP	Histone chaperone	+	(Maksakova et al., 2013)
miR-344	microRNA	+	(Yang et al., 2020)
Obox4	Transcription Factor	+	(Guo et al., 2022)
Foxd3	Transcription Factor	+	(Puri et al., 2021)
Gata2	Transcription Factor	+	(Choi et al., 2017)

REX1	Transcriptional repressor	-	(Schoorlemmer et al., 2014)
Nucleolin	Ribosomal DNA/RNA binding protein	-	(Percharde et al., 2018)
Table 1. Positive and negative regulators of 2CLC induction			

Induction of a splicing ‘failure’ has been shown to cause cells to de-differentiate to mimic earlier embryonic states (Wyatt et al., 2022). In the same vein, since the discovery of 2CLCs, several other in vitro alternatives for emulating cells undergoing zygotic genome activation have been developed, such as Totipotent blastomere-like cells (TBLCs) which are induced by inhibiting the spliceosome with pladienolide B (Shen et al., 2021).

While 2CLCs have been cemented as an effective and experimentally efficient model for the 2C stage of early embryogenesis, there remain caveats in their usage for studying ZGA, primarily because one single 2CLC cannot give rise to an entire embryo, therefore, is not considered totipotent by that definition. Additionally, because they derive from transcriptionally active cells (mESCs), and are not ‘switching on’ their genome for the first time. As a more minor point, they cannot mirror anterior-posterior patterning since they exist as single cells within a culture, and they are not in the same extracellular environment as 2-cell embryos (Denker, 2016; Torres-Padilla, 2020).

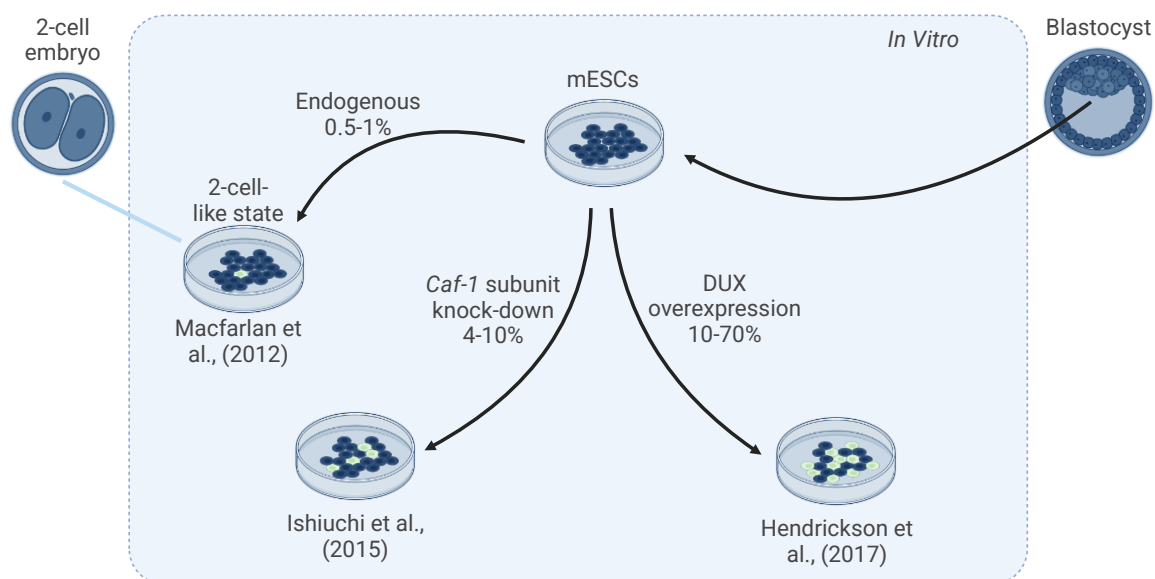


Figure 1.2: 2CLC induction. Source of mES cell cultures and proportion of endogenous 2CLCs with different induction techniques. Adapted from Eckersley-Maslin et al. (2018) and Fu et al. (2019)

1.2 Transposons in early embryogenesis

1.2.1 Transposable element phylogeny

As aforementioned, only 1.5% of the genome comprises protein-coding genes. However, repetitive elements make up roughly 45% of the genome, first identified by Barbara McClintock in maize, in 1950 (Figure 1.3). The vast majority of these repetitive elements are transposable elements, or TEs, (Bourque et al., 2018). Originally, a phylogeny of TEs was derived through the study of their mobilisation mechanism only, but as more and more TEs have been discovered, TEs are classified by their sequence similarities and abundances (Arkhipova, 2017, 2018). TEs can be first classified into two classes; class 1 TEs (retrotransposons), which transpose using an RNA intermediate and subsequent reverse transcription, and class 2 TEs (DNA transposons), which excise out of the genome before re-integration (summarised by Feschotte and Pritham, 2007). Approximately 1% of the genome encodes for DNA transposons since their cut-and-paste replication mechanism results in no net change in copy number post-transposition. The majority of the TE-derived genome encompasses retrotransposons (roughly the remaining 44% of the entire genome) (Figure 1.4).

The re-integration mechanism of retrotransposons can divide this class into further subclasses, depending on the presence of long terminal repeats (LTRs) flanking the TE, therefore grouping retrotransposons into LTR retroelements or non-LTR retroelements (Wicker et al., 2007). In this study, I focus on LTR retroelements only. Transcription of the LTR retroelements begins in the 5' LTR and terminates in the 3'LTR, after which the resulting mRNA is transported to the cytoplasm and reverse transcribed, before import back into the nucleus for re-integration into the genome. LTR retroelements contain a proviral sequence in between the LTRs; these provide the

machinery required for replication and reintegration. This proviral sequence has evolutionary roots in exogenous retroviruses (Benachenhou et al., 2013; Eickbush and Jamburuthugoda, 2008; Lefkowitz et al., 2018; Wicker et al., 2007) and encodes for a polycistronic mRNA of several proviral proteins which are post-translationally cleaved (Wells and Feschotte, 2020). The exact proteins per ERV vary according to subfamily; for example, the *env* protein sequence is frequently either lost or mutated, and this appears to correlate with their expansion probability (Magiorkinis et al., 2012). Some LTR retroelements are theorised to be a chimera between domains of certain DNA transposons and non-LTR retroelements (Malik and Eickbush, 2001), whereas the subset of LTR retroelements called Endogenous retroviral elements (ERVs) are known to be a footprint of ancient viral integrations, endogenized through inheritance from parent to offspring (Lefkowitz et al., 2018; Medstrand et al., 2002). Moreover, LTR retroelements may also undergo homologous recombination of their two LTR arms, resulting in a solo LTR remaining and the removal of the proviral sequence for this specific insertion (Coffin, 1979). Finally, TEs are then classified into families according to sequence similarity, following an 80-80-80 rule: two elements are of the same family if they are both at least 80bp in length and share 80% sequence homology over 80% of their aligned sequence in either or both of their internal domains OR terminal repeats (Wicker et al., 2007).

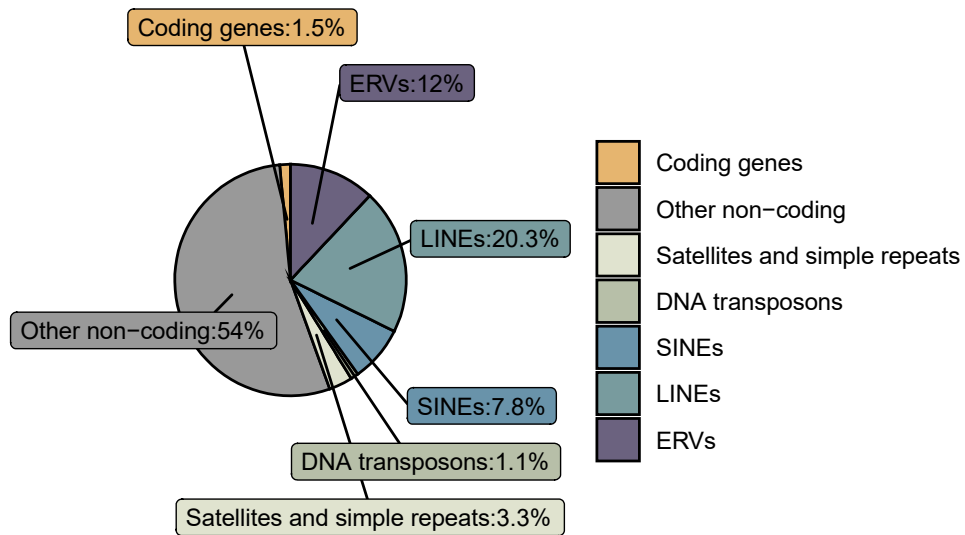


Figure 1.3: Proportion of different classes and families of TEs in the mm10 genome assembly. Updated to the mm39 assembly in 2020. Relative to coding genes and non-repetitive non-coding regions. Adapted from Smit (2013).

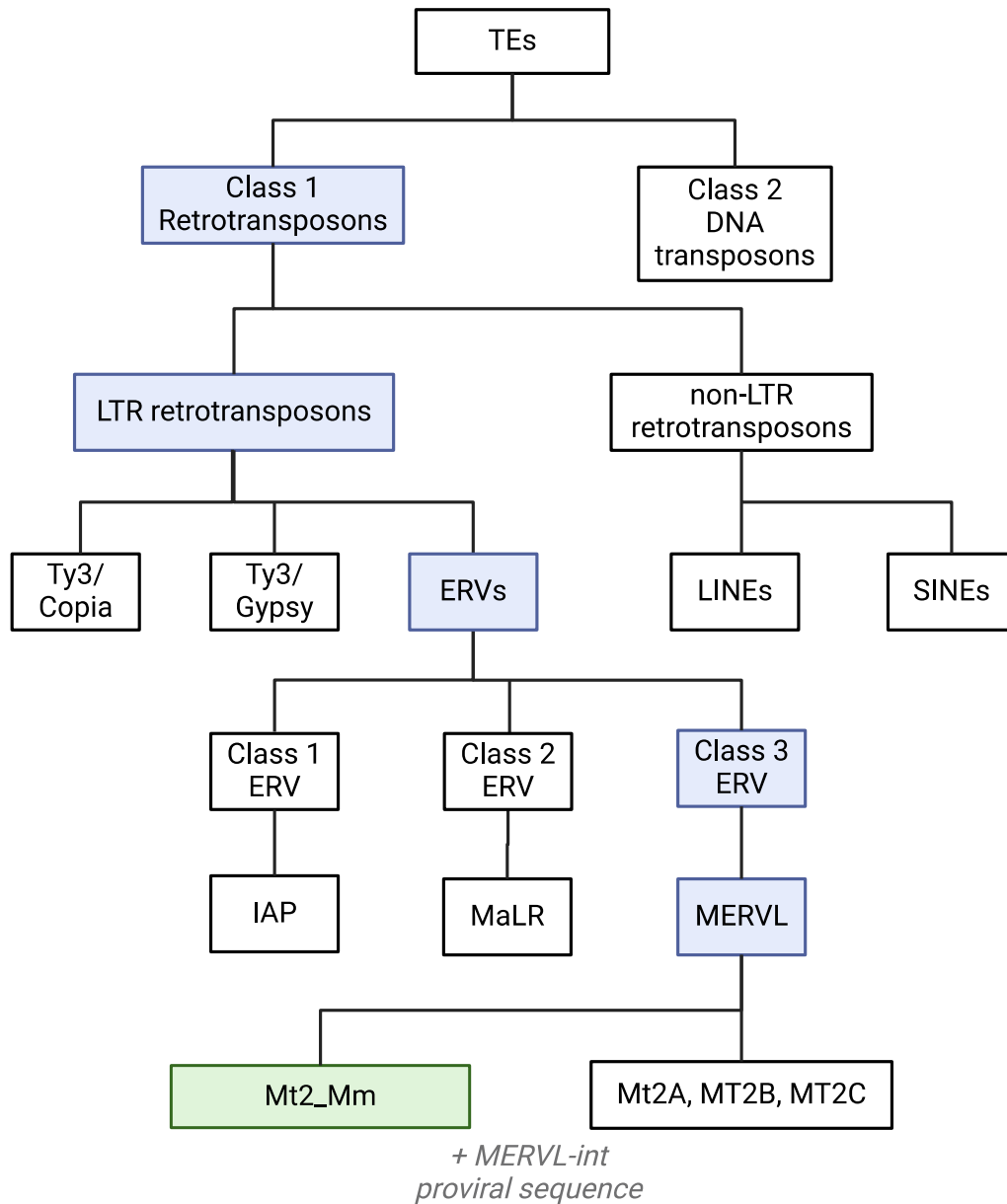


Figure 1.4: Phylogeny of TE classes, families, and subfamilies, highlighting the categorisations of MERV1 elements. Blue for

class/family/subfamily and green for canonical MERVLs. Adapted from (Bourque et al., 2018; Stocking and Kozak, 2008; Storer et al., 2021)

1.2.2 Transposon profiles in preimplantation development

The highly coordinated changes in the 3D genome structure, DNA methylation levels, chromatin accessibility, and histone marks occurring during early embryogenesis result in carefully regulated TE expression profiles. In mature mouse oocytes, MalR elements transcripts are highly abundant (these are retrotransposons closely related to the MERVL family (Stocking and Kozak, 2008)). In the mouse zygote, the earliest TE expression in the early embryos is of the MERVL transposable elements, both solo LTRs and full MERVLs, concomitant with the minor and major waves of ZGA (Kigami et al., 2003; Kruse et al., 2019). By the major wave of ZGA, LINE1 elements also are derepressed, (Berrens et al., 2021; Jachowicz et al., 2017; Percharde et al., 2018), and by the morula and blastocyst stages, SINE B1, SINE B2, and IAP elements are activated (summarised by Rodriguez-Terrones and Torres-Padilla, 2018) (Figure 1.5).

Each TE family active during embryogenesis has roles specific to the stage it is expressed or derepressed. LINE1s are key in the exit from the 2-cell stage, specifically in recruiting KAP1 to the *Dux* gene (Percharde et al., 2018). SINE elements are known to be significant donors of CTCF binding sites and therefore key in the establishment of 3D genome structure (Choudhary et al., 2023; Macia et al., 2011; Schmidt et al., 2012), as have MERVL sequences of the MT2_Mm group, specifically in 2C embryos and 2CLCs (Kruse et al., 2019). In 2011, Macfarlan et al., found that MERVL insertions (both canonical and non-canonical MERVLs) happened to be in significantly gene-poor regions and that the few genes that were within 200bp of the MERVLs were 2-cell specific genes (Macfarlan et al., 2011), suggesting a possibility that MERVLs have been co-opted into a gene regulatory function specific to murine ZGA. Most recently,

it was shown that the process of transcribing MERVL elements was indispensable to ZGA (Sakashita et al., 2023).

Transposable element families are silenced through epigenetic mechanisms, namely DNA methylation and histone mark deposition. Long-term silencing of retrotransposons is generally regulated by DNA methylation (Reik, 2007). DNA methylation is theorised to have initially evolved as a protective mechanism to silence TEs and was subsequently co-opted as a mechanism for maternal imprinting (Ondiřová et al., 2020). However, histone modifications at TE loci are also utilised as a more transient method of silencing, for example, H4K20me3 at IAP LTRs (Martens et al., 2005), and trimethylation of H3K9, H3K36, and H4K20, resulting in the induction of transcriptionally silenced heterochromatic regions (Sripathy et al., 2006). MERVLs, in contrast, are silenced after the 2-cell stage through the deposition of H3K9me2 marks (Macfarlan et al., 2011; Schoorlemmer et al., 2014), mediated by the histone chaperone FACT (Chen et al., 2020). This histone methylation mark is not mediated by SETDB1, but it is dependent on G9A and GLP, both of which are key lysine demethylases active at the ZGA stage (Maksakova et al., 2013). The methylation of TE DNA and associated histones is mediated by the recruitment of KAP1 by KRAB zinc-finger proteins (Imbeault et al., 2017; Iouranova et al., 2022; Rowe et al., 2010)

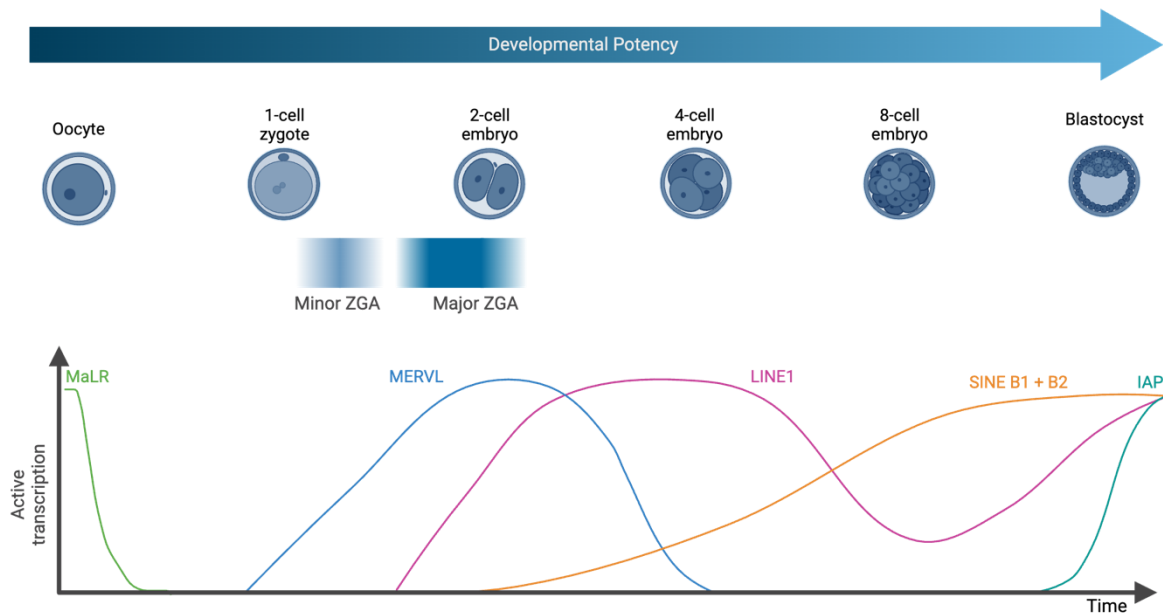


Figure 1.5: Profiles of TE expression across murine embryogenesis.

Showing peaks of active transcription relative to cell stage and potency

1.2.3 Co-option

Co-option is the process by which a feature (e.g., a genetic element) adapted to function in a specific context, is harnessed to fulfil the same function in another context. The more stringent definition of co-option states that co-option occurs only when there is a notable biological change in a system, specifically beneficial to the host organism's fitness (Fueyo et al., 2022), whereas the more flexible studies suggest that conferred molecular function is sufficient (Peaston, 2008; De Souza et al., 2013). The second definition is often used interchangeably with exaptation (whereby a feature is used for a different function to which it had originally evolved for (Gould and Vrba, 1982)). Therefore, it is arguable that TEs can be either co-opted into a host genome (if their impact on the host genome is beneficial), or exapted (since their function within the host genome is different to that of the TE itself, for instance in how endogenous retroviruses are now not producing infectious viral particles, but for example may act as *cis*-regulatory elements for a host gene). To date, a plethora of studies have further confirmed that TEs are frequently co-opted into the genomes of many metazoans, often as *cis*-regulatory elements (Table 2).

The role of TEs as regulatory elements as a principle has been of interest for some time (Banville and Boie, 1989; Britten, 1997). In humans, HERVK insertions correlate strongly with gene transcriptional units (Brady et al., 2009). The presence of promoter sequences and splice donor sites in some transposons such as ERVs is of particular relevance here, since their integration into a gene regulatory unit allows the TE to act as an alternative promoter, and thus, for splicing between the TE and the body of the gene in the battery. This phenomenon has also been observed many times with ERV elements, both canonical and non-canonical, due to the presence of promoter sequences in their LTRs; indeed, class 3 ERVL LTRs contribute the most 5' exons out

of the major families of retrotransposons (Franke et al., 2017; Modzelewski et al., 2021). After examining the proximity of TEs to certain human genes (Medstrand et al., 2002), some of the first complete assessments of TE-promoted genes were undertaken by van de Lagemaat et al., in 2003, scanning through existing databases of UTRs and promoter sequences to find TE hits in these regions, unspecific to cell type.

Species	Type	Transposon	Target Gene	Biological process	Reference
Human	Exonisation	HERV-W	Syncytin	Placental adhesion	Mi et al., (2000), Nature
Drosophila	Exonisation	Tc-1	PAX (DNA- binding domain)	Several developmental processes	Franz et al., (1996) Proc Natl Aca Sci
Jawed vertebrates (proved in lancelets)	Exonisation	TIR	RAG1/RAG2	VDJ recombination	Evidenced in Huang et al., (2016) Cell. Hypothesised by Thompson (1995) Immunity, and McBlane et al., (1995) Cell
Human	Enhancer	ERV-9	Beta-globin	Erythrocyte function	Pi et al., (2010) PNAS
Human	Enhancer	LTR5HS- HERVK	several, including RBBP4 and KZFPs	Gastrulation	Pontis et al., (2022), Nat Commun
Humans	Insulator	SINE- B2	N/A	3D genome organisation	Schmidt et al., (2012) Cell
Hominids	Promoter	HERV-9	P63	Proapoptotic, tumour suppression	Beyer et al., (2011) PNAS

Mouse	Promoter	MT LTR	D6Ert527e, fusion of Gfpt1 and Antxr1 genes	N/A	Franke et al., (2017) Genome Res
Mouse	Promoter	MT2B2	CDK2AP1	Cell proliferation in implantation development	Modzelewski et al., (2021), Cell
Human, mouse	lncRNA	ERV1	N/A	NFκB signalling	Zhou et al., (2019), MBio

Table 2. Examples of transposon co-option

1.3 MERVL elements and transcriptional control

1.3.1 Murine Endogenous retroviruses - L

Murine ERVs can be classified into 3 groups: class 1 of MERVL elements relating to retroviral leukaemia, class 2 of Intracisternal A-type particle (IAPs) and mammary tumour elements, and class 3 of MERVL (murine ERVs with a leucine tRNA primer) and MaLR (Mammalian-Apparent Long-Terminal Repeat) elements (McCarthy and McDonald, 2004). MERVL elements and MaLR elements share a strong sequence similarity in their LTR sequences, suggesting they share a common insertion event (Stocking and Kozak, 2008). Despite sharing a common ancestor with HERVLs, which were inserted during the lineage specification of placental mammals (Stocking and Kozak, 2008), MERVL insertions are the result of a recent expansion, and a substantial proportion of insertions are still retrotransposition competent (Bénit et al., 1999; Costas, 2003). As such, many MERVL insertions are not present in every murine lineage due to both insertions after speciation (Keane et al., 2011; Nellåker et al., 2012) and structural variations as part of speciation (Ferraï et al., 2023).

This study focuses on the promoter elements within the LTR of Murine Endogenous Retroviral elements with Leucine primers (MERVLs). These are a class 3 ERV (Figure 2B), with leucine PBS sites (as mentioned above). MERVL elements lack an *env* domain in their proviral sequence (Nit et al., 1997). Classically, MERVL elements are grouped by the sequence of their LTR; the canonical MERVL LTR is noted as MT2_Mm in the literature (for example in Choi et al., 2017), but several other MERVL LTRs exist (MT2A, MT2B, and MT2C). These can occur in the genome as solo LTRs, which are not retrotransposition competent, or in conjunction with their proviral sequences as full MERVLs. The internal proviral sequence of the MERVL is

denoted MERVL-int, and though it is not a transposition-competent sequence in itself, and is mostly annotated in isolation.

1.3.2 MERVL LTR promoters

As aforementioned, the structure of full ERVs is of two identical LTR sequences flanking a proviral sequence. The anatomy of each LTR is generally well conserved, between different ERV subfamilies (Benachenhou et al., 2009, 2013), and is as follows: a region of enhancer and promoter sequences called U3 (unique 3'), a short repetitive region of variable length dependent on the ERV subfamily, and a U5 (unique 5' region) with a splice donor site. Interestingly, MERVL LTRs are observed to lack a splice acceptor site entirely (Franke et al., 2017). The 3' end of the LTR contains a PBS (primer binding site). The U3 region is the least conserved between ERV subfamilies, with the majority containing promoter TATA and CAAT motifs (Figure 1.6).

How exactly the MERVL LTR promoter functions at a molecular level (for both the 5' or 3' LTRs) remains slightly obscure since no studies have fully dissected the genetic or epigenetic aspects of MERVLs alone. Broader works have focused more on either the sequences of ERVs across all metazoans (Benachenhou et al., 2009, 2013), or the molecular biology of simian and avian retroviruses. From this, we know of the presence of more DNaseq hypersensitive sites in the 5' LTR than in the 3' LTR, and that on activation of the 5'LTR as a promoter, the 3' LTR is 'shut down' and cannot drive transcription (Guntaka, 1993; Herman and Coffin, 1986; Swain and Coffin, 1993). Most works considering MERVL expression mechanisms have examined this in the context of ZGA or preimplantation development; MERVLs are confirmed to harbour binding sites for *Dux*, *Obox*, *Rex1*, *YY1/YY2*, *Gata2*, and *G9a* (Choi et al., 2017; Hendrickson et al., 2017; De Iaco et al., 2020; Kruse et al., 2019; Maksakova et

al., 2013; Schoorlemmer et al., 2014). Moreover, MERVL LTRs have recently been confirmed to also act as enhancers during the exit from totipotency (Li et al., 2023).

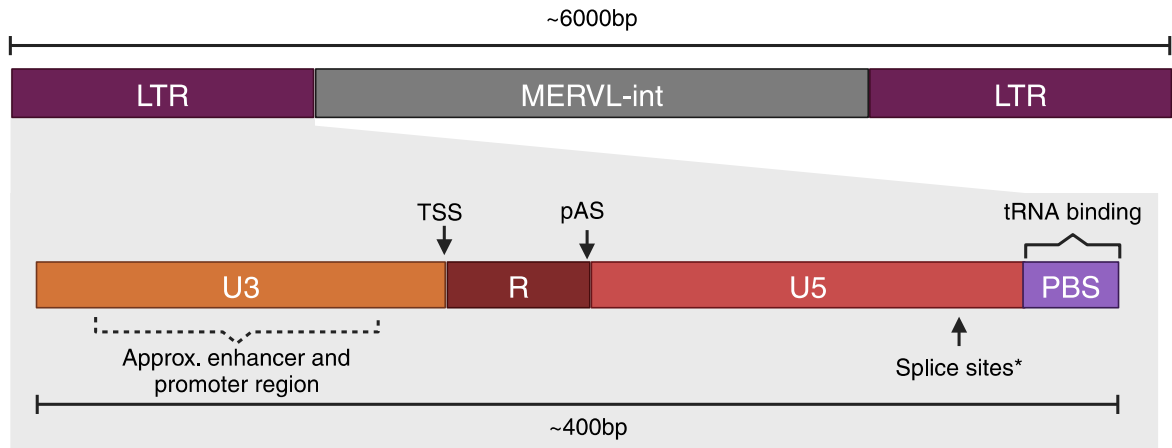


Figure 1.6: Structure of MERVL elements and regulatory elements within the LTR. Highlighting general promoter regions and transcriptional control elements: GTP cap site, polyadenylation signal, splice donor site, and Primer Binding Site (tRNA binding site for replication).

1.3.3 MERVLs in preimplantation development

MERVLs are some of the first transcripts to be expressed after fertilisation, with RNA being observed as early as the G2 phase of the 1st cell cycle (Kigami et al., 2003). After this, they peak in expression during the 2C stage, with active transcription ceasing at the end of the 2C stage, but mRNA persisting until the 8C stage and depleting entirely by the blastocyst (Sakashita et al., 2023). Their expression is known to be crucial for ZGA and later development to correctly proceed; knockdowns and knockouts confirm that absence of MERVL transcription and translation result in embryonic development arresting (Choi et al., 2017; Rosa et al., 2024; Sakashita et al., 2023).

Work by Peaston et al., 2004 was some of the first to validate multiple TE-promoted transcripts in murine embryogenesis using expressed sequence tags in mice, defining several isoforms promoted by several retrotransposons (MERVLs, ORRs, IAPs, and LINEs). Several studies have carried this further to explore the long non-coding products and protein products of MERVL-derived alternative promoters in several stages of mouse development, including in MII oocytes pre-embryogenesis, for example, a truncated isoform of the Dicer protein involved in RNA interference in the developing oocyte (Franke et al., 2017; Karlic et al., 2017). More recently, the genes *Cdk2ap1* (Modzelewski et al., 2021) and *Prmt6* (Honda et al., 2023) have been exemplified as having alternative transcript isoforms driven by MERVL LTR (specifically, MT2B2 LTRs) alternative promoters during embryogenesis, with impacts on embryo viability if the alternative isoform is depleted or knocked out entirely. However, to date, no known example of an MT2_Mm-promoted alternative transcript isoform translated to a functional protein isoform at the 2C stage exists. Furthermore, previous studies into alternative transcript isoforms have not explored

any possible differences between the promoter function of solo LTRs compared to complete MERVL elements, despite the transcriptional signal being stronger in complete MERVL elements compared to solo LTRs (Kruse et al., 2019).

While the role of the MERVL promoter starting to be understood, the role of the MERVL-int encoded proteins, specifically the MERVL-Gag protein, is contested. Until recently, the translation of the MERVL-Gag was considered not to have any impact on embryogenesis (Sakashita et al., 2023) and has mainly only been used as an indicator for correct MERVL expression (transcription and translation, for example Chen et al., 2020, Guo et al., 2022, and Xie et al., 2022). Unexpectedly, recently a study showed a crucial role for the MERVL-Gag in the exit from totipotency, whereby it bound the prefoldin protein Uri to allow for Oct4 and Sox2 degradation in 2CLCs, and the loss of MERVL-gag allowed for the exit from 2CLCs (Rosa et al., 2024). This contradictory result warrants future examination.

1.4 Sequencing technologies

1.4.1 Short- and Long-read sequencing

A link between promoter choice and downstream alternative splicing patterns has been known for some time (Pecci et al., 2001), with more recent comprehensive studies about this disequilibrium in a range of experimental organisms (Alfonso-Gonzalez et al., 2023). A recent classification method by the ENCODE consortium pushed for a move towards an isoform-based view of transcriptional complexity from the previous more gene-centric perspective, to allow for cross-dataset unification and combat the difficulties in annotating alternatively processed isoforms (Reese et al., 2023). This possible shift in perspective requires the use of long-read sequencing, since the majority of transcripts are larger than the read lengths of the current gold standard of short-read sequencing, resulting in difficulties in complete transcript structure resolution from promoter to polyA tail terminus. Long-read sequencing (developed by Oxford Nanopore Technologies, and PacBio sequencing) allows for the resolution of the exact structure of transcript isoforms, by virtue of its capture of the entire RNA (or cDNA) molecule; this enables the coupling of promoters to alternatively spliced exons and polyA tails (Amarasinghe et al., 2020; Chen et al., 2021; van Dijk et al., 2018; Kuo et al., 2020; Wang et al., 2021b). Moreover, this can overcome the 3' bias present in short-read sequencing (Wang et al., 2009).

Oxford Nanopore Technology (ONT) sequencing works by ratcheting DNA, RNA, or cDNA molecules through a nanopore protein embedded on a chip (Figure 1.7). A consistent voltage is applied to the protein, which is differentially disrupted when each base passes through (driven by a motor protein) (summarised by van Dijk et al., 2018 and Wang et al., 2021b). The changes in voltage are recorded in a 'squiggle plot'. The bases are then called using neural network base-calling tools, such as Albacore,

Guppy, Scrappie, and Flappie (Wick et al., 2019). ONT's technologies sequence a DNA or RNA molecule from beginning to end, allowing for exploration of the combinatorial effects of alternative gene promoters, alternative splicing, and more definitive assessment of changes in polyA tail length (Chen et al., 2021; Krause et al., 2019; Olsen et al., 2016). Moreover, m6A modifications to mRNA can be assayed if direct RNA is sequenced (Leger et al., 2021), giving additional insight into RNA processing RNA (Aw et al., 2020). While several studies have utilised long-read sequencing protocols to examine the transcriptomics of mouse development stages (Berrens et al., 2021; Fan et al., 2020; Qiao et al., 2020) and one of human preimplantation development (Torre et al., 2023), there appears not to have been a transcript-level integration across datasets in any study to date incorporating both long- and short-read RNA-seq data in early mouse embryogenesis.

1.4.2 CAGE-seq

Examination of promoter structures requires mapping of the 5' end of mRNA molecules exactly. CAGE-seq is a modification to short-read RNA-seq, whereby instead of capturing reads across the length of the RNA transcript, the reads with a 5' m7G are biotinylated and pulled down using streptavidin-coated magnetic beads. Through this cap-trapping method, the exact transcriptional start site of RNA can be captured, and thus the promoter can be mapped (Haberle and Lenhard, 2016; Shiraki et al., 2003; Takahashi et al., 2012) (Figure 1.8). Downstream analysis can be completed with R-based tools such as CageR and CAGEfightR (Haberle et al., 2015; Thodberg et al., 2019) to correct and assign clustered transcriptional start sites. This overcomes the 3' bias common in short-read RNA-seq (Wang et al., 2009). Recently, long-read CAGE-seq has also been developed, utilising PacBio sequencing to capture the entire mRNA molecule (Maeng et al., 2023). Initial CAGE-seq protocols required

a large amount of input RNA (Haberle and Lenhard, 2016; Takahashi et al., 2012); however, more recent developments in CAGE-seq protocols have optimised this for super-low input (Cvetesic et al., 2018). However, caveats remain in CAGE-seq reads; these can capture either only mature RNAs or RNA from a specific cell fraction, and only RNAs from reads initially 100nt or longer (Haberle and Lenhard, 2016; Takahashi et al., 2012). In this study, I utilise SLIC-CAGE data; this is a Super-Low Input Carrier CAGE-seq protocol; by the addition of “carrier” RNA molecules with endonuclease sites in their sequences (which are degraded just before sequencing), lower quantities of sample RNA can be selected during the 5’ cap trapping step. As a result, this technique can capture nascent RNAs (i.e., those not polyadenylated yet), and is not restricted to any cellular fraction.

Studies into the promoter architecture utilising CAGE-seq data have defined certain ‘grammatical’ rules during oocyte development and zygotic genome activation. The dynamics from oocytes to zygotic promoters are known to be conserved at least between zebrafish and mice (Cvetesic et al., 2020a; Haberle et al., 2014); in oocyte-specific promoters (active in oogenesis), sharp promoters with several TATA boxes clustered close together are active, but by ZGA and embryogenesis, broader promoters are activated, dependant more on nucleosome positioning than alternative promoter sequence (Cvetesic et al., 2020a; Haberle et al., 2014). While this is a deep examination of the regulatory logic of promoter choice, no such study examines TE-derived promoters, or indeed, MERVL LTR promoters, in the context of ZGA itself.

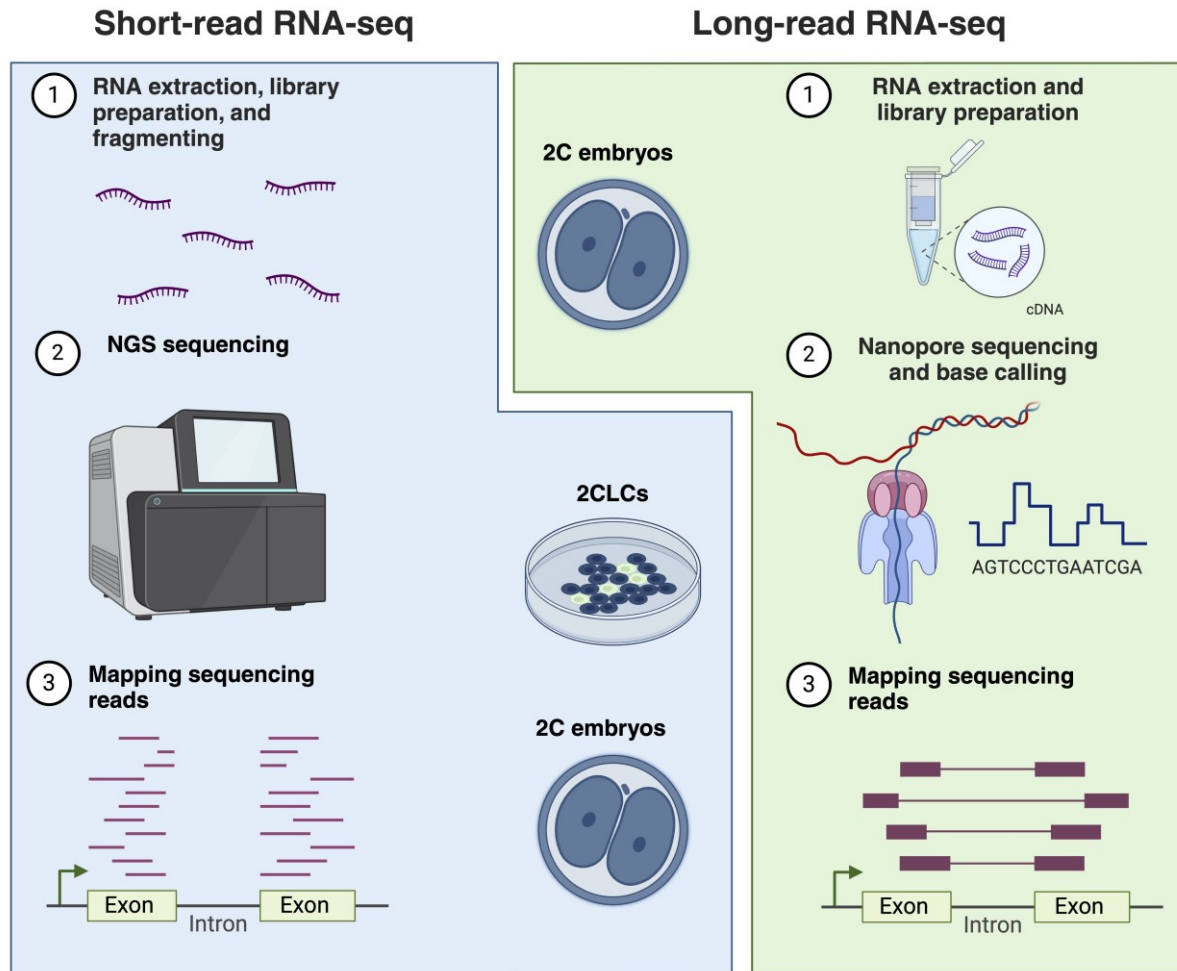
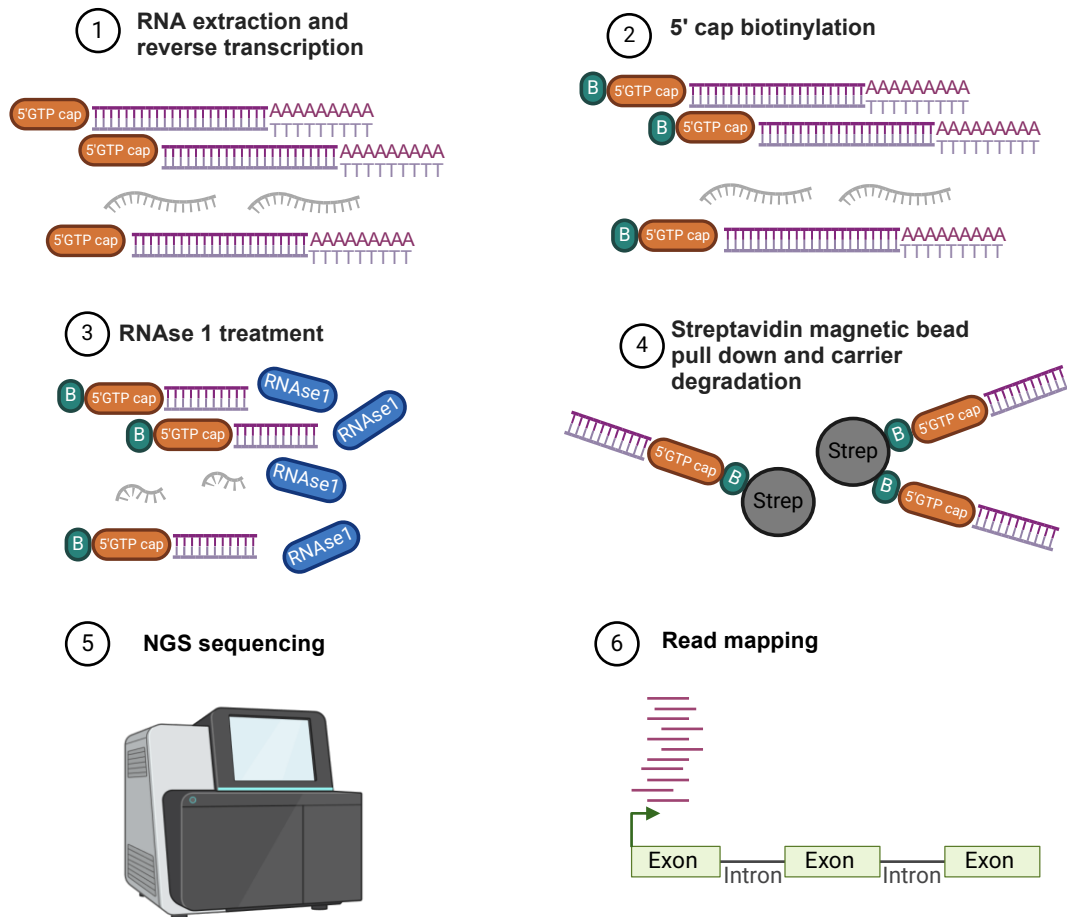


Figure 1.7: Simplified workflow (library preparation, sequencing, and read alignment) for sequencing technologies used in this study. Short-read RNA-seq library preparation shown in blue, and long-read RNA-seq library preparation (Oxford nanopore cDNA-PCR sequencing) shown in green. Cell types sequenced with each technology shown centrally.

SLIC-CAGE



Developmental stages used

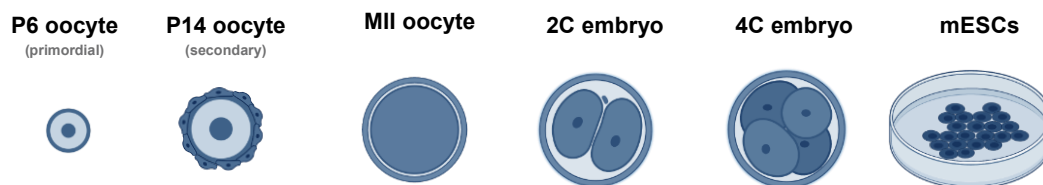


Figure 1.8: Simplified workflow for library prep, sequencing, and read alignment using SLIC-CAGE (Super Low Input Carrier Cap-analysis gene expression sequencing). Cell types sequenced using SLIC-CAGE shown below the workflow.

1.5 Aims

While studies have begun assessing the role of MERVL insertions into the mouse genome, and the importance of MERVL element expression is confirmed in ZGA, the role of the chimeric transcript isoforms transcribed is not confirmed in ZGA. To address this question, I will investigate the usage of MERVLs as alternative promoters on genes during early embryogenesis using publicly available short-read RNA-seq data, long-read RNA-seq data, and CAGE-seq data. I will analyse MERVL expression in early embryogenesis as well as in differently generated 2CLCs. Finally, I will select candidate genes for which MERVL-promoted transcripts are produced and have coding potential, and I will assess the functional impact of MERVL-promoted isoforms on 2CLCs as a model for the 2-cell stage of embryogenesis. This study opens doors to experimental validation of these isoforms, and potentially, a deeper understanding of the mechanisms regulating ZGA.

Aim 1. To define and characterise a repertoire of high-confidence MERVL-promoted alternative transcripts isoforms in 2CLCs.

For which genes do MERVLs serve as alternative promoters in 2C embryos and 2CLCs? How do the characteristics of MERVL-promoter transcripts differ from reference transcripts of the same genes (promoter location, length, exon count)? This will indicate whether any consistent characteristics of transcription are noted in MERVL-promoted alternative transcripts.

- i. Exploration of selected MERVL element signals as promoters using publicly available CAGE-seq data for early embryonic developmental stages and an in vitro system (2C, 4C, 8C, and mESC).

- ii. Assembly of transcript isoforms from several publicly available short-read and long-read RNA-sequencing datasets covering a range of early embryonic stages and in vitro models for the 2-cell stage.
- iii. Identification of supported transcript isoforms between datasets using transcript isoforms assembled using long-read RNA-sequencing data.
 - i. Comparison to published MERVL-gene chimaera studies for MERVL elements at the 2C stage.

Aim 2. To understand the functional impact of selected MERVL-promoted alternative protein isoforms in 2CLC function.

Do alternative transcript isoforms driven by MERVL alternative promoters have any consistent changes in expression and translated structure, relative to their reference isoform? Do MERVL-promoted isoforms have coding potential? Which protein domains are modified if a transcript is driven by a MERVL element? This will indicate how many MERVL-promoted isoforms result in translated proteins.

- iv. Characterisation of alterations to coding potential, open reading frame, and protein domains of MERVL-promoted isoforms relative to reference isoforms of the same genes.
- v. Assessment of high-confidence transcript expression across progressing early embryonic developmental stages (oocyte, zygote, early 2C, late 2C, 4C, 8C, morula, early blastocyst inner cell mass, early blastocyst TE trophectoderm, late blastocyst inner cell mass, late blastocyst trophectoderm, and mESC) using long-read data.
- vi. Analysis of coding potential of transcript isoforms, and homology search for any potential changes to protein domains.

2. Materials and Methods

Code availability.

Custom analysis scripts, package versions, and IGV sessions are available at:

https://github.ic.ac.uk/sa4321/MERVL_isoforms

Data availability.

Short-read RNAseq data was obtained from the European Nucleotide Archive ((Hendrickson et al., 2017): PRJNA293908, (Ishiuchi et al., 2015): PRJEB6962, (Sakashita et al., 2023): PRJNA805243). Long-read RNA-seq data was also obtained from the European Nucleotide Archive ((Berrens et al., 2021): PRJEB40562, (Fan et al., 2020): PRJNA616184). CAGE-seq data was obtained from ArrayExpress (MII oocyte, 2C, and 4C, (Cvetesic et al., 2020b): E-MTAB-9757, mESC, 5ng and 10ng samples only (Cvetesic et al., 2018): E-MTAB-6519). Data details are included in table 3. The mm39 reference genome assembly and reference fasta were obtained from UCSC, as was the repeatmasker track.

Dataset	Cell type	Sequencing technology	Read details	Accession code
(Hendrickson et al., 2017)	2CLC: DUX overexpression - mESC	Illumina HiSeq 2500	125bp, PE, bulk RNA-seq , polyA selected, stranded	PRJNA338965
(Ishiuchi et al., 2015)	2CLC: Caf-1 p150 subunit siRNA 2CLC: Caf-1 p60 subunit siRNA - mESC	Illumina HiSeq 2500	50bp, PE, bulk RNA-seq , rRNA depleted, unstranded	PRJEB6962
(Sakashita et al., 2023)	- siRNA control 2C embryo - CRISPRi control 2C embryo	Illumina HiSeqX	150bp, PE, bulk RNA-seq , rRNA depleted, stranded	PRJNA805243
(Berrens et al., 2021)	2C embryo	MinION R9.4.1, Oxford Nanopore	SE, CELLO-seq protocol (cDNA-PCR), polyA selected, unstranded	PRJEB40562
(Fan et al., 2020)	- MII oocyte - Zygote - Early 2C embryo - Late 2C embryo	MinION 9.4.1, Oxford Nanopore	SE, SCAN-seq protocol (cDNA-PCR), polyA selected, unstranded	PRJNA616184

	• 4C embryo • 8C embryo • Morula • Early blastocyst ICM • Early blastocyst TE • Late blastocyst ICM • Late blastocyst TE • mESC			
(Cvetesic et al., 2018)	• mESC	Illumina HiSeq 2500	50bp, SE, Super Low Input Carrier.CAGE-seq (SLIC-CAGE), unstranded	PRJEB25104
(Cvetesic et al., 2020a)	• MII oocyte • 2C embryo • 4C embryo	Illumina HiSeq 2500	50bp, SE, Super Low Input Carrier.CAGE-seq (SLIC-CAGE), unstranded	PRJEB41307

Table 3: datasets analysed in this study

MT2-Mm MERVL locus definition.

BED files for conserved full MERVL and solo LTR loci (MT2-Mm only) with and without *Dux* binding were obtained via personal communication with Dr Kai Kruse (initially curated as per methods in (Kruse et al., 2019)). These were lifted over from mm10 to mm39 in the command line using `UCSC liftOver`, and combined into one BED file containing locations of *Dux*-bound full MERVLs (n = 369) and solo LTRs (604). Loci of 5' and 3' MT2-Mm LTRs of full MERVLs were extracted by overlapping full MERVL loci with MT2-Mm loci in the UCSC RepeatMasker GTF, removing mis-annotated 5' and 3' MT2_Mms. MERVL-int exact starts and ends were obtained by overlapping full MERVLs with MERVL-int annotations in the UCSC RepeatMasker GTF, removing MERVL-int sequences which had MT2_Mm loci removed from the prior analysis.

CAGE-seq processing

CAGE-seq reads read quality was initially assessed with `FastQC` (Andrews, 2020). All raw sequencing files in fastq format were trimmed with `fastp` (Chen et al., 2018) using the following options: `--cut_tail_window_size 1, --cut_tail_mean_quality 20, -3, --correction, --trim_poly_x`. Trimmed reads were aligned with `Bowtie2` (Langmead and Salzberg, 2012) using all default parameters, and sorted and converted from SAM to BAM on-the-fly. This was done for retaining all multimapping reads (maximum k 5), retaining the single best alignment (SBA) for reads, and filtering for only uniquely-mapping reads. Biological and technical replicates were merged at this stage.

Short-read RNA-seq processing

Sequencing read quality was initially assessed with `FastQC` (Andrews, 2020).

All raw sequencing files in fastq format were trimmed with `fastp` (Chen et al., 2018)

using the following options: `--cut_tail_window_size 1, --`

`cut_tail_mean_quality 20, -3, --correction, --trim_poly_x.`

Preliminary pseudo-alignment of short-read RNA-seq datasets was completed using

`Salmon v1.5.2` (Patro et al., 2017). `STAR` (Dobin et al., 2013) was then used to align

the data, keeping only single best aligning reads, with its 2-pass alignment strategy

per sample, with the following options: `--outSAMunmapped Within, --`

`outSAMattributes NH HI AS nM jM jI, --sjdbGTFfile`

`mm39.ncbiRefSeq.gtf, --winAnchorMultimapNmax 50, --`

`outFilterMultimapNmax 1, --chimOutType WithinBAM.` Since all short-

read data was paired-end, SAM `XS` tags were added to the BAM files using `STAR's`

supplementary `tagXSstrandedData.awk` script. Post-alignment filtering to retain

only uniquely mapping reads was completed with `Samtools` (Li et al., 2009), to

filter out any remaining non-primary alignments, supplementary alignments, and

unpaired reads as defined by the `Samtools` format. Biological replicates were merged

with `Samtools` for visualisation purposes.

Long-read RNA-seq processing

Sequencing read quality was initially assessed with `FastQC` (Andrews, 2020).

Out of the eight CELLO-seq RNA-seq samples available (Berrens et al., 2021), two

had adapter content over 20% and thus were discarded from further use. No samples

were discarded from the SCAN-seq dataset (Fan et al., 2020). Long-read sequencing

data was aligned to the mm39 genome using `Minimap2` (Li, 2018), using the

following options: `-K 200M --eqx --MD -ax splice --secondary=no`. Per condition, resulting BAM files were merged with `Samtools` (Li, 2011) for visualisation purposes.

Isoform assembly and definition of MERVL-promoted isoforms.

Transcriptomes were assembled with `Stringtie` (Pertea et al., 2015), inputting single replicate BAM files of uniquely mapping reads, with the minimum predicted transcript threshold set to 50bp. Biological replicates were merged post-assembly with `Stringtie` in merge mode (Pertea et al., 2016). `GFFcompare` was used (Pertea et al., 2020), to annotate isoforms according to their parent genes, and to classify the variation of each isoform intron chain relative to its parent gene. Custom `R` scripts were utilised to define MERVL-promoted isoforms by selecting the first exon of each transcript and overlapping it with conserved, *Dux*-bound full MERVLs and solo LTRs. Mono-exonic transcripts and transcripts with exonised MERVL elements (i.e., those where the isoform had a class code '=', since these are identical to known transcript isoforms) were removed from further analysis, as were transcripts classified with `GFFcompare` as "s", "x", "i", "e", "r", "u", "p", "y". Genes with novel MERVL-promoted transcript isoforms were selected and compared between conditions, and selected for further analysis if they were present in either of the 2C long-read transcriptomes and one short-read transcriptomes (2CLC or 2C). Transcript isoforms were extracted for these reads from both long-read transcriptomes and merged with `Stringtie` in merge mode (Pertea et al., 2016) to produce one high-confidence list of MERVL-promoted isoforms transcribed at the 2C stage.

Expression analysis of novel isoforms

Transcriptomes for long-read datasets (CELLO-seq, (Berrens et al., 2021) and SCAN-seq (Fan et al., 2020)) were merged with `Stringtie` in merge mode (Pertea et al., 2016) to produce a novel 2C specific transcriptome, using default parameters, including the UCSC mm39 reference genome. `GFFread` was used to retrieve mRNA sequences from this novel transcriptome GTF in fasta format. For long-read data (all developmental time points present in (Fan et al., 2020) dataset), aligned BAM files of each blastomere were input to `Salmon v1.5.2` (Patro et al., 2017) using the `--ont` flag to denote nanopore sequenced long reads and quantified using the novel transcriptome fasta, as defined above. Since the custom transcriptome's transcript IDs are produced after merging the long-read transcriptomes, the transcript IDs are unique to this study's assembly. The output was fed into `DESeq2` (Love et al., 2014) to assess differences between conditions and expression levels of MERVL-promoted novel transcripts. MERVL-promoted transcripts and reference counterparts were defined from the custom transcriptome GTF by inputting the GTF of high-confidence MERVL-promoted isoforms as the 'reference' to `GFFcompare` and extracting transcripts with `class_code '='` to be MERVL-promoted.

Gene ontology analysis

Gene Ontology analysis for high-confidence genes with MERVL-promoted transcripts was completed with `clusterProfiler` (Wu et al., 2021). Gene universes were set as the default mm39 reference genome, the novel 2C transcriptome, or a mESC-specific transcriptome constructed by merging assembled transcriptomes for mESC samples from Hendrickson et al., (2017) and Ishiuchi et al., (2015) only, using `Stringtie` in merge mode.

Signal visualisation at MERVL loci

For all RNA-seq and CAGE-seq data, expression analysis for visualisation purposes was completed with tools in the `deepTools` suite (Ramírez et al., 2016). BigWigs were generated for merged biological and technical replicates using `bamCoverage`, normalised using BPM and with a bin size of 10bp. `computeMatrix` was used to generate count matrices for full-MERVL and solo LTR loci, as well as for individual 5' and 3' Mt2_Mms and MERVL-int sequences as defined above, with reference points 'TSS' (i.e. the 5' end of the MERVL element). All heatmaps were plotted with `plotHeatmap`, including metaplots with window size 1kb upstream of the 5' terminus of the 5'LTR/solo LTR and 20kb downstream, with `--zmax` set to 4 for colour scaling.

Transcript isoform visualisation

Splice junctions were extracted from BAM files of all merged biological replicates all RNA-seq data (short- and long-read both). These were annotated according with their transcript isoform using `RegTools Junction Extract` (Cotto et al., 2023), and reformatted to a link file format with a custom bash script. Only splice junctions supported by over 5 split reads were retained. Link files, bigwig files, and GTF files of the same samples were input into `pyGenomeTracks` (Lopez-Delisle et al., 2021), with link files' thickness proportional to number of reads supporting the splice junction and Y-axis height formatted for appearance (units: TPM). The bigwig file for 2C CAGE-seq data was visualised with `pyGenomeTracks` also.

Coding potential examination, ORF prediction, and protein domain homology search

High-confidence transcript isoforms were input to `GFFread` with the mm39 genome fasta (Pertea et al., 2020) to extract mRNA sequences, and subsequently into `CPAT` (Wang et al., 2013) to calculate coding potentials for each transcript isoform. Custom `R` scripts were used to examine coding potential distributions, predicted ORFs, and integrate expression data from the long-read expression analysis. The MERV1-promoted open reading frame with the highest coding potential per gene was taken as the ‘best’ open reading frame; if more than one open reading frame had an equal highest coding potential, the open reading frame with the highest average TPM in late 2C embryo samples from SCAN-seq (Fan et al., 2020) was taken. This was repeated for the reference open reading frame for each of these genes. Finally, the mRNA sequences output by `CPAT` for these open reading frames were translated using the `Biostrings::translate` in `R`, and input to InterProScan (Jones et al., 2014) with the application options `CDD`, `COILS`, `Gene3D`, `HAMAP`, `MobiDBLite`, `PANTHER`, `Pfam`, `PIRSF`, `PRINTS`, `PROSITEPATTERNS`, `PROSITEPROFILES`, `SFLD`, `SMART`, `SUPERFAMILY`, and `TIGRFAM`. Domains were then filtered manually for visualisation within peptides.

Chapter 3: Results

3.1 MERVL elements are independently expressed at the 2-cell stage embryo

Previous work has shown that during embryogenesis, the active expression of MERVL elements initiates with minor ZGA at the S-phase of the 1-cell stage embryo and peaks at the 2-cell stage of mouse embryogenesis, with transcripts dropping by the 4-cell stage and being fully degraded by the 16-cell stage (Hendrickson et al., 2017; Kigami et al., 2003; Kruse et al., 2019; Sakashita et al., 2023).

Numerous works have found chimeric transcripts originating in MERVL elements (including solo MT2_Mm, MT2A, MT2B, and MT2C LTRs and full MERVL elements, where these LTRs flank a MERVL-int sequence) which read through into downstream genes (Evsikov et al., 2004; Honda et al., 2024; Macfarlan et al., 2012; Modzelewski et al., 2021; Peaston et al., 2004). Out of these, only solo LTRs and full MERVL elements comprising MT2_Mm sequences are considered canonical MERVL sequences (as referenced in Choi et al., (2017)), since they contain the MERVL consensus sequence. All MERVL elements peak in expression at the 2-cell stage (Evsikov et al., 2004; Honda et al., 2024; Macfarlan et al., 2012; Modzelewski et al., 2022; Peaston et al., 2004); however, in this analysis, I will assess only MT2_Mm canonical MERVL solo LTRs and full elements present in both the C57Bl6/J and the 129P2/Ola mouse lines, since these strains are used in the datasets analysed in this study.

While careful analysis has been conducted using short-read RNA-seq technologies to assess MERVL expression and resulting transcripts during murine ZGA, no analyses to date have mapped the independent promoter activity of MERVL

loci at this developmental time point, nor have they examined the transcriptional signal of the entire chimeric mRNA molecule. To this end, I utilise CAGE-seq technology, and both short- and long-read RNA-seq technology.

3.1.1 Curation of a list of MERVL elements

Prior work by the lab has curated a list of full canonical MERVL elements (MT2_Mm LTRs flanking MERVL-int sequences) for the mm10 mouse genome version, accounting for polymorphic inter-strain TE insertions (Kruse et al., 2019; Nellåker et al., 2012). In this analysis, only insertions of canonical full MERVL elements and solo MT2_Mm LTRs common to C57Bl6/J mice (the mouse strain for the reference genome) and 129P2/Ola (for which our datasets' E14 mESCs, and hence 2CLCs, are derived) (Hendrickson et al., 2017; Kruse et al., 2019) were selected, and lifted over to the mm39 genome version, yielding 973 total elements (369 full MERVL elements, and 604 solo MT2_Mm LTRs). During lift-over, one MERVL element was removed since it now fell at the end of a chromosome.

I examined these MERVL elements case by case to find other possible MERVLs to remove. For comparison, a case where a full MERVL element with 2 intact LTRs and an intact MERVL-int segment is also shown (Figure 3.1, top). One case was found of a full MERVL which was fused; three solo LTRs flanked 2 MERVL-int segments (Figure 3.2, bottom). I chose to retain these, since technically, there was at least 1 full MERVL here. Further examination illuminated one instance where the 3' LTR was from an MT2C MERVL, not an MT2_Mm (Figure 3.2, top). One final case was when the 3' LTR was present, but the annotated full MERVL range extended past it into another transposon (Figure 3.2, bottom). After removing these, this resulted in a final number of 367 total full MERVL elements for this study. While the average sizes for MERVL elements are in line with the literature (Figure 3.3), the

minimum size for full MERVL elements varies greatly, attributable to varying lengths of MERVL-int segments.

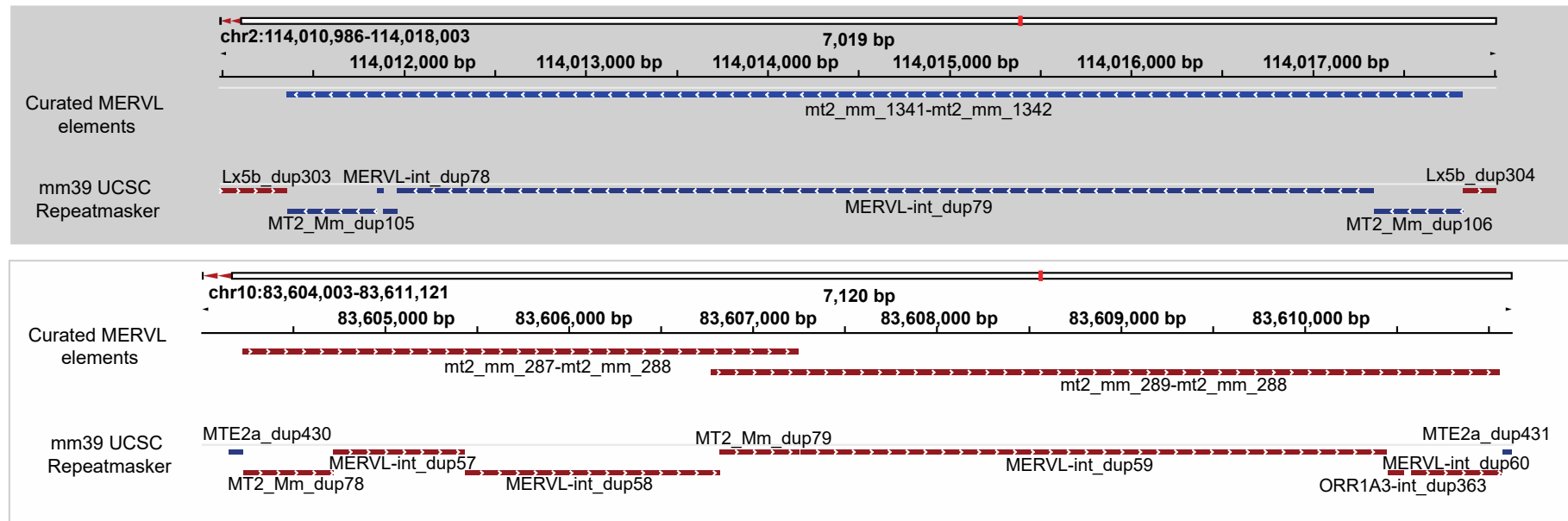


Figure 3.1: IGV tracks of MERVL elements conserved between C57Bl6J and 129P2/Ola genomes. Top (grey): example full MERVL element with 5' and 3' LTRs intact, flanking a MERV-int segment. Tracks top-down: Original MERVL elements curated by Kruse et al., (2019) lifted over to the mm39 genome, and the UCSC RepeatMasker transposable element track. Bottom (white): Fused MERVL element with two MERV-int segments between 3 LTRs. Tracks top-down: Original MERVL elements curated by Kruse et al., (2019) lifted over to the mm39 genome, and the UCSC RepeatMasker transposable element track.

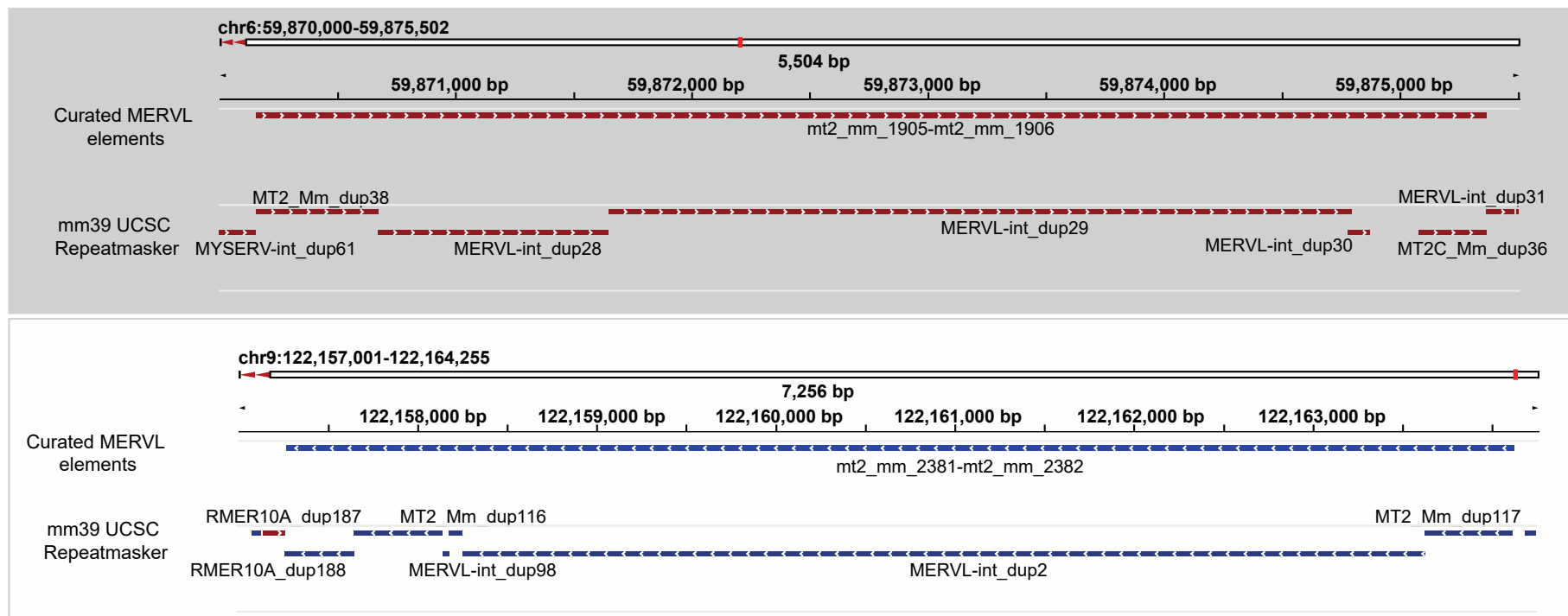


Figure 3.2: IGV tracks of MERV elements conserved between C57Bl6J and 129P2/Ola genomes. Top (grey): MERV element excluded from further analysis, since its 3' LTR is an MT2C_Mm element and not an MT2_Mm element. Tracks top-down: Original MERV element file from Kruse et al., (2019) lifted over to the mm39 genome; mm39 reference transcriptome (no gene models are present in this example), and the UCSC RepeatMasker transposable element track. Bottom (white): MERV element excluded from further analysis since the full element extends past the 3' LTR. Tracks top-down: Original

MERVL element file from Kruse et al., (2019) lifted over to the mm39 genome; mm39 reference transcriptome (no gene models are present in this example), and the UCSC RepeatMasker transposable element track.

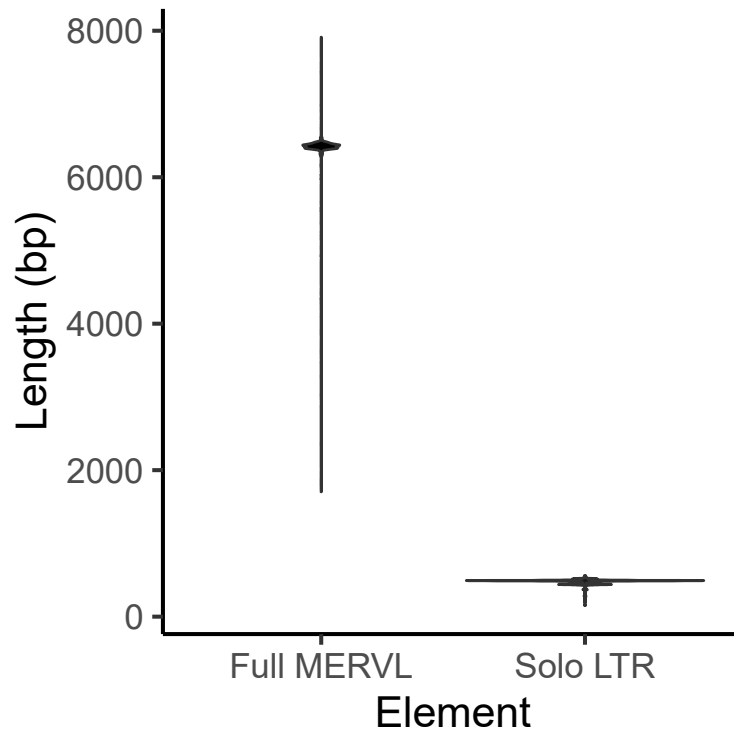


Figure 3.3. Distribution of lengths of MERVL elements conserved between C57Bl6J/129P2/Ola strains. Full MERVL elements (LTR=MERVL-int+LTR) and solo LTRs (LTR only), lengths in base pairs.

3.1.2 MERVLs show transcriptional signals in vivo and in vitro at the 2C stage.

Transcriptional signal initiating in MERVL elements and reading through into downstream genes has been previously confirmed (Evsikov et al., 2004; van de Lagemaat et al., 2003; Macfarlan et al., 2011, 2012; Modzelewski et al., 2021); however, this has not been fulfilled by utilising long-read data to confirm results extracted from short-read data. To this end, transcriptomes for 2C embryos were assembled from both short- and long-read (Oxford Nanopore) RNA-seq datasets, and for two methods of 2CLC induction (*Dux* overexpression and knock-down of two major subunits of *Caf-1*) (table 3). Post-trimming with Fastp (Chen et al., 2018) short-read RNA-seq data was aligned with STAR (Dobin et al., 2013), and only uniquely mapping reads were retained (Appendix figure 1.1). Long-read data was aligned with minimap2 (Li, 2018).

Short-read RNAseq data at 2C embryos and in 2CLCs exhibits a strong signal of transcription, initiating within MERVL elements at their LTRs, known to contain promoter sequences, and driving through MERVL-int segments of full MERVL elements), in 2C embryos and 2CLCs alike (Figure 3.4). MERVL expression is either entirely null or negligible at LTRs and internal sequences in all mESC samples, as is expected since MERVL expression is known to drop past the 4-cell stage in embryogenesis (Sakashita et al., 2023).

Transcription visibly reads through past the termini of the 3' LTRs (in the case of full MERVL elements) and the termini of solo LTR elements in the *Dux* overexpression 2CLC dataset, as previously observed by Kruse et al., in 2019. This expression pattern is also notable in the *Caf-1* knockdown 2CLCs, and in the 2C embryos (both short- and long-read sequencing), albeit to a lesser level (Figure 3.4).

This is not attributable to sequencing depth since the *Dux* overexpression dataset is the least deeply sequenced but has the highest level of read-through transcriptional signal (Appendix Figure 1.1). Thus, this highlights the differences in *Dux* overexpressed 2CLCs compared to *Caf-1* knockdown as a model for 2C embryos, possibly in that overexpression systems can exhibit ‘leaky’ phenotypes, especially for systems overexpressing a transcriptional regulator (Lee et al., 2012; Razo-Mejia et al., 2018).

Previous studies have seen numerous instances of transcription reading through from ERVs into genes, not only into downstream intergenic regions (van de Lagemaat et al., 2003; Macfarlan et al., 2011; Modzelewski et al., 2021; Peaston et al., 2004). In long-read RNA-seq data of 2-cell embryos, punctate regions of increased signal are observed downstream of the MERVL elements in all RNA-seq datasets (Figure 3.5). This is likely attributable to the signal for exonic loci only. Since both long-read RNA-seq library preparation methods involved polyA-selection of RNA (table 3), only mature transcripts are retained, therefore intronic RNA is absent in this data. Hence, I conclude that this analysis has recapitulated prior results that transcription occurs in MERVL elements and reads through to downstream genes.

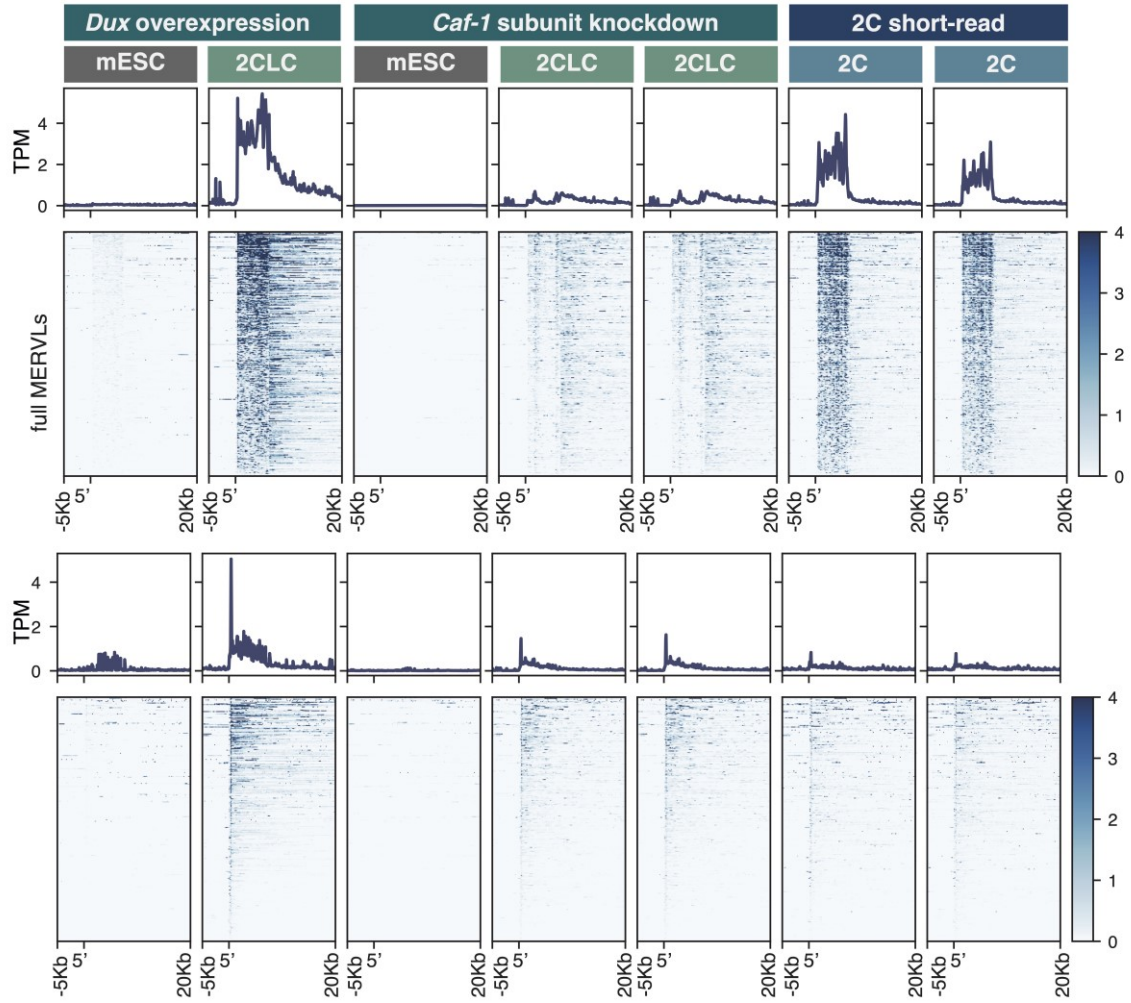


Figure 3.4: MERVL expression in 2CLCs, mESCs, and 2C embryos using short-read data. Dataset and sample indicated above each metaplot and heatmap. Short-read RNAseq signal of full MERVL elements (top) and solo LTRs (bottom) in average TPM (mean). All transposons are oriented 5'-3' irrespective of origin strand. Panel shows 5Kb upstream of element 5' terminus, and 20Kb downstream of 5' terminus. Data analysed for *Dux* overexpressed control mESCs, *Dux* overexpressed 2CLCs, *Caf-1* knockdown control mESCs, *Caf-1* p150 subunit knockdown 2CLCs, *Caf-1* p60 subunit knockdown 2CLCs, and 2C embryos (control for MERVL knockdown and control for MERVL CRISPRi, in order). Data analysed from Hendrickson et al., (2017), Ishiuchi et al., (2015), and Sakashita et al., (2023).

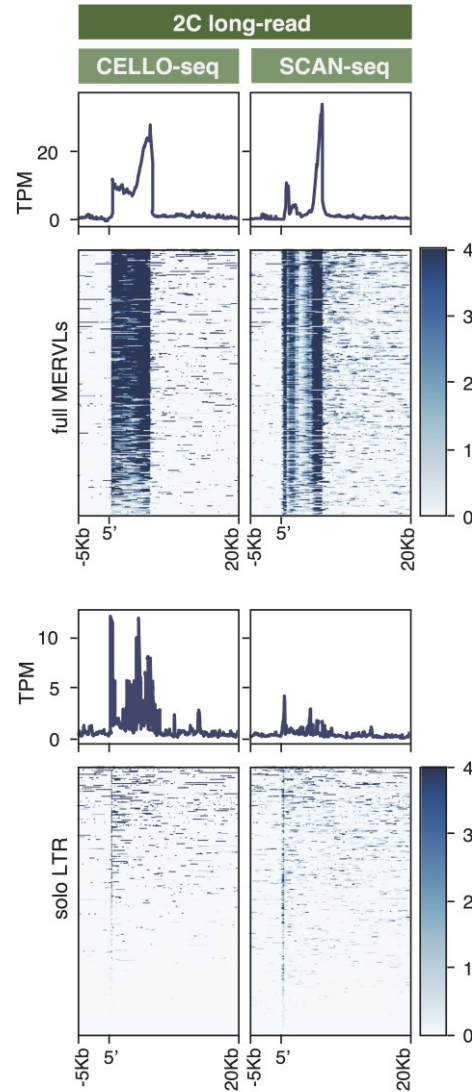


Figure 3.5: MERVL expression in 2C embryos using long-read data.

Dataset and sample indicated above each metaplot and heatmap. CELLOseq and SCANseq long-read RNA seq signal of full MERVL elements (top) and solo LTRs (bottom) in average TPM (mean). All transposons are oriented 5'-3' irrespective of origin strand. Panel shows 5Kb upstream of element 5; terminus, and 20Kb downstream of 5' terminus. Data analysed from Berrens et al., (2020) and Fan et al., (2020).

3.1.3 CAGE-seq data shows MERVL expression initiates in LTRs

Prior work has begun confirming the role of MERVL elements as alternative promoters for genes in mouse preimplantation development (Evsikov et al., 2004; van de Lagemaat et al., 2003; Macfarlan et al., 2011, 2012; Modzelewski et al., 2021). However, CAGE-seq data has not been utilised to assess the promoter signal within MERVL LTRs as yet. CAGE-seq data for the earliest developmental stages (2C stage, 4C stage, and mESCs) and for follicular oocytes (from postnatal day P6 and P14 mice) and mature MII oocytes (Cvetesic et al., 2018, 2020a) was trimmed with Fastp (Chen et al., 2018) and aligned with Bowtie2 (Langmead and Salzberg, 2012).

Initial assessment of uniquely mapped reads (i.e., those solely aligning to one location, with no equal mapping score to align elsewhere with Bowtie2) shows sparse, minimal transcriptional start site signal at either the full MERVL elements or solo LTRs (Figure 3.6, top). This brings to light the technical issues with aligning 50bp length read, single-end sequenced data to highly repetitive elements which are longer than the CAGE-seq reads. However, it is still possible to conclude that certain full MERVL elements and solo Mt2_Mm LTRs have sequences unique enough to have reads mapped to their loci. Retaining the single-best alignment (i.e., reads mapped only once, but with the capacity to be mapped to ≥ 1 location) shows a stronger transcriptional signal at all MERVL loci (Figure 3.7, bottom). Given the highly repetitive nature of MERVL elements, it stands to reason that retaining these reads gives a practical workaround to the technical limitations of CAGE-seq data (primarily short read length, but also the low sequencing depth resultant of cap-trapping). The 2C and 4C stages revealed consistent signals at MERVL LTRs, supporting the notion that transcription starts at the LTRs during ZGA. As expected, a minimal promoter signal is seen at mESCs, for both full MERVLs and solo LTRs,

where the MERVL transcriptional signal is reported to have dropped to null (Figure 3.6, bottom) The solo LTR transcriptional signal at 2C stage embryos is comparable to that of the 5' LTR in both CAGEseq and RNAseq (Figure 3.7, top and bottom).

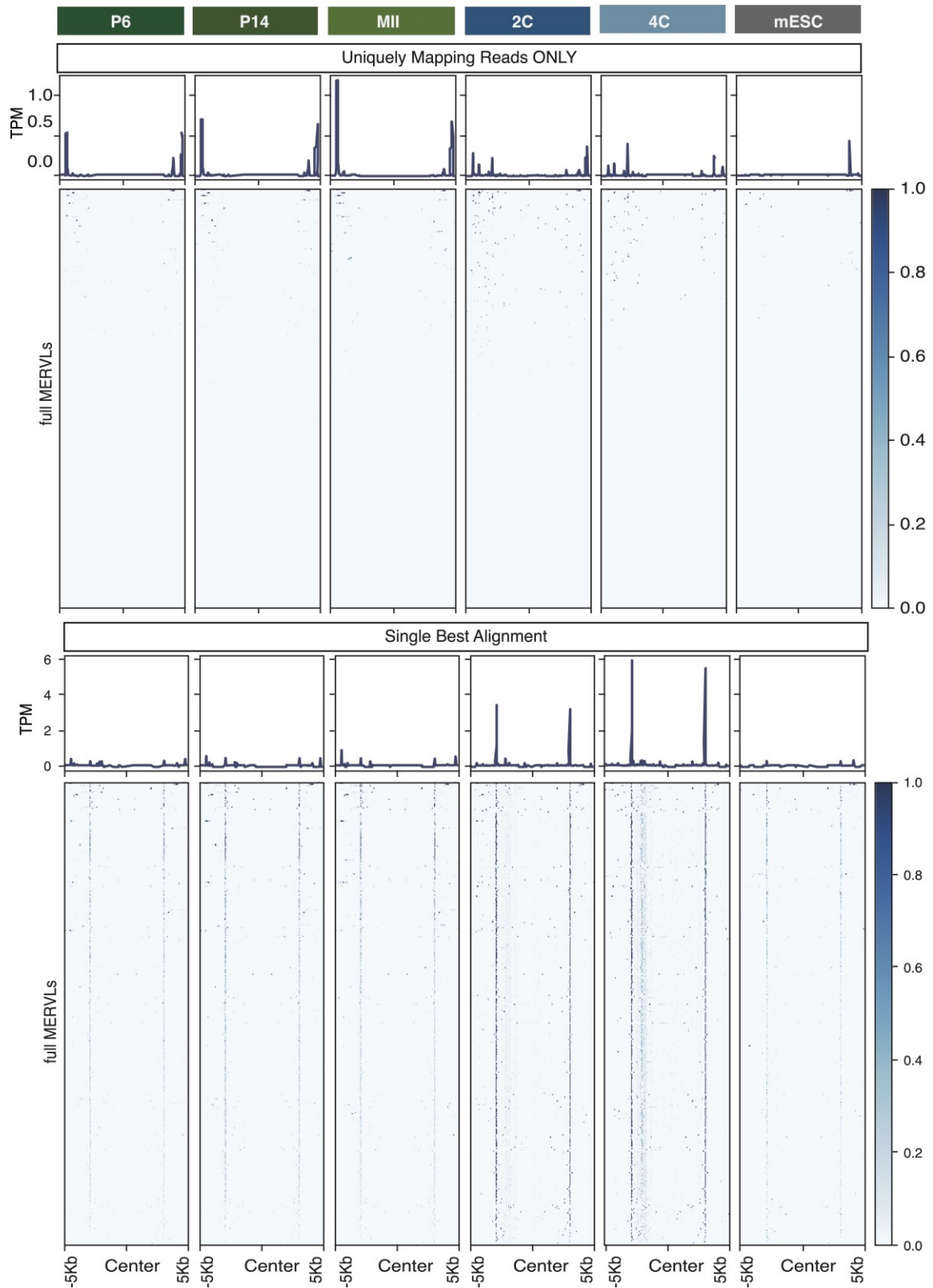


Figure 3.6. Transcription initiates in MERVL LTRs across late oocyte maturation and early embryogenesis. Sample indicated above each metaplot

and heatmap. CAGE-seq signal of **full MERVL** elements in average TPM (mean). Top panel shows uniquely mapping reads only, and bottom shows single best alignment for all reads. Panel shows 5Kb flanking windows either side of MERVL elements centre. Data analysed for P6 oocytes, P14 oocytes, MII oocytes, 2C stage embryos, 4C stage embryos, and mESCs. Data analysed from Cvetesic et al., (2018) and (2020).

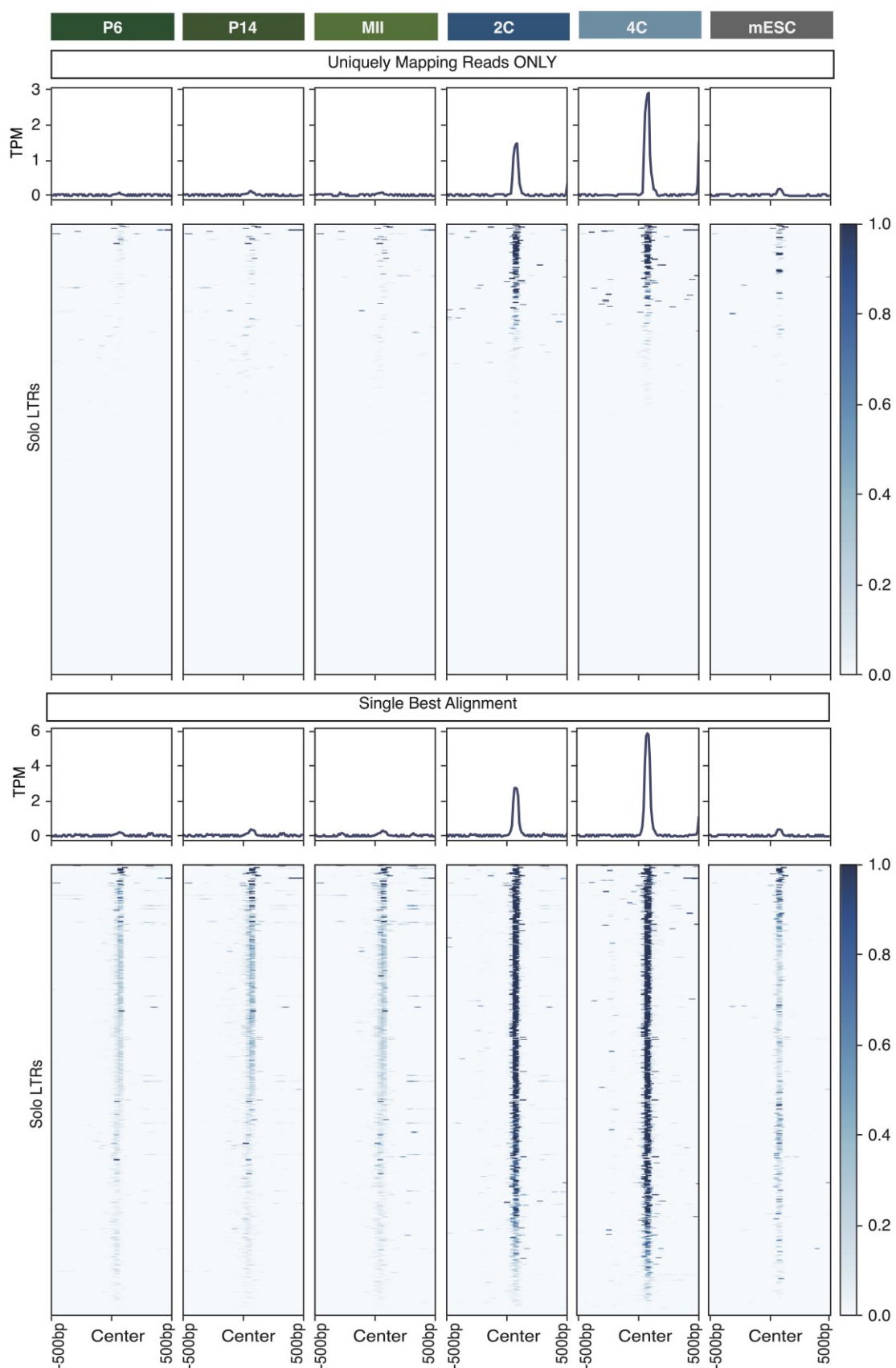


Figure 3.7 Transcription initiates in MERVL LTRs across late oocyte maturation and early embryogenesis. Sample indicated above each metaplot and heatmap. CAGE-seq signal of **solo LTR** elements in average TPM (mean). Top panel shows uniquely mapping reads only, and bottom shows single best alignment for all reads. Panel shows 5Kb flanking windows either side of MERVL elements centre. Data analysed for P6 oocytes, P14 oocytes, MII oocytes, 2C stage embryos, 4C stage embryos, and mESCs. Data analysed from Cvetesic et al., (2018) and (2020).

3.1.4 Transcriptional signal initiating in LTRs is biased to the sense strand

Unexpectedly, there appears to be minimal CAGE-seq signal in MII oocytes; since these are transcriptionally dormant, CAGE-seq data from follicular oocytes was also processed, revealing transcription initiating weakly in MT2_Mms and full MERVLs at P6 follicular oocytes (Figure 3.6 bottom, and Figure 3.7, bottom). While there is known transcriptional de-repression between the follicular stage of oocyte development and the final arrest at the MII phase (Schultz et al., 2018), this contradicts prior work observing that MERVL expression is absent in developing and mature oocytes (Peaston et al., 2004; Yu et al., 2020). To explore this further, the 5' capped transcriptional signal of MERVL elements on the forward and reverse DNA strands was analysed (Figure 3.8-3.9), since this allowed the possibility that in oocytes, perhaps MERVL LTRs were acting solely as enhancers, which would not have been captured by prior studies (Li et al., 2023; Peaston et al., 2004; Yu et al., 2020).

Initial analysis was on uniquely-mapping reads only, for both full MERVL elements (Figure 3.8) and solo LTRs (Figure 3.9). However, as before, with only uniquely mapping reads retained, little-to-no transcriptional signal is seen (as expected since each panel is only half of this subset of reads). Examination of the single-best alignment is more illuminating; primarily plus strand reads mainly align to full MERVL elements and solo MT2_Mm LTRs on the plus strand of DNA, and vice; versa for minus strand reads. However, an unexpected faint but non-negligible transcriptional initiation signal is visible on the opposite strand to each TE, suggesting possible transcription in the opposite direction of enhancer RNAs

(eRNA), still biased towards the sense strand (summarised by Sartorelli and Lauberth, 2020).

In summary, this analysis recapitulates known transcriptional and promoter patterns of MERVL elements (both full MERVLs and solo LTRs) in early embryonic development, specifically, in 2C embryos and 2CLCs. I confirm that MERVLs are expressed in 2C embryos and 2CLCs, that the LTRs of MERVL elements exhibit promoter activity, and that read-through signal initiates in MERVL elements and transcribes into downstream genes, in 2CLCs and 2C embryos, though the extent varies between datasets. I find very faint transcriptional signal initiating in MERVL elements' LTRs in MII and follicular oocytes contrary to the literature. Moreover, there is a pale but still visible signal in the opposite direction of the MERVL elements. This raises the possibility that in follicular and mature oocytes, MERVL promotes are both expressed and act also as enhancers, which warrants further investigation falling beyond the scope of this project.

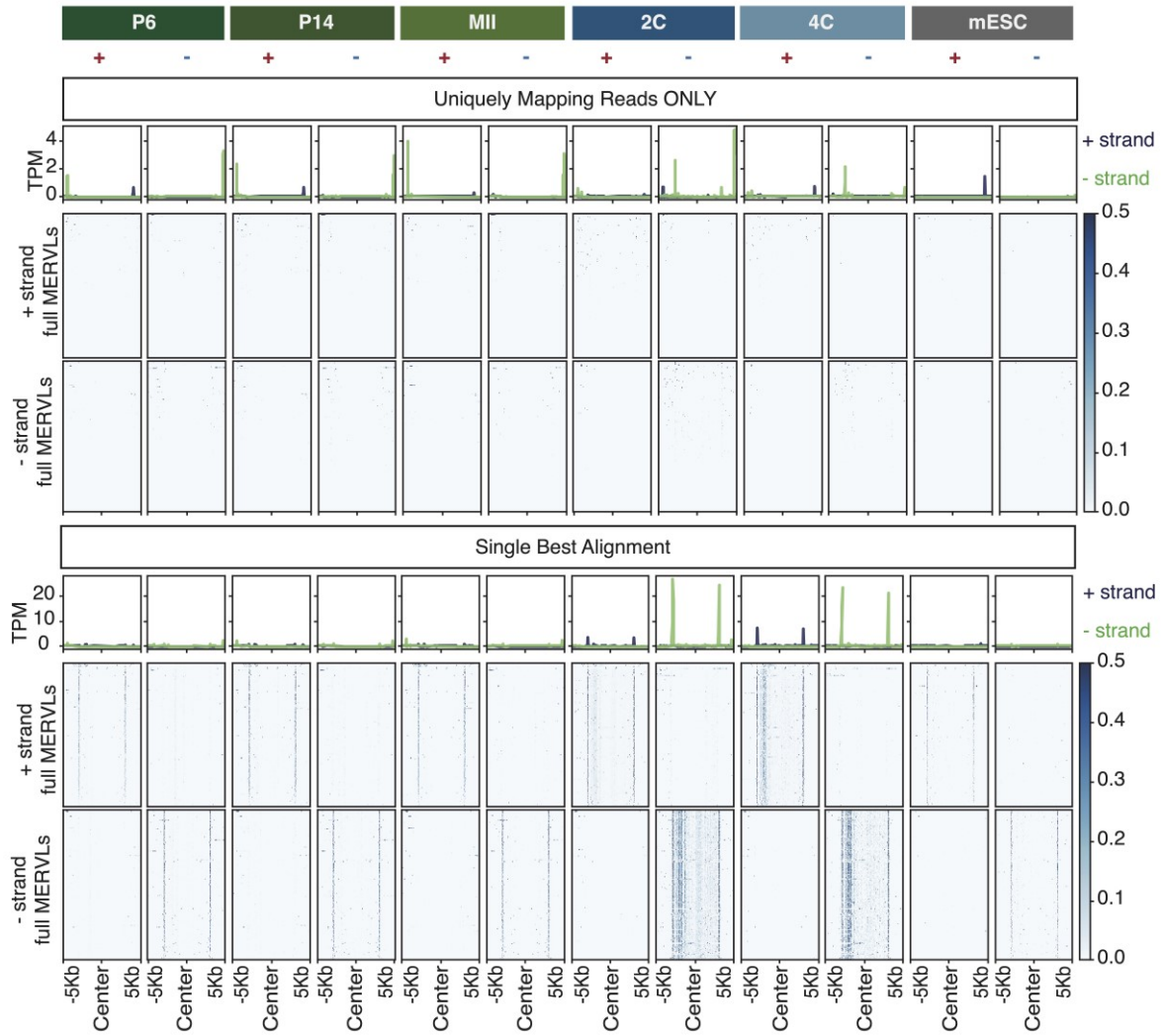


Figure 3.8: Strand-based examination of transcription initiation in full MERV LTRs across late oocyte maturation and early embryogenesis.

Sample indicated above each metaplot and heatmap. CAGE-seq signal of full MERV elements in average TPM (mean). Top panel shows uniquely mapping reads only, and bottom shows single best alignment for all reads. Plus-strand MERV element signal shown above and minus strand element signal below. Panel shows 5Kb flanking windows either side of MERV elements centre. Data analysed for P6 oocytes, P14 oocytes, MII oocytes, 2C stage embryos, 4C stage embryos, and mESCs. Data analysed from Cvetesic et al., (2018) and (2020).

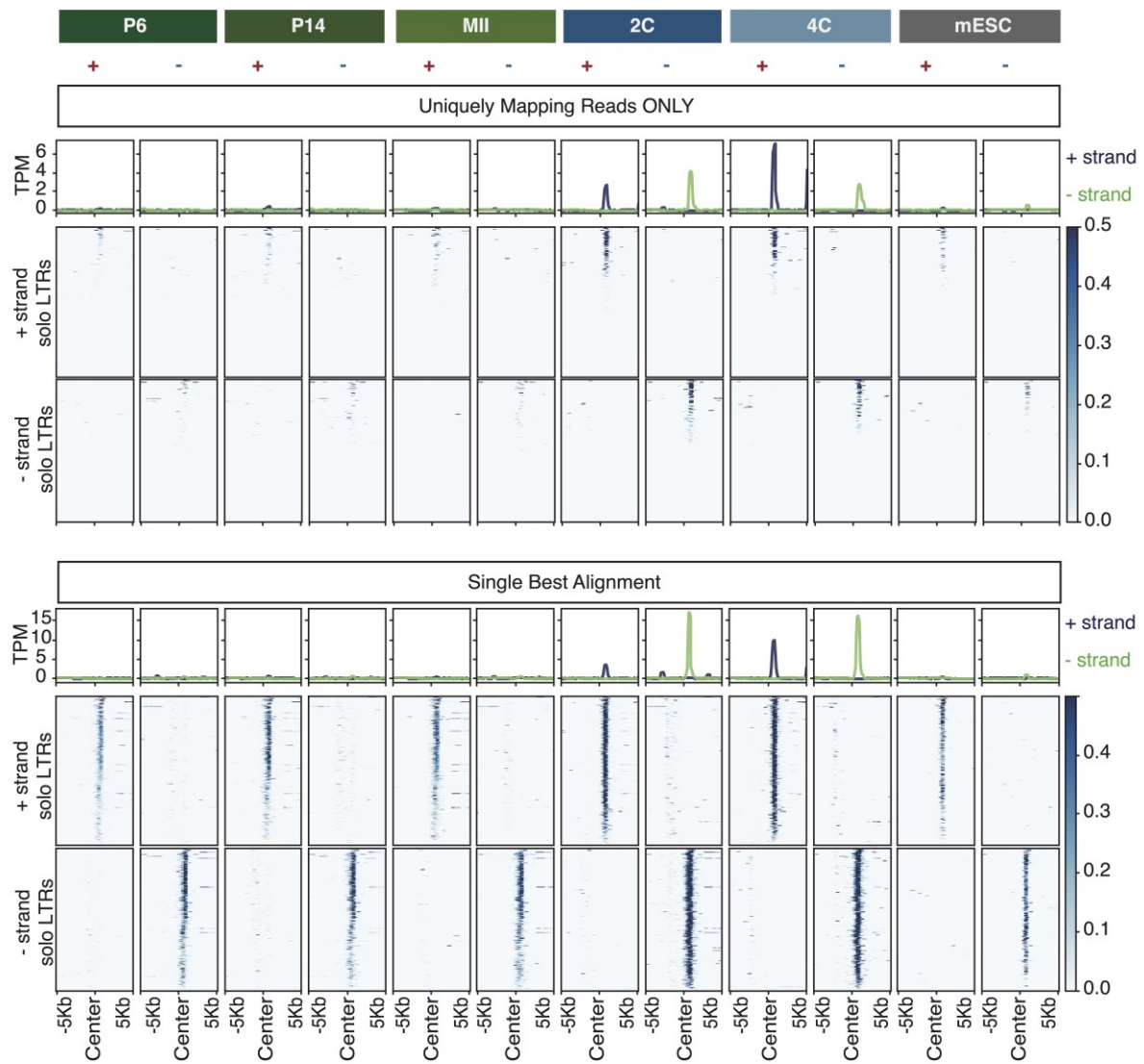


Figure 3.9: Strand-based examination of transcription initiation in solo MERVL LTRs across late oocyte maturation and early embryogenesis.

Sample indicated above each metaplot and heatmap. CAGE-seq signal of solo LTR elements in average TPM (mean). Top panel shows uniquely mapping reads only, and bottom shows single best alignment for all reads. Plus-strand MERVL element signal shown above and minus strand element signal below. Panel shows 5Kb flanking windows either side of MERVL elements centre. Data analysed for P6 oocytes, P14 oocytes, MII oocytes, 2C stage embryos, 4C stage embryos, and mESCs. Data analysed from Cvetesic et al., (2018) and (2020).

3.2 Examining the impact of MERVL promoters on the 2C transcriptome.

Several studies have confirmed the presence of MERVL-promoted transcript isoforms. Peaston et al., utilised expressed sequence tags and the BLAST database in 2004 to validate several retroelement-promoted isoforms in mouse embryogenesis. More recently, the initial 2CLC study (Macfarlan et al., 2012) examined MERVL alternative promoters extracting spliced reads between all MERVL elements in the C57Bl6/J mouse genome. These genes have been examined further by other studies (Honda et al., 2024; Sakashita et al., 2023).

The most comprehensive study to date, utilising more sophisticated deeply-sequenced short-read RNAseq, was by Modzelewski et al., in 2021. Here, they assemble several transcriptomes of several mouse preimplantation development stages to extract novel MERVL-promoted transcripts. Combining multiple assembled transcriptomes to find shared MERVL-promoted transcript isoforms and the genes to which they belong (their ‘parent’ genes) poses the challenge of constructing *identical* isoforms between publicly available datasets with batch effects. In their study, this is overcome by examining shared genes first, then extracting transcript structures from one dataset only.

Moreover, while this is a fairly extensive analysis, transcripts can be assembled with more certainty with the additional use of long-read RNA-seq data. In my project, I combine both deeply sequenced short-read data and comprehensive long-read data, to first extract high-confidence ‘parent’ genes, and extract the

transcript isoform structures using a merged novel transcriptome of the long-read datasets.

3.2.1 Identifying MERVL-promoted transcripts.

Having observed transcription proceeding past the end of the MERVL elements, I sought to assess the instances where this downstream transcription reads through into genes, to define the MERVL-promoted gene chimaeras generated which resulted in novel transcripts. To address this, transcriptomes for 2C embryos (long- and short-read) and 2CLCs and mESCs were assembled with Stringtie (Pertea et al., 2015) and transcripts were annotated according to gene and isoform type with GFFcompare (Pertea et al., 2020). To assess MERVL-promoted transcripts, the first 10 bp of each transcript's first exon was extracted and overlapped with the complete MERVLs and solo LTRs present in both the C56Bl6/J genome and the 129p2/Ola genomes on a per-dataset basis (Figure 3.10). Mono-exonic transcripts, transcripts not in the sense direction of the MERVL promoter, and transcripts with inaccurate gene assignments were removed from further analysis (see methods). Transcripts were then categorised into reference-matching non-MERVL promoted, novel non-MERVL-promoted, reference-matching MERVL-promoted (i.e., cases of exonised MERVL elements), and novel MERVL-promoted transcripts.

Initial comparisons of these transcripts between datasets show comparable proportions of reference transcripts expressed and assembled across all 2CLC and 2C embryo conditions, and between the two mESC conditions also (Figure 3.11). The number of novel transcripts is notably higher in 2C embryos and 2CLCs than it is in mESCs. This can be attributable to the process of ZGA, and in 2CLCs, this reflects the model's ability to recapitulate the transcriptional state of 2C embryos (with the caveat that 2CLCs are already transcriptionally active). The number of novel

transcripts is variable even within the 2C and 2CLC conditions (ranging between 11,372 to 119,986), corroborating previous studies exploring increased complexity of transcriptional output during early embryogenesis (Abe et al., 2015; Revil et al., 2010; Tian et al., 2021; Wyatt et al., 2022), and highlighting batch effects between studies.

The numbers of novel, MERVL-promoted transcripts range from 1,062 to 2,790 in total, across all 2C embryo and 2CLC datasets. This further encapsulates the increase in transcriptional complexity at this developmental time point, and that 2CLCs recapitulate this (as prior studies have confirmed). A minimal number of MERVL-promoted transcripts are noted in the mESC conditions (14 and 13 in the control mESC of the *Dux* overexpression and the *Caf-1* knockdown study, respectively), in line with the previous observation of minor MERVL expression in *Dux* overexpression control mESCs (Figure 3.4).

Finally, I filtered out transcripts with class codes designating “inappropriate” transcripts (i.e., those predicted by GFFcompare to be mono-exonic, PCR duplicates or artefacts, undefinable transcripts, transposon-only reads, and mapping errors), resulting in a total of 274-652 transcripts per 2C embryo/2CLC dataset. 27 and 0 transcripts were retained in the mESC control samples (*Dux* overexpression and *Caf-1* knockdown, respectively) (Figure 3.11). These came from 145-218 genes (Figure 3.12).

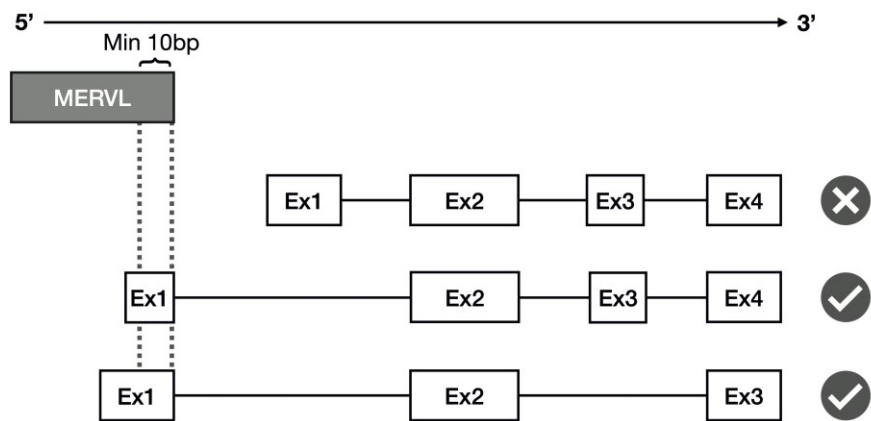


Figure 3.10. Schematic outlining definition of MERVL-promoted transcripts. If the first exon of a transcript overlaps a MERVL element by at least 10bp then it is considered to be promoted by the MERVL element.

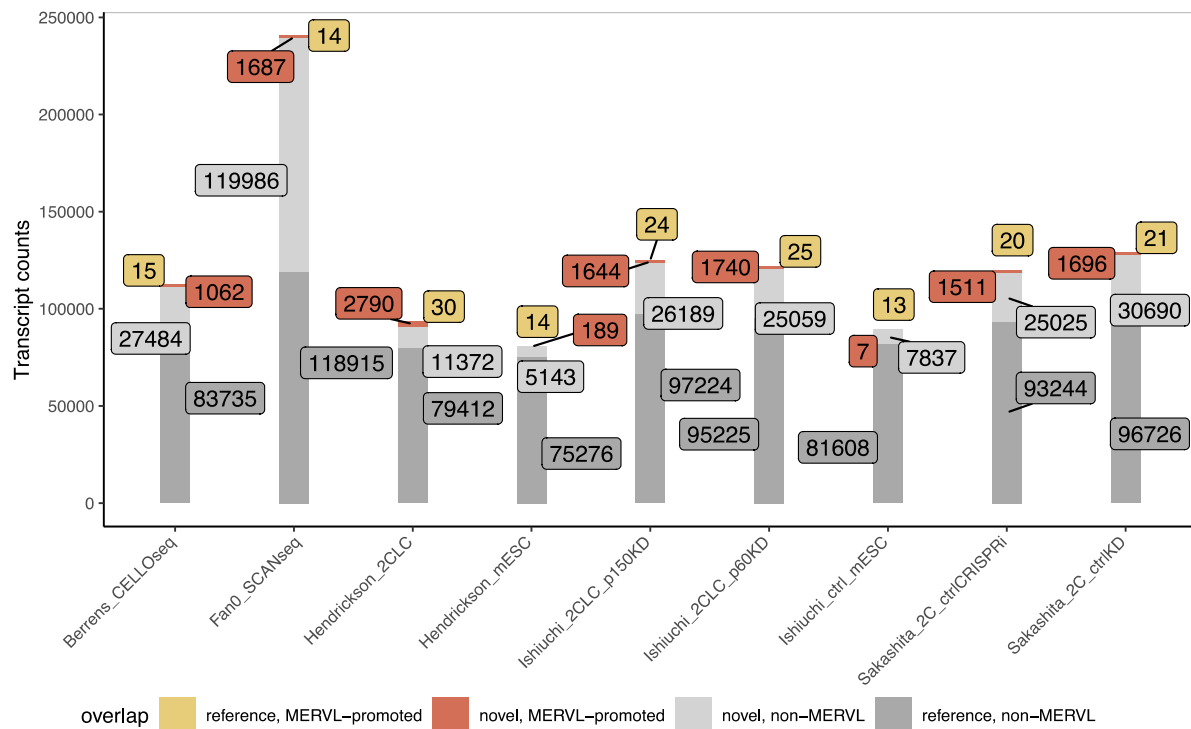


Figure 3.11. Breakdown of transcript type assembled per dataset.

Reference non-MERVL-promoted transcripts, novel non-MERVL-promoted transcripts, reference MERVL-promoted transcripts, and novel MERVL-promoted transcripts categorised. Data analysed from Berrens et al., (2020), Fan et al., (2020), Hendrickson et al., (2017), Ishiuchi et al., (2015), and Sakashita et al., (2023).

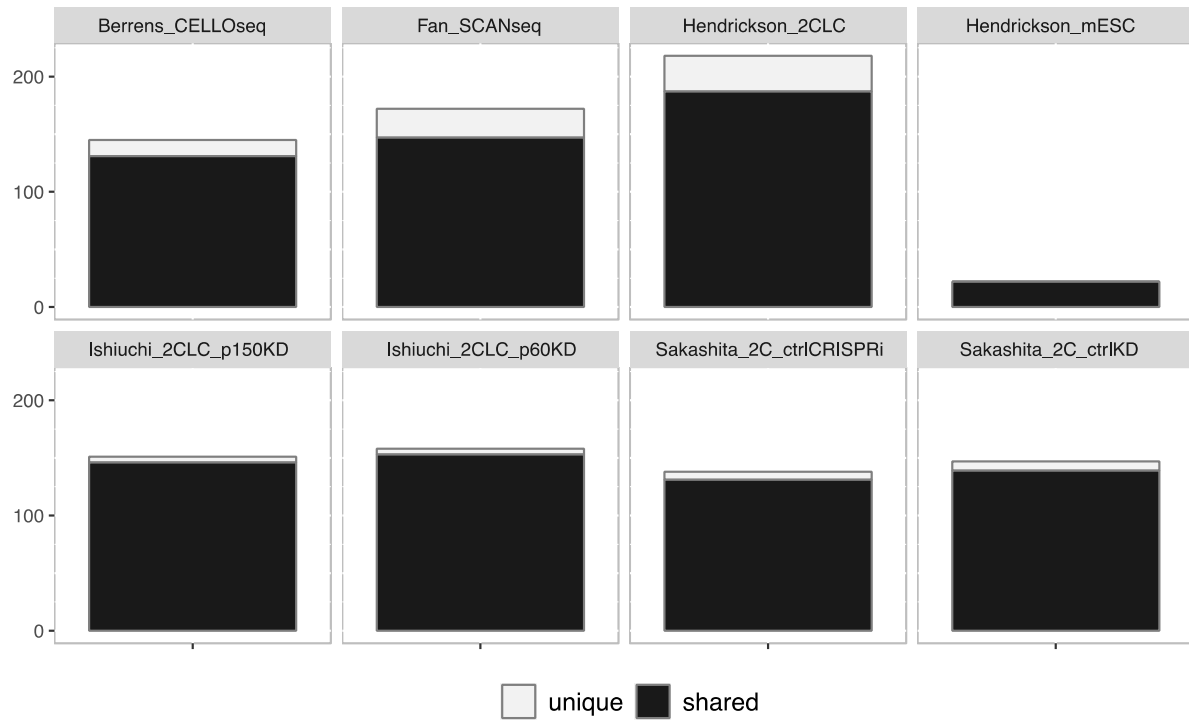


Figure 3.12. Number genes (parent genes) with MERVL-promoted transcripts per dataset unique to each dataset and shared between datasets. Data analysed from Berrens et al., (2020), Fan et al., (2020), Hendrickson et al., (2017), Ishiuchi et al., (2015), and Sakashita et al., (2023).

Class	Berrens	Fan	Hendrickson	Hendrickson	Ishiuchi	Ishiuchi	Ishiuchi	Sakashita 2C	Sakashita
code	CELLOseq	SCANseq	2CLC	mESC	2CLC	2CLC	mESC	ctrlCRISPRi	2C ctrlKD
					p150KD	p60KD			
=	140152	139737	140258	140293	140236	140314	140351	140144	140121
c	341	342	318	318	320	231	215	342	342
e	1	3	0	0	1	0	0	0	0
i	3290	35364	2811	632	9699	9225	1832	4571	6065
j	14551	28891	2599	2046	3031	2863	1729	8633	10302
k	1471	2384	1131	940	956	975	895	1859	1842
m	568	382	154	320	425	344	1146	166	196
n	665	931	101	121	339	313	996	148	158
o	545	2206	308	105	229	202	80	574	605

p	291	3565	161	57	1068	1080	251	246	373
s	97	207	3	5	1	0	1	68	85
u	6042	42578	6120	1407	11507	11222	1276	8565	11083
x	1514	5619	1176	176	965	957	153	2065	2120
y	144	569	137	29	170	186	35	267	205

Table 4: breakdown of GFFcompare class codes assigned to all assembled transcripts relative to mm39 NCBI RefSeq genome assembly. Data analysed from Berrens et al., (2020), Fan et al., (2020), Hendrickson et al., (2017), Ishiuchi et al., (2015), and Sakashita et al., (2023).

3.2.2 Constructing a list of high-confidence ‘parent’ genes.

Having confirmed the presence of novel, MERVL-promoted transcripts assembled in the transcriptomes of the 2C stage, I sought an understanding of these isoforms relative to reference isoforms. The Hendrickson (2017) short-read 2CLC dataset (*Dux* overexpression 2CLCs) and the Fan et al., (2020) long-read 2C embryo dataset (SCAN-seq) have the highest number of genes with MERVL-promoted transcripts that are unshared with any other dataset (Figure 3.12); this is in line with my previous analysis showing that these are the datasets with the most novel MERVL-promoted transcripts (Figure 3.11).

Thus, a list of high-confidence genes and isoforms was required, which I constructed by extracting the genes with MERVL-promoted isoforms per dataset, overlapping the common genes, and selecting those genes present in either or both of the two long-read RNA-seq datasets, (Berrens et al., 2021; Fan et al., 2020), and one or more of the short-read RNA-seq datasets (Hendrickson et al., 2017; Ishiuchi et al., 2015; Sakashita et al., 2023) (Appendix 2, figure 1). Hereafter, these genes are referred to as parent genes. The two long-read datasets were then filtered for these 172 genes, and the output was filtered GTFs merged using Stringtie in merge mode, resulting in a list of 576 transcripts. This approach has two benefits; first, the discrepancies between the two long-read datasets are offset by combining the parent genes and selecting those MERVL-promoted isoforms remaining post-merging retaining only those with strong enough coverage. Secondly, it can best harness the benefits of long-read sequencing in defining exact and reliable mRNA transcript structures while using deeply sequenced short-read RNA-seq data to support these transcripts. To retrieve the final lists of transcripts and genes, the two long-read total transcriptomes (CELLOseq by Berrens et al., 2021 and SCANseq by Fan et al., 2020)

were merged using Stringtie in merge mode, and annotated with GFFcompare using the list of merged list of MERVL-promoted transcripts from the prior analysis. This outputs 167 genes, with 549 transcripts. The differing 5 genes (*Apol7b*, *Diaph3*, *Gm6916*, *Iars*, and *Trak2*) and their associated transcripts were lost during the merge since the MERVL-promoted isoforms for these were no longer abundant enough (Appendix 2, figure 2).

In summary, this analysis yields 167 high-confidence parent genes of 549 MERVL-promoted transcripts for use in further analysis of expression and exploration of protein domain changes.

3.2.3 Comparison of parent genes retrieved in this study to prior works.

Previous studies have attempted to construct similar lists of MERVL-promoted isoforms; Macfarlan et al., (2012) and Maksakova et al., (2013) used RNA-seq data sequenced from 2C stage embryos and 2CLCs in their respective studies, and Modzelewski et al., (2021), combined several published short-read RNA-seq studies. Modzelewski et al., and Macfarlan et al., find 307, and 306 genes with MERVL-promoted transcripts respectively; however, the majority of these are driven by MERVL LTRs which are not MT2_Mms (i.e., MT2A/MT2B/MT2C); only 57 and 36 genes respectively from these studies are MT2_Mm promoted. Maksakova et al., find 42 genes, and it is not confirmed whether these initiate in MT2_Mm MERVLs or all MERVL LTRs (i.e., including MT2A/B/C LTRs).

A comparison of these published gene lists to the 167 parent genes detected in this project shows that the majority of the genes are unique to each study (153 genes in this study, 25 genes from the Macfarlan list, 39 from the Maksakova list, and 16 genes from the Modzelewski list) . Only 2 genes are shared in this study, Macfarlan et al., and Modzelewski et al.,; *Ddit4l* (Appendix 3, figure 1) which is involved in the

DNA damage response, and ribosomal stalk protein gene *Rplpo* (Appendix 3, figure 2). Notably, *Rplpo* is considered relatively stable and often used for transcript and protein normalisation (Aw et al., 2020; Biasini and Marques, 2020; Garland et al., 2022). With this study, Maksakova et al., and Modzelewski et al., *Ppig* is shared (Appendix 3, figure 3). *Ppig* is Peptidyl-prolyl cis-trans isomerase G and is linked to protein folding and pre-mRNA splicing (via similarity in Uniprot, (The Uniprot Consortium, 2019)). Between this study, Macfarlan et al., and Maksakova et al., only *Pemt* is shared (Appendix 3, figure 4), regulating phosphatidylcholine synthesis, and with links to the alternative splicing of homeostasis proteins (Johnson et al., 2020; Zhou et al., 2022).

Contrary to these published works (in particular, Macfarlan et al., (2012)), no GO terms were enriched in this study's list of 167 parent genes, as compared to a background of all genes in the reference genome, or genes expressed either mESCs only or 2C embryos/2CLCs.

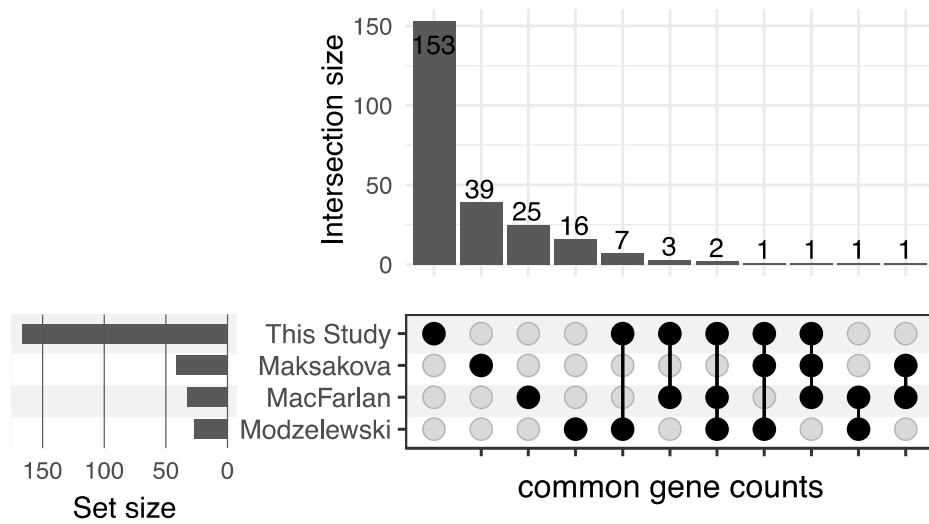


Figure 3.13: Comparison of high-confidence parent genes in this study to other similar works. Data analysed from Maksakova et al., (2013), Macfarlan et al., (2012), and Modzelewski et al. (2021).

3.2.4 Characterising MERVL-promoted transcripts in the context of the 2C transcriptome.

The initial model for chimeric transcript assembly based on prior studies included an intergenic, upstream MERVL element splicing to the 2nd exon of a gene, to result in an alternative transcript (as shown by Honda et al., 2023; van de Lagemaat et al., 2003; Modzelewski et al., 2021; Peaston et al., 2004) (Figure 3.14, top). However, assessing the locations of the MERVL elements relative to the 169 high-confidence genes shows that 76 of 167 MERVL-promoted transcripts result from intronic MERVLs acting as alternative promoters (Figure 3.15, left), supplementing the original model (Figure 3.14, bottom). Since MERVL elements lack splice acceptor sites, these cannot be read-through transcription; additionally, the analysis in subchapter 3.1 shows a null transcriptional signal upstream of MERVL elements reading into the MERVL bodies.

Analysis of the parent genes of MERVL-promoted transcripts showed that 60/167 are unconfirmed genes (i.e., genes with LOC, Rik, or GM-based gene symbols) (Figure 3.15, right). Examination of the raw numbers of isoforms per gene shows that reference transcripts outnumber MERVL-promoted transcripts on average by a magnitude of 3 (Figure 3.16), in line with previous results showing that more reference isoforms are retrieved in 2C embryo and 2CLCs than novel isoforms on the whole (Figure 3.11).

Finally, I reviewed exon lengths and intron lengths for MERVL-promoted transcripts, relative to their reference counterparts. The average exon length is much smaller than intron lengths (Figure 3.17). This further supports a previous result, of speckly transcriptional signal visible in long-read RNAseq data (Figure 3.5). The

signal is only present at maturely spliced regions mapped back to the genome, explaining the gaps in signal as introns. In the Caf-1 knockdown 2CLCs (Figure 3.4), the expression pattern downstream appears more shaded than punctate, in line with the rRNA library preparation method which can select for pre-mRNAs also (table 3).

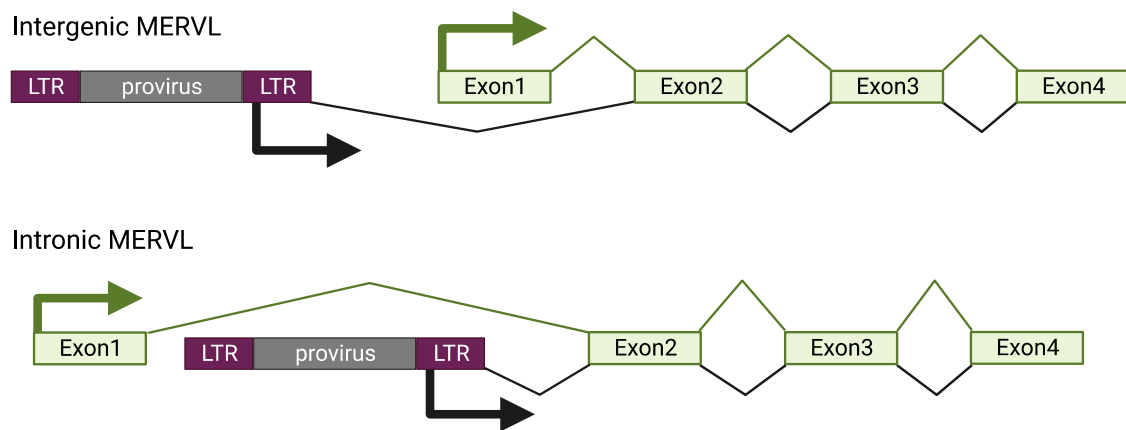


Figure 3.14: Working model of MERVL-promotion of transcripts.

MERVL alternative promoters result in alternative isoforms, either by having the MERVL alternative promoter upstream of the gene (top) or inserted in a gene intron (bottom). MERVL elements lack splice acceptor sites and so they cannot be spliced to, thus they cannot be exonised without mutation.

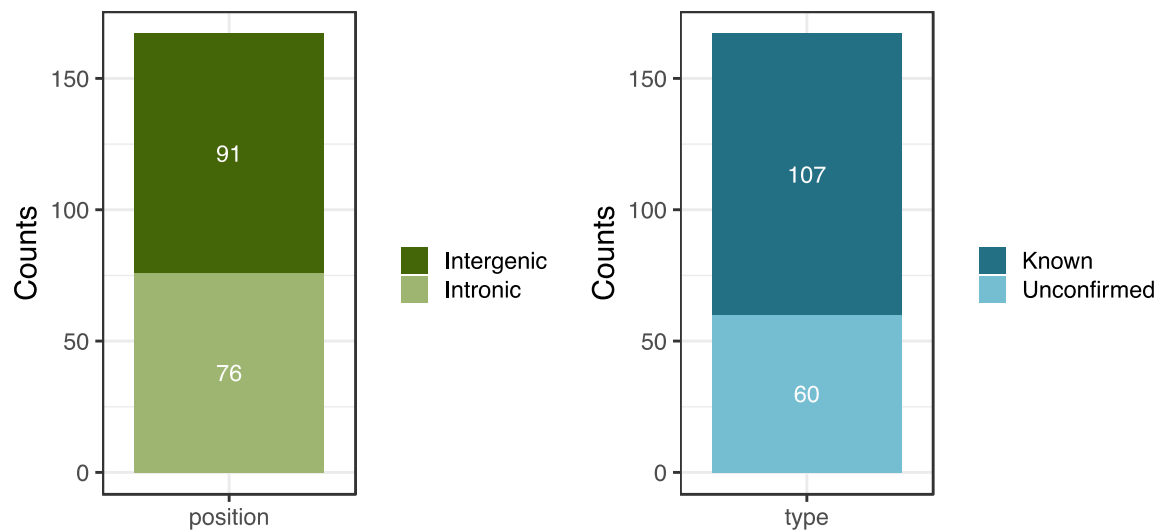


Figure 3.15: MERVL elements can promote transcription from introns, for confirmed and unconfirmed genes. The number of parent genes with intronic versus intergenic MERVLs driving alternative transcripts, and number of parent genes with known/assigned gene product, and genes with no assigned symbol (i.e., Gm, Loc, Rik, or pseudogenes).

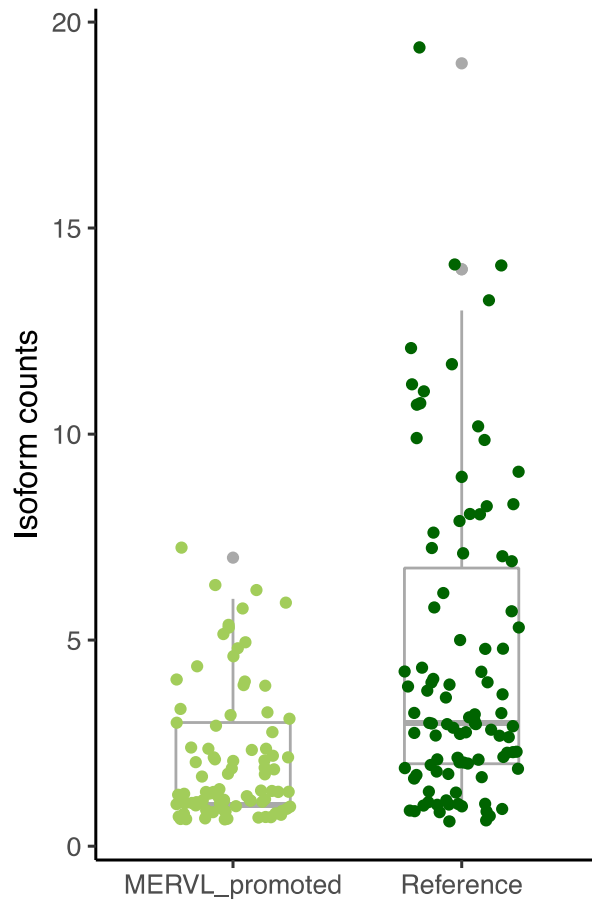


Figure 3.16: Reference transcripts outnumber MERVL-promoted transcripts. Distribution of MERVL-promoted and reference transcripts predicted per parent gene. Wilcoxon statistic $W = 2281$, $p\text{-value} = 1.215\text{e-}08$.

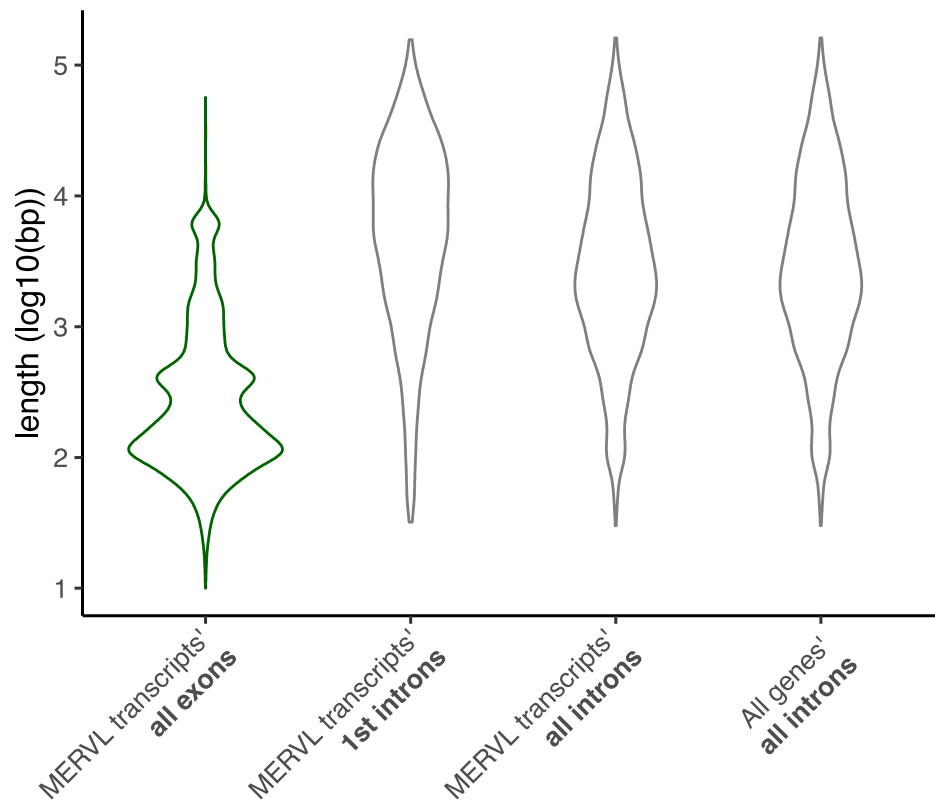


Figure 3.17: Exon lengths are shorter than intron lengths. Distributions of exon and intron lengths of MERVL-promoted transcripts and reference transcripts, log2 transformed.

3.3 Cellular dynamics of MERVL-promoted isoforms.

The peak of MERVL expression is from the S-phase of the 1-cell stage (i.e., minor ZGA, (Kigami et al., 2003) through the major ZGA wave, to the end of the 2-cell stage (Evsikov et al., 2004; Honda et al., 2024; Macfarlan et al., 2012; Modzelewski et al., 2022; Peaston et al., 2004). Whether this extends to chimeric transcripts initiating in MERVL elements has only been assessed superficially (Macfarlan et al., 2012; Sakashita et al., 2023). Conversely, there is mounting evidence suggesting that non-canonical MERVL elements (i.e., MTA/B/C ERV elements and other MT2 MERVL elements) are expressed from oocyte maturation and onwards until the blastocyst stage (Franke et al., 2017; Modzelewski et al., 2021; Sakashita et al., 2023), as are the chimeric transcripts they may promote. While it is unexpected that the chimeric transcripts can be expressed when the promoting element for these is observed to be inactive (such as in the developing oocyte), examination of specifically MT2_Mm-promoted transcripts throughout preimplantation development is as yet uncompleted, and thus still required.

Whether these MERVL-promoted transcripts have any impact on early embryogenesis depends on whether these transcripts are translated, and whether the protein produced undergoes any functional changes. To this end, the coding potential of MERVL-promoted transcripts is required, to examine whether the transcripts produced can be translated at all. Subsequently, extraction of possible protein domain sequence homologies is necessary, to understand the changes to a protein isoform relative to the reference protein isoform, when an alternative promoter drives the transcript isoform.

3.3.1 Expression patterns of MERVL-promoted transcripts using long-read data.

First, I examined the expression patterns of the novel transcripts using long-read data (Fan et al., 2020). cDNA-PCR long-read RNAseq data produced with the SCAN-seq library preparation method for preimplantation development was aligned using Minimap2 (Li, 2018) and quantified using Salmon (Patro et al., 2017) in alignment mode with nanopore-specific options, at the transcript level. Expression analysis was completed using DESeq2 (Love et al., 2014).

Principal component analysis (Figure 3.18) of the expression of all transcripts in the novel transcriptome shows a horseshoe clustering of cell types (MII oocytes, 1-cell zygotes, early and late 2-cell stage embryos, 4-cell embryos, 8-cell embryos, 16-cell morula, early blastocysts' trophoctoderm and inner cell mass cells, and late blastocysts' trophoctoderm and inner cell mass cells, as well as mESCs), as expected for a developmental time course. In general, the clustering pattern is consistent with the original study, with most cells clustering by type, and the biggest difference in cell types is before and after zygotic genome activation initiates (between the 1-cell zygote and the early 2-cell stage cells). However, unlike the original work, the early and late 2-cell stage cells (n=13 and n=18) and the 4-cell stage cells (n=13) disperse a little more across PC1; PC1 appears to describe the developmental progression from oocyte to late blastocyst, which suggests this result shows difference even within cell types for cell age. As expected, mESCs are the most similar to the late blastocyst cells.

Of the 549 MERVL-promoted transcripts, 47 have null expression in any cell type and, thus, are lost when row normalising expression. These belong to 33 genes (*4930578G10Rik*, *Abcb5*, *Antxr1*, *Aqr*, *Ccser1*, *Csnk2a1*, *D17Ertd648e*, *Dcbld1*, *Dhcr7*, *Echdc3*, *Gm19792*, *Gm30220*, *Gm31624*, *Gm35689*, *Gm4011*, *Gm40910*,

Gm40954, Gm46534, Gm5577, LOC118567331, LOC118568063, LOC118568292, Mob3b, Nat8f2, Olfr881, Pemt, Ric3, Sgk3, Tent2, Ube2w, Vmn1r227, Zc3h6, Zswim2), of which all but *D17Ertd648e, Gm35689, Vmn1r227, and Zc3h6* have more than 1 MERVL-promoted transcript. These lost transcripts are likely to be highly expressed in the CELLOseq dataset, which was also used to construct the novel transcriptome, or are Stringtie merging artefacts.

The remaining 502 transcripts can be k-means clustered into 4 groups to assess expression (Figure 3.19). Cluster 1 captures transcripts with expression peaking in the MII stage, 1-cell zygote stage, and some expression in the 2-cell stages (n=109). The next cluster (n=111) is of 2-cell-specific transcripts, with some latent expression present in the 48 cell stages. Cluster 3 is of transcripts with non-specific transcription across the 2-cell stage to the late blastocyst (n=95). Cluster 4 is the largest, but capturing transcripts with inconsistent signal within cell types, and a subset of transcripts with mESC-specific expression. Cluster 2 exhibits the expected expression pattern, with transcripts initiating in expression with minor ZGA, and persisting until the 4-8 cell stage.

MERVL elements peak in expression after minor ZGA, and prior work finds no expression of MT2 elements in maturing oocytes, only MTA/B/C elements (Peaston et al., 2004; Yu et al., 2020). In this study, a faint MERVL element transcriptional signal is captured with CAGE-seq (Figure 3.6-3.7, and 3.8-3.9) during late oocyte maturation (P6 and P14 stages), and in MII oocytes. In theory, these transcripts could all have belonged to one gene and been transcript assembly artefacts by Stringtie (as was evidenced in section 3.2.1) since the MERVL promoter is not strongly active in oocytes. To confirm whether or not the presence of the oocyte cluster (cluster 1) was artefactual, I next examined the genes in the oocyte cluster and the 2C cluster (cluster 2). 109 transcripts were in the oocyte cluster, with 111 in the

2C cluster (Figure 3.20); while none of the transcripts is shared, these come from 49 genes and 41 genes respectively, 14 of which overlap between clusters (Figure 3.21). Comparing the transcripts per the 14 shared genes, almost all genes have a maximum of 5 transcripts per gene EXCEPT *Gm40910*, which has 10 transcripts in the oocyte cluster. Despite this outlier, no genes have so many transcripts that they could be skewing the oocyte cluster, so this is not a technical artefact. (Figure 3.22). I next examined the expression of the 167 parent genes of MERVL-promoted transcripts, by aggregating the expression of all transcripts by gene using `tximport::summarizeToGene`. All of the 167 genes are expressed in the 2-cell stages, as expected and consistent with the literature (Figure 3.23). However, a significant subset of genes does have expression in MII oocytes, as seen at the transcript level. This analysis confirms that while some genes have rather high expression at the 2C stages (Appendix 3 figure 2), they do not peak in expression at the 2C stage. Additionally, this study finds that overall, expression varies at the transcript level, and this granularity is obscured in gene-level comparison.

Finally, a comparison to the Modzelewski et al., (2021) list of MT2_Mm-promoted genes (i.e., the transcripts in this study's novel transcriptome belonging to the Modzelewski et al., list of genes) shows a similar pattern, with a large subset of transcripts with expression in oocytes (Figure 3.24). It should be noted here that *Rplpo* (Appendix 3 figure 2) is a parent gene shared between this study and the Modzelewski et al., (2021) dataset, and is therefore a key example of a MERVL-promoted gene with substantial support which is not 2C-specific. From this, I conclude that some MERVL-promoted transcripts have higher expression in oocytes than in the 2C stage, and therefore may be maternally inherited, contradictory to prior results positing that there is no MERVL activity in oocyte development (Peaston et al., 2004). However, this would require further investigation by way of

examining more transcript structures and spliced reads between MERVL elements and gene bodies, case by case, as completed for *Rplpo*. Longitudinal analyses clustering transcript types may also be enlightening.

2C-specific transcripts were then defined as having Log2 fold change (Log2FC) expression >1 over Oocytes AND Morula stage embryos; this is because MERVL expression is known to drop to null by morula stage, and this can remove any maternal transcripts fully. This results in 41 transcripts from 23 genes (*Abcb5*, *AU019990*, *Csrnp3*, *Fam172a*, *Farp1*, *Fundc1*, *Gm10865*, *Gm11784*, *Gm30220*, *Gm33923*, *Gm4011*, *Gm40910*, *Gm42808*, *Gm5309*, *Gnpnat1*, *LOC118568063*, *P4ha2*, *Plaat3*, *Ppig*, *Prex2*, *Sgk3*, *Ube2w*, *Zfp560*), 13 of which are confirmed genes, specific to the 2C stage (Oocyte<late 2C stage>morula) (Figure 3.25-3.26).

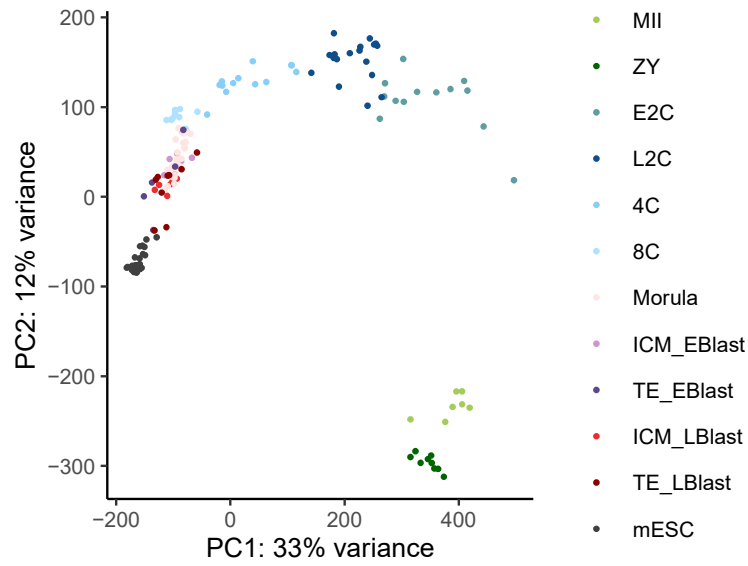


Figure 3.18: PCA plot for SCAN-seq (Fan et al., 2020) long-read RNA-seq dataset. All transcripts in the novel transcriptome assembled in 3.2.2, with expression in any cell type, were used (total 549).

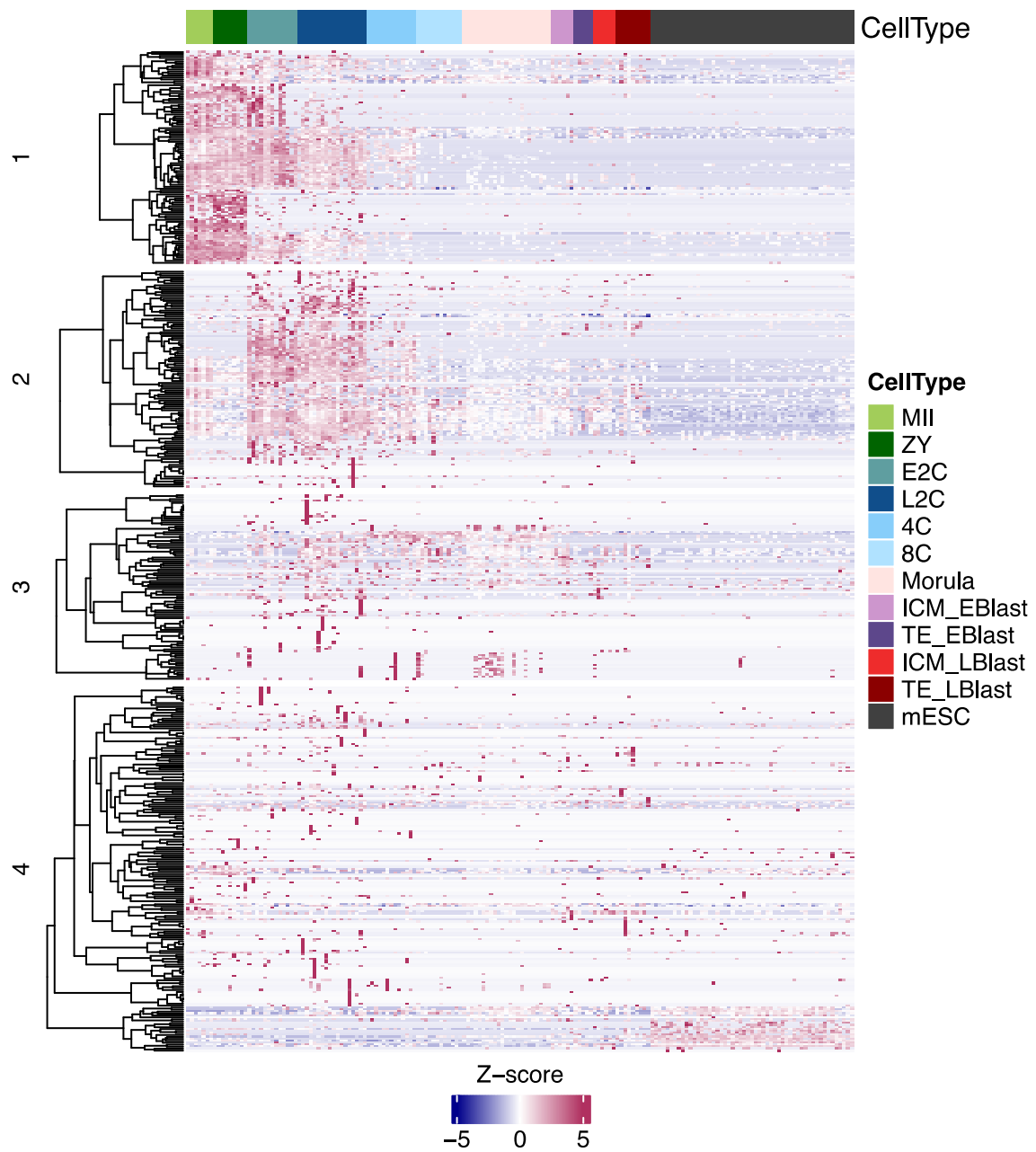


Figure 3.19: Expression analysis of MERVL-promoted transcripts across preimplantation development. Z-scored heatmap of 549 MERVL-promoted transcripts across MII oocytes (MII), 1C zygotes (ZY), 2C stage embryos (2C), 4C stage embryos (4C), 8C stage embryos (8C), 16 cell stage embryos (Morula), Early blastocyst inner cell mass (ICM_Eblast), Early blastocyst trophoctoderm (TE_Eblast), Late blastocyst inner cell mass (ICM_LBlast), Late

blastocyst trophectoderm (TE_LBlast), and mESCs. Transcripts are k-means clustered into 4 clusters.

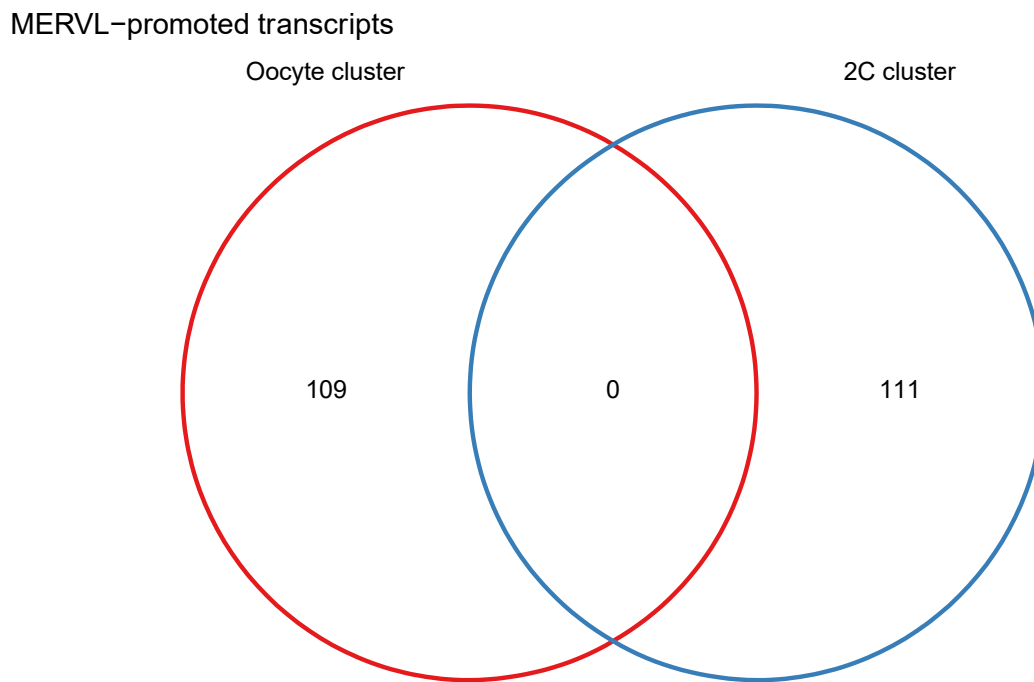


Figure 3.20. No MERVL-promoted transcripts are shared between clusters. Venn diagram comparing MERVL-promoted transcripts from Oocyte-specific cluster and 2C-specific cluster.

Genes with MERVL-promoted transcripts

Oocyte cluster

2C cluster

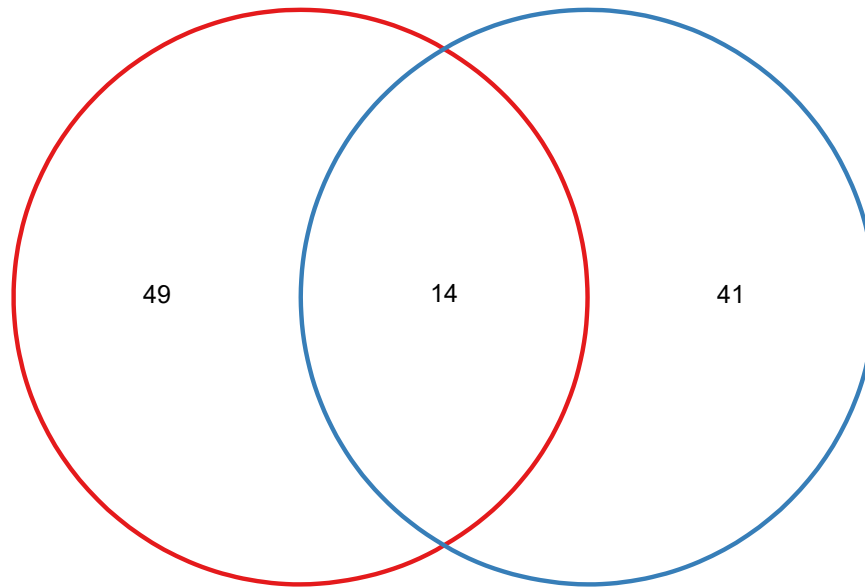


Figure 3.21: Few parent genes are shared between clusters. Venn diagram comparing parent genes of MERVL-promoted transcripts from more Oocyte-specific cluster and more 2C-specific cluster.

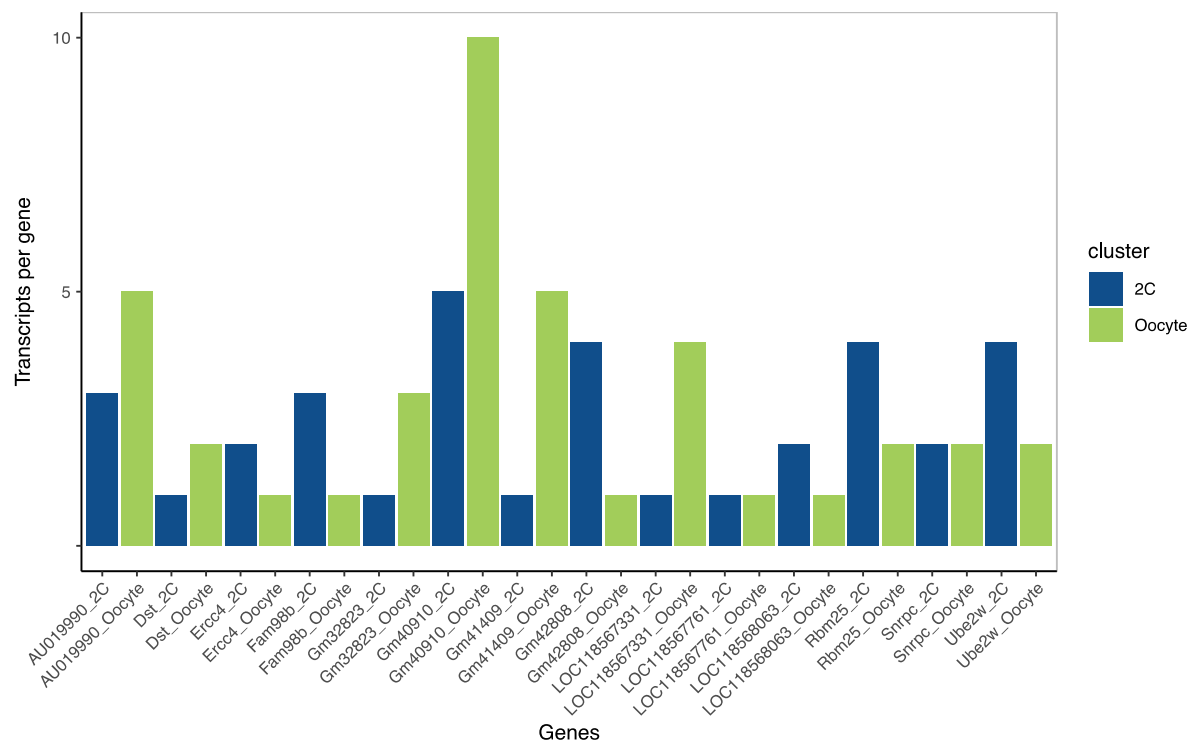


Figure 3.22: Comparison of MERVL-promoted transcript numbers from parent genes shared between the oocyte-specific cluster and the 2C specific cluster. 2C-specific MERVL-promoted transcripts are in blue, and oocyte-specific transcripts are in green.

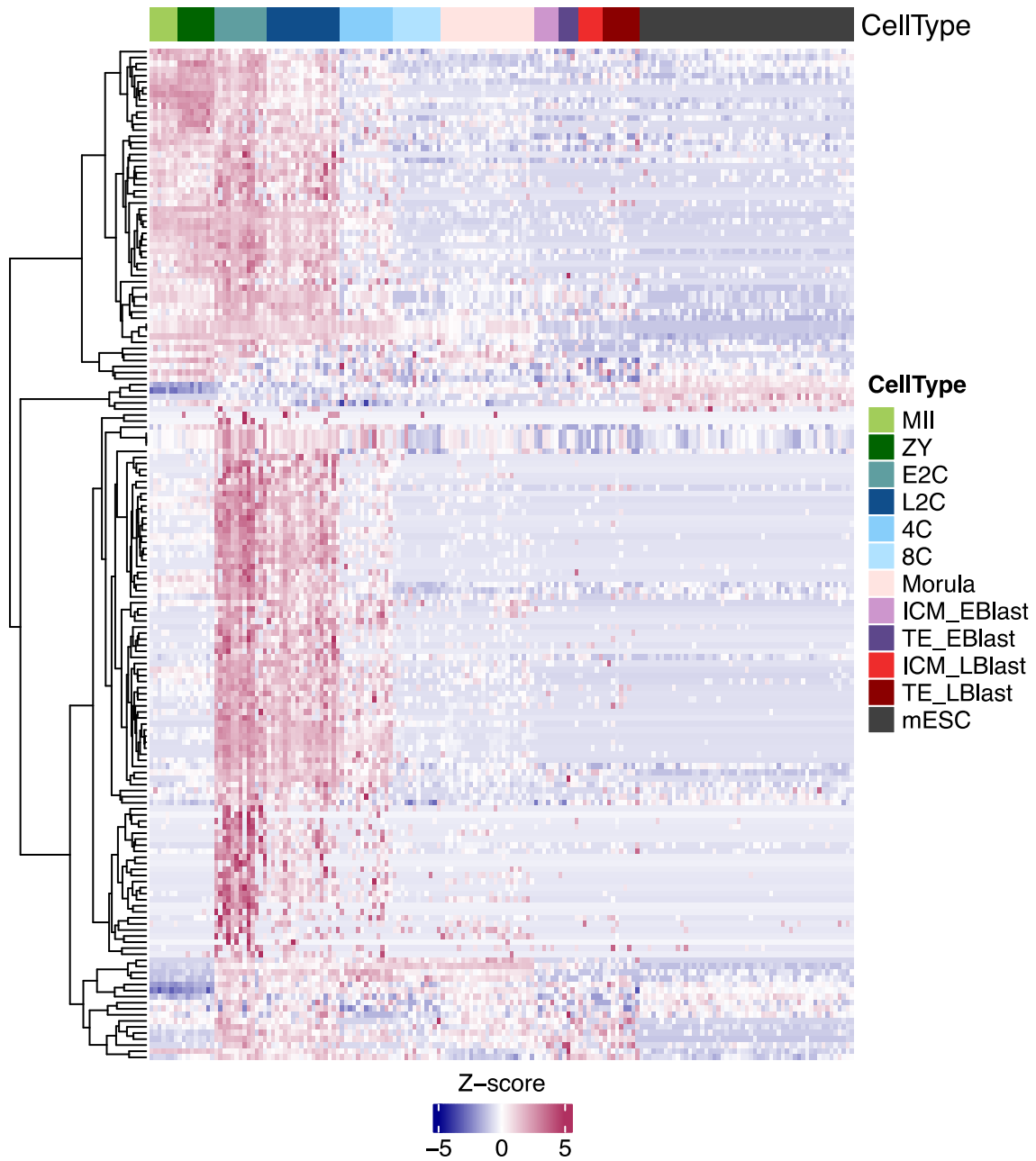


Figure 3.23: Summarised expression analysis of parent genes of MERVL-promoted transcripts across preimplantation development. Z-scored heatmap of 549 MERVL-promoted transcripts across MII oocytes (MII), 1C zygotes (ZY), 2C stage embryos (2C), 4C stage embryos (4C), 8C stage embryos (8C), 16 cell stage embryos (Morula), Early blastocyst inner cell mass (ICM_Eblast), Early blastocyst trophectoderm (TE_Eblast), Late blastocyst inner cell mass (ICM_Lblast), Late blastocyst trophectoderm (TE_Lblast), and mESCs.

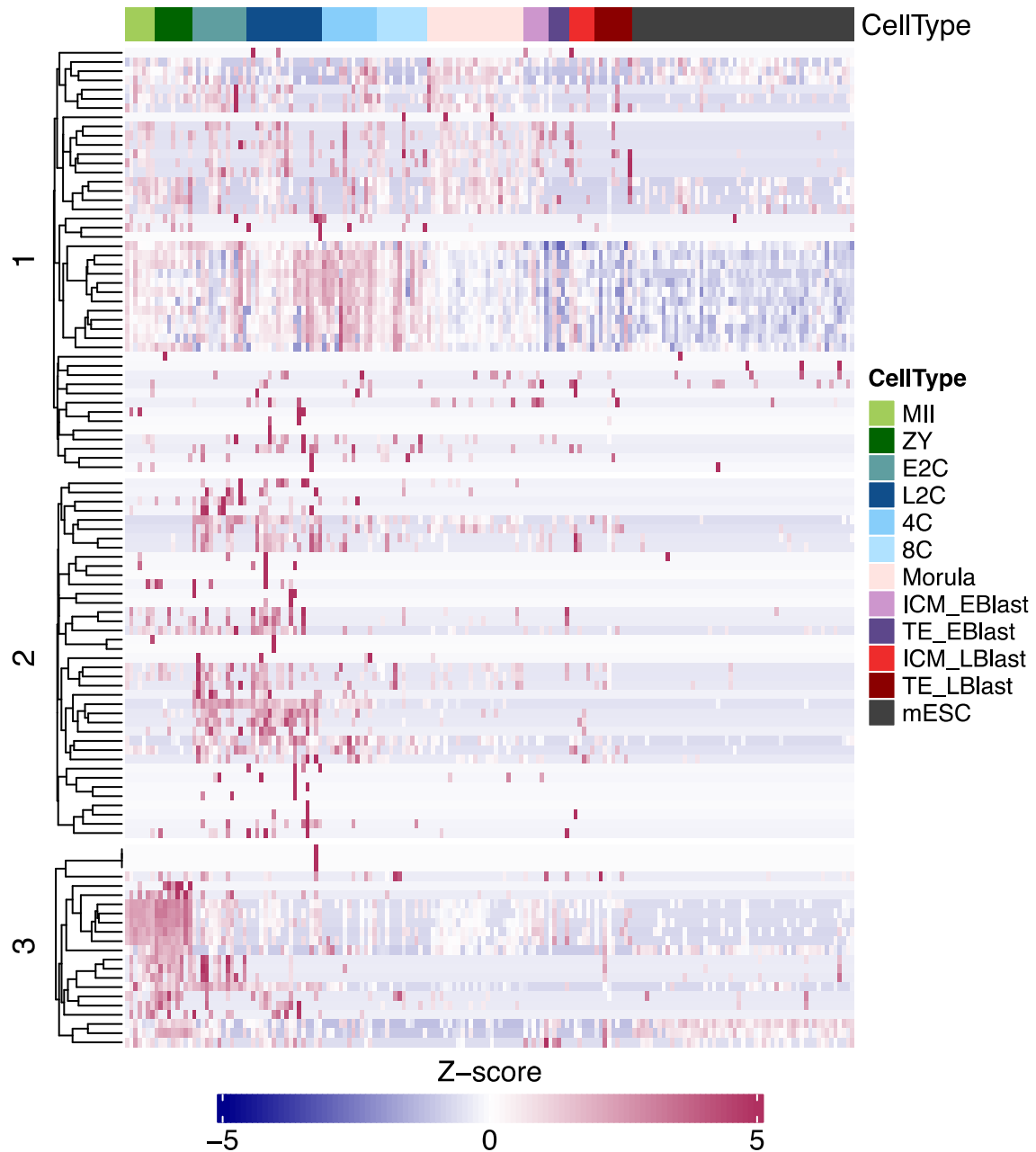


Figure 3.24: Expression analysis of transcripts assembled in this study belonging to parent genes of MERVL-promoted transcripts defined by Modzelewski et al., (2021). Z-scored heatmap of MERVL-promoted transcripts across MII oocytes (MII), 1C zygotes (ZY), 2C stage embryos (2C), 4C stage embryos (4C), 8C stage embryos (8C), 16 cell stage embryos (Morula), Early blastocyst inner cell mass (ICM_Eblast), Early blastocyst trophectoderm

(TE_Eblast), Late blastocyst inner cell mass (ICM_LBlast), Late blastocyst trophoctoderm (TE_LBlast), and mESCs.

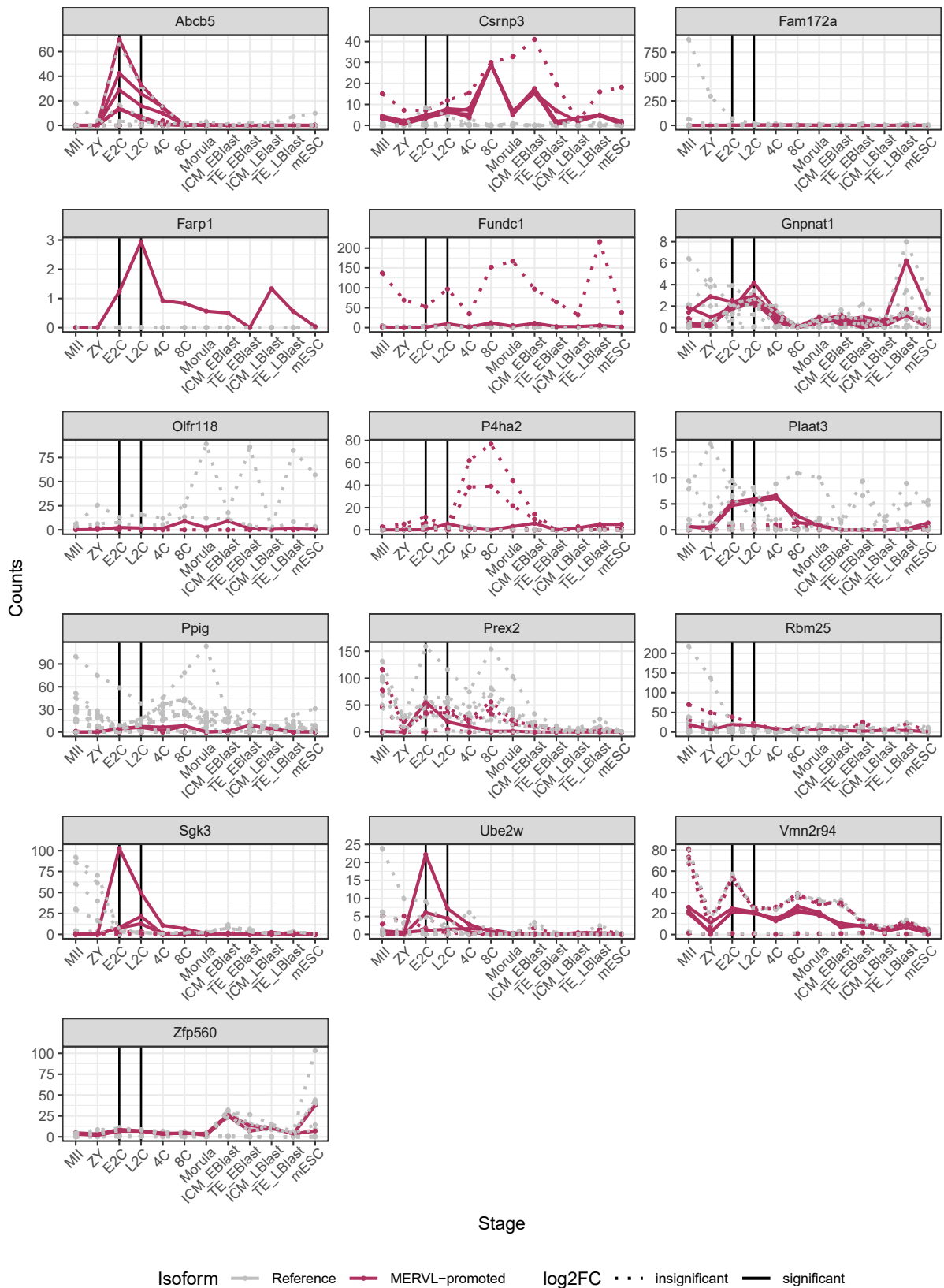


Figure 3.25: Expression patterns across preimplantation development of 2C-specific MERVL-promoted transcripts in context of their parent genes, for known genes. Transcript raw counts shown across MII oocytes (MII), 1C zygotes (ZY), 2C stage embryos (2C), 4C stage embryos (4C), 8C stage embryos (8C), 16 cell stage embryos (Morula), Early blastocyst inner cell mass (ICM_Eblast), Early blastocyst trophectoderm (TE_Eblast), Late blastocyst inner cell mass (ICM_LBlast), Late blastocyst trophectoderm (TE_LBlast), and mESCs. MERVL-promoted transcripts in magenta, with reference transcripts in grey, and 2C-significant ($\text{Log}_2\text{FC} > 1$ for 2C over Oocytes and Morulae) in solid lines and insignificant transcripts in dotted lines.

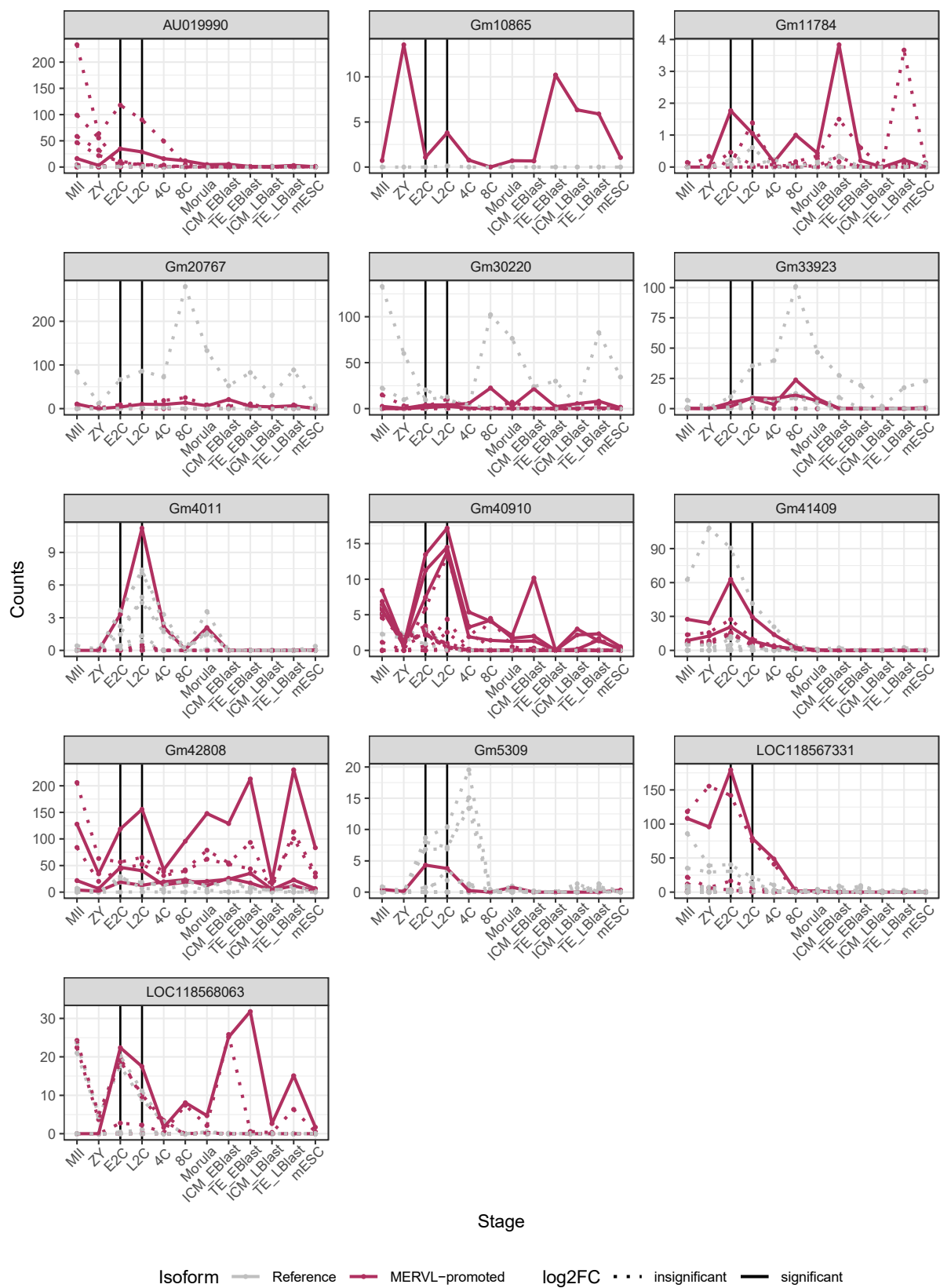


Figure 3.26: Expression patterns across preimplantation development of 2C-specific MERVL-promoted transcripts in context of their parent genes, for unconfirmed genes. Transcript raw counts shown across MII oocytes (MII), 1C zygotes (ZY), 2C stage embryos (2C), 4C stage embryos (4C), 8C stage embryos (8C), 16 cell stage embryos (Morula), Early blastocyst inner cell mass (ICM_Eblast), Early blastocyst trophectoderm (TE_Eblast), Late blastocyst inner cell mass (ICM_LBlast), Late blastocyst trophectoderm (TE_LBlast), and mESCs. MERVL-promoted transcripts in magenta, with reference transcripts in grey, and 2C-significant ($\text{Log}_2\text{FC} > 1$ for 2C over Oocytes and Morulae) in solid lines and insignificant transcripts in dotted lines.

3.3.2. MERVL-promoted transcripts are statistically less likely to be coding compared to reference transcripts.

To understand whether a gene (or novel gene) results in a functional protein, it is important to consider the likelihood of translation in conjunction with transcript expression data. In particular, this is important since one transcript may have several potential open reading frames (ORFs), which could be translated. Therefore I extracted the mRNA sequences of the 549 MERVL-promoted transcripts and calculated the coding potentials of their predicted ORFs using CPAT (Wang et al., 2013). CPAT assesses coding potential by integrating information on hexamer sequence usages in known coding and non-coding transcripts with predicted open reading frame length and coverage, to output several ORFs per transcript with coding potentials. For this analysis, the *Igk* locus, *Gm32432*, and *Dscam* genes' transcripts were removed, since these all have up to 98 transcripts and several genes nested within large, complex loci (Appendix 2 figure 3).

I then aimed to define the general characteristics of ORFs predicted from parent genes of MERVL-promoted transcripts, and their reference counterparts. The remaining 161 parent genes (after removing the aforementioned genes) with 535 transcripts resulted in 11241 ORFs, of which 2116 have a coding potential of over 0.44 (the CPAT threshold (Wang et al., 2013) (Figure 3.27). MERVL-promoted ORFs are statistically less likely to be coding overall, despite the offset in total ORF numbers (Figure 3.28, left). Additionally, MERVL-promoted parent genes overall ALL have at least one non-coding MERVL-promoted ORF, and the majority have both coding and non-coding MERVL-promoted ORFs. However, no parent genes have *only* coding MERVL-promoted ORFs (Figure 3.29) While the overall shape of the distributions of transcript coding potential is bimodal for both MERVL-promoted transcripts and

reference transcripts of the same genes (Figure 3.28, left), the cumulative distribution of coding potentials shows a shift towards lower coding potentials for MERVL-promoted ORF relative to the parent genes' reference ORF (Figure 3.28, right).

Next, I examined the change in coding potential relative to transcript length and predicted ORF length (Figure 3.30, top), to confirm any correlation between these. Transcript length shows a bimodal distribution, with some transcripts of long lengths (>1750nt) predicted to have very high or low coding potentials. The majority of transcripts are under 750nt in length. As expected, there is a strong correlation between transcript length and open reading frame length (Figure 3.30, bottom).

Finally, I selected the transcript with the ORF with the highest coding potential, per gene, to select potential candidates for examining protein domain changes and protein structure (Appendix 4). Extracting the transcript with the ORF of the highest coding potential resulted in 91 transcripts (and thus 91 genes). At this point, I integrated the expression data analysed in 3.3.1, to build a complete picture of which MERVL-promoted transcripts may influence cell function during ZGA. Intersecting this list of 91 transcripts with the genes which have 2C-specific transcripts (as completed in 3.3.1) resulted in a total of 15 genes which have 2C-specific coding MERVL-promoted ORFs (*Sgk3*, *P4ha2*, *Abcb5*, *Gm40910*, *Gnpnat1*, *Plaat3*, *Csrnp3*, *Ppig*, *Gm33923*, *Gm11784*, *Gm42808*, *Gm5309*, *Gm4011*, *Zfp560*, *Fundc1*) (Figure 3.31). Of these, at least 1 transcript is 2C specific and has an ORF predicted to be coding; there may be other MERVL-promoted transcripts with coding ORFs that are not 2C specific.

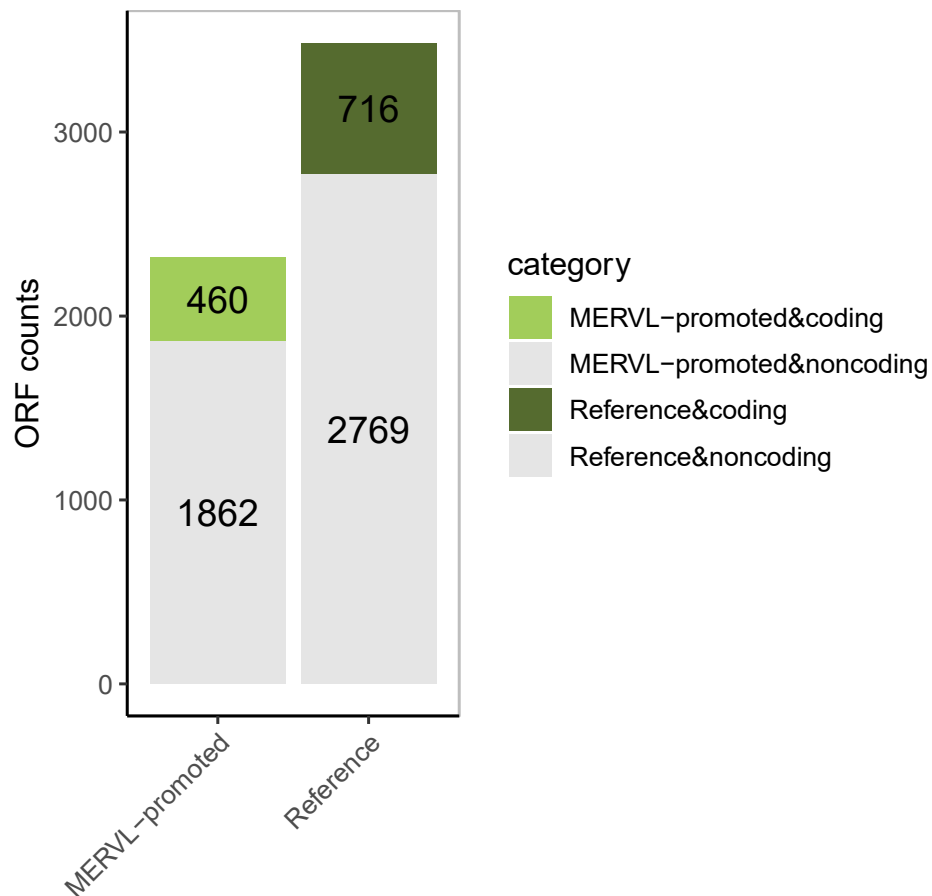


Figure 3.27: Counts of coding and non-coding ORFs derived from all MERVL-promoted and reference transcripts. Coding ORFs counts from MERVL-promoted and reference transcripts in light and dark green respectively, and non-coding transcripts in grey.

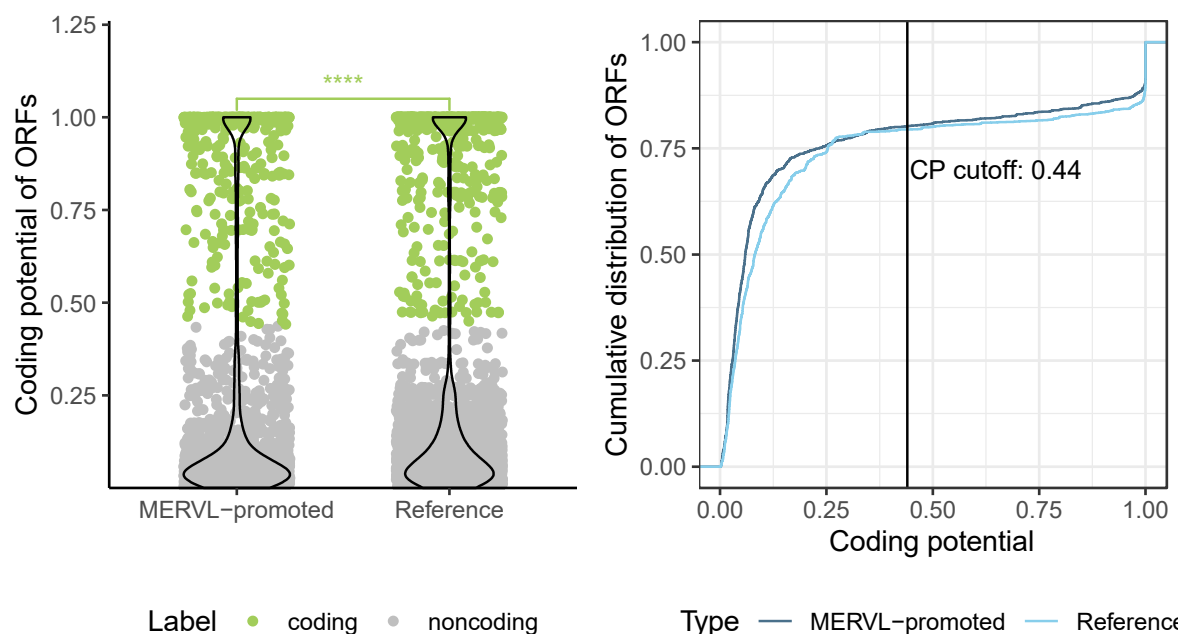


Figure 3.28: Examination coding and non-coding ORFs derived from MERVL-promoted and reference transcripts. Top: Coding potential distribution for MERVL-promoted ORFs versus reference ORFs, in green and grey for coding and non-coding ORFs respectively, Komolgorov-Smirnov P-value 0.00000000000000022. Bottom: Cumulative distribution of ORFs and coding potentials, with MERVL-promoted ORFs in dark blue and reference ORFs in light blue; proportionally, more MERVL-promoted ORFs are non-coding compared to reference ORFs.

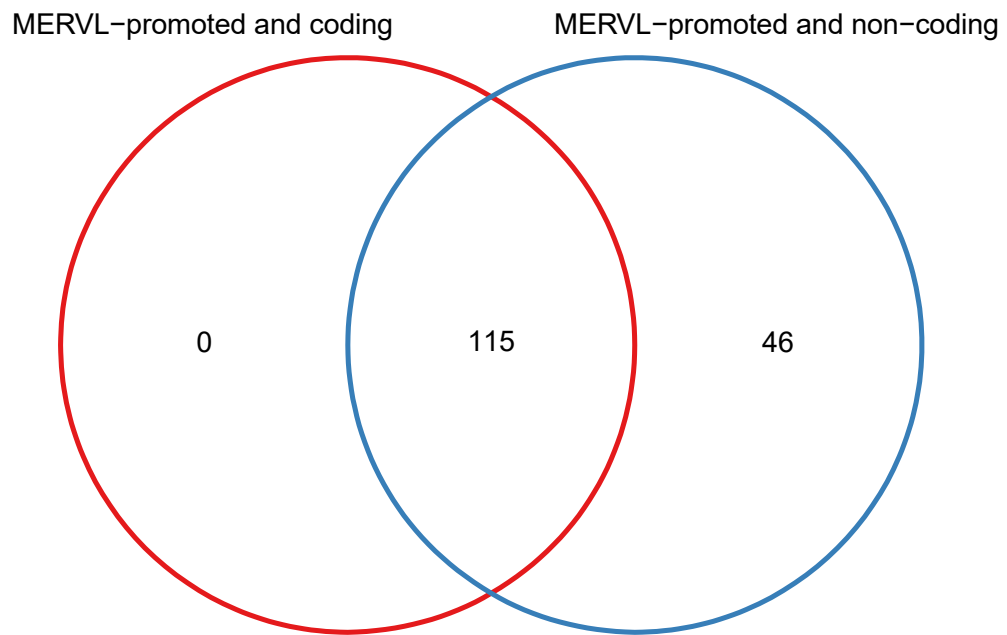


Figure 3.29: Venn diagram comparing whether a parent gene has coding or non-coding ORFs. No parent genes have only coding ORFs; they each have at least one non-coding ORF, if not all non-coding.

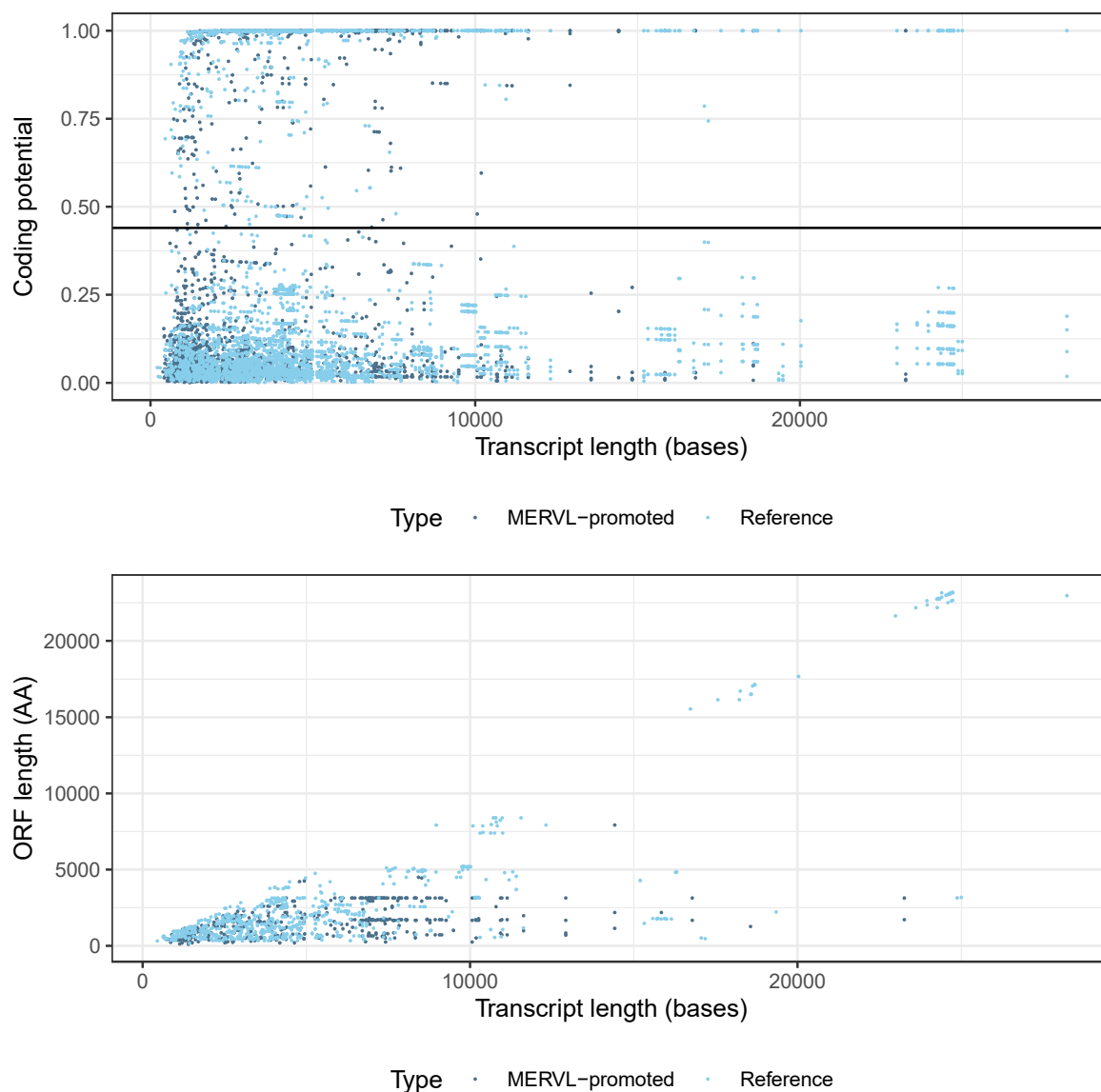


Figure 3.30: Comparison of coding potential, mRNA transcript length, and ORF length. Top: Coding potential for mRNAs plotted against the coding potential for their derived ORFs, with longest ORFs with highest coding potentials labelled, MERVL-promoted transcripts in dark blue and reference transcripts in light blue. Horizontal line at 0.44 denoting coding potential cut-off. Bottom: ORF length plotted against transcript length, (Kendall's Tau 0.3935847, positive

correlation) with longest ORFs and longest transcripts labelled, MERVL-promoted transcripts in dark blue and reference transcripts in light blue.

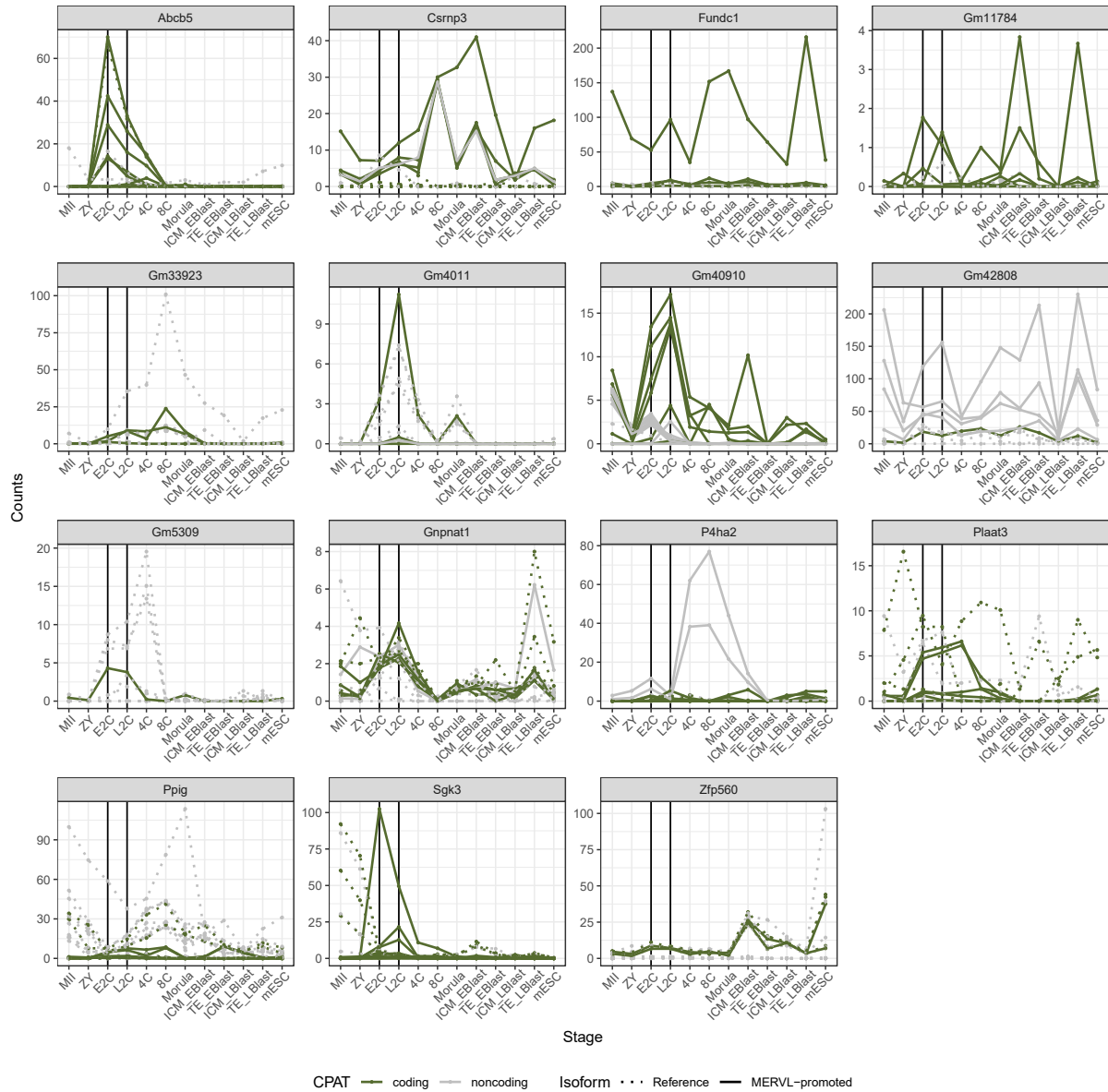


Figure 3.31: Expression patterns across preimplantation development of 2C-specific MERVL-promoted transcripts in context of their parent genes, for genes with coding transcripts. Transcript raw counts shown across MII oocytes (MII), 1C zygotes (ZY), 2C stage embryos (2C), 4C stage embryos (4C), 8C stage embryos (8C), 16 cell stage embryos (Morula), Early blastocyst inner cell mass (ICM_Eblast), Early blastocyst trophectoderm (TE_Eblast), Late blastocyst inner cell mass (ICM_Lblast), Late blastocyst trophectoderm (TE_Lblast), and mESCs. Coding transcripts in dark green, with

non-coding transcripts in grey, and MERVL-promoted transcripts in solid lines and reference transcripts in dotted lines.

3.3.3. Protein domain homology in MERVL-promoted transcripts.

Having defined that ORF lengths differ between MERVL-promoted transcripts and that certain exons are skipped, I next examined the changes in protein domains between reference isoforms and MERVL-promoted isoforms. To complete a 1:1 examination, I selected the ORF with the highest coding potential per parent gene for both MERVL-promoted transcripts and reference transcripts. I translated the mRNA sequences of these ORFs to amino acids using the `Biostrings` package in R, and input these to InterProScan (Jones et al., 2014).

Out of the 9 known genes with 2C-peaking MERVL-promoted transcripts, 6 of them returned a change in protein domain between the selected MERVL-promoted ORF and reference open reading frame, *Abcb5*, *Csrnp3*, *Fundc1*, *Gnpnat1*, *Sgk3*, and *Zfp560* (Table 5). Each of these genes lost at least one domain in the MERVL-promoted isoform. *Fundc1*, *Gnpnat1*, and *Zfp560* all gained domains also, but in the case of *Fundc1* and *Zfp560*, these gained domains were those encoded in the MERVL-int segments. *Fundc1* and *Zfp560* were the only genes with alternative isoforms promoted by complete MERVL elements; the four other genes were all promoted by solo LTRs. Perplexingly, *Gnpnat1*'s gained domain in its MERVL-promoted isoform is the functional Gcn5-related N-acetyltransferase (GNAT) domain.

While a net change analysis is illuminating to see potential changes in functionality, many domains were retained. For example, both the MERVL-promoted isoform and reference isoform selected for *Sgk3* retained a RIBOSOMAL PROTEIN S6 KINASE domain (Table 5). To compare these directly, I manually examined the domains, filtered for redundancy (e.g., for *Zfp560*, collapsed

overlapping “C2H2-type zinc finger” and “C2H2-type Zinc-finger” domains into one “C2H2-type zinc finger” annotation), and visualised these along peptide lengths

Abcb5 shows a case of domain shuffling, where some domains are retained (e.g., P-loop containing nucleoside triphosphate hydrolase), but in a different organisation between the reference and MERVL-promoted isoforms (Figure 3.32). *Gnpnat1* shows a similar domain shuffling pattern, where the domains are common between the isoforms, but the order is not recognisable (Figure 3.34). The MERVL-promoted isoform and the reference isoforms of *Csrnp3* share the consensus disorder prediction, Cysteine/serine-rich nuclear protein N-terminus, and the TGF-BETA INDUCED APOPTOSIS PROTEIN domains, but the MERVL-promoted isoform truncates prematurely, thus skipping Cysteine/serine-rich nuclear protein (CSRNP) family signature and the coiled domain (Figure 3.33). *Sgk3* shows the classic, expected result, with the MERVL-promoted isoform skipping the PX_2 and PX_CISK domains, but retaining all others further down the peptide (Figure 3.36). Finally, both *Fundc1* and *Zfp560* retain no domains specific to their parent genes; *Fundc1* loses the FUN14 DOMAIN and the FUN14 family signature, and *ZFP560* loses the entire zinc finger array and bed5 domains (Figures 3.34, 3.37). However, as aforementioned, these both retain endogenous retrovirus associated domains (e.g., dUTPase and dUTPase-like, RNaseH and Ribonuclease H).

Overall, from this analysis it is clear that domain changes occur in MERVL-promoted transcripts compared to reference isoforms. This implies potentially different functions of gene products on MERVL promotion. In the cases of domain retention between the reference isoform and the MERVL-promoted isoform, the peptides may have a similar function, but the differences may result in different localisation, reduced function, or entire functional loss or degradation

Gene	Domains in MERVL isoform	Domains in Reference isoform	Net domain change	gained_domains	lost_domains
Abcb5	ABC transporter,ABC transporter integral membrane type-1 fused domain profile.,ABC transporter transmembrane region,ABC transporter type 1, transmembrane domain,ATP-DEPENDENT PERMEASE MDL1, MITOCHONDRIAL,P-	AAA_5,ABC transporter,ABC transporter integral membrane type-1 fused domain profile.,ABC transporter transmembrane region,ABC transporter type 1, transmembrane domain,ABC transporters family signature.,ABC_6TM_P	-6		AAA_5,ABC transporters family signature.,ABC_6TM_P M_Pgp_ABCB1_D1_gp_ABCB1_D2_lik e,ABC_MTABC3_ MDL1_MDL2,ATP-binding cassette, ABC transporter-type domain profile.

loop containing	gp_ABCB1_D1_like,AB
nucleoside triphosphate	C_6TM_Pgp_ABCB1_D
hydrolase,P-loop	2_like,ABC_MTABC3_
containing nucleoside	MDL1_MDL2,ATP-
triphosphate hydrolases	binding cassette, ABC
	transporter-type
	domain profile.,ATP-
	DEPENDENT
	PERMEASE MDL1,
	MITOCHONDRIAL,P-
	loop containing
	nucleoside triphosphate
	hydrolase,P-loop
	containing nucleoside
	triphosphate hydrolases

Csrnp3	consensus disorder prediction,Cysteine/serine-rich nuclear protein N-terminus,TGF-BETA INDUCED APOPTOSIS PROTEIN	Coil,consensus disorder prediction,Cysteine/serine-rich nuclear protein (CSRNP) family signature,Cysteine/serine-rich nuclear protein N-terminus,TGF-BETA INDUCED APOPTOSIS PROTEIN	-2		Coil,Cysteine/serine-rich nuclear protein (CSRNP) family signature
Fundc1	dUTPase-like,ENDOGENOUS RETROVIRUS GROUP K MEMBER POL PROTEIN,Integrase catalytic domain profile.,Integrase core	FUN14 DOMAIN CONTAINING,FUN14 family	6	dUTPase-like,ENDOGENOUS RETROVIRUS GROUP K MEMBER POL PROTEIN,Integrase catalytic domain profile.,Integrase core	FUN14 DOMAIN CONTAINING,FUN14 family

	domain,Ribonuclease H superfamily,Ribonuclea se H-like,RNase H,RNase H type-1 domain profile.		domain,Ribonuclease H superfamily,Ribonucle ase H-like,RNase H,RNase H type-1 domain profile.
Gnpnat1	-,Acetyltransferase (GNAT) family,Acyl- CoA N-acyltransferases (Nat),Gcn5-related N- acetyltransferase (GNAT) domain profile.,GLUCOSAMINE 6-PHOSPHATE N- ACETYLTRANSFERAS E	-,Acetyltransferase (GNAT) family,Acyl- CoA N-acyltransferases (Nat),GLUCOSAMINE 6-PHOSPHATE N- ACETYLTRANSFERAS E	1 Gcn5-related N- acetyltransferase (GNAT) domain profile.

P4ha2	- ,2OG-Fe(II) oxygenase superfamily,Fe(2+) 2- oxoglutarate dioxygenase domain profile.,p4hc,PROLYL 4-HYDROXYLASE ALPHA SUBUNIT,Prolyl 4- Hydroxylase alpha- subunit, N-terminal region,q2cbj1_9rhob like domain,Tetratricopeptid e repeat domain,TPR repeat profile.,TPR-like	- ,2OG-Fe(II) oxygenase superfamily,Fe(2+) 2- oxoglutarate dioxygenase domain profile.,p4hc,PROLYL 4-HYDROXYLASE ALPHA SUBUNIT,Prolyl 4- Hydroxylase alpha- subunit, N-terminal region,q2cbj1_9rhob like domain,Tetratricopeptid e repeat domain,TPR repeat profile.,TPR-like	0
--------------	---	---	---

Plaatz	endopeptidase domain like (from Nostoc punctiforme),HRAS-LIKE SUPPRESSOR - RELATED,Lecithin retinol acyltransferase,LRAT domain profile.	endopeptidase domain like (from Nostoc punctiforme),HRAS-LIKE SUPPRESSOR - RELATED,Lecithin retinol acyltransferase,LRAT domain profile.	0
Ppig	consensus disorder prediction,Cyclophilin peptidyl-prolyl cis-trans isomerase signature,Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD,Cycloph	consensus disorder prediction,Cyclophilin peptidyl-prolyl cis-trans isomerase signature,Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD,Cycloph	0

	ilin-like,Cyclophilin-like domain	ilin-like,Cyclophilin-like domain		
	superfamily,Cyclophilin	superfamily,Cyclophilin		
	-type peptidyl-prolyl cis-trans isomerase domain	-type peptidyl-prolyl cis-trans isomerase domain		
	profile.,Cyclophilin-type	profile.,Cyclophilin-type		
	peptidyl-prolyl cis-trans isomerase	peptidyl-prolyl cis-trans isomerase		
	signature.,PEPTIDYL-PROLYL CIS-TRANS ISOMERASE	signature.,PEPTIDYL-PROLYL CIS-TRANS ISOMERASE		
Sgk3	AGC-kinase C-terminal domain	AGC-kinase C-terminal domain	-5	PX domain,PX domain profile.,PX domain
	profile.,consensus	profile.,consensus		
	disorder	disorder		superfamily,PX_2,
	prediction,Phosphorylas	prediction,Phosphorylas		PX_CISK

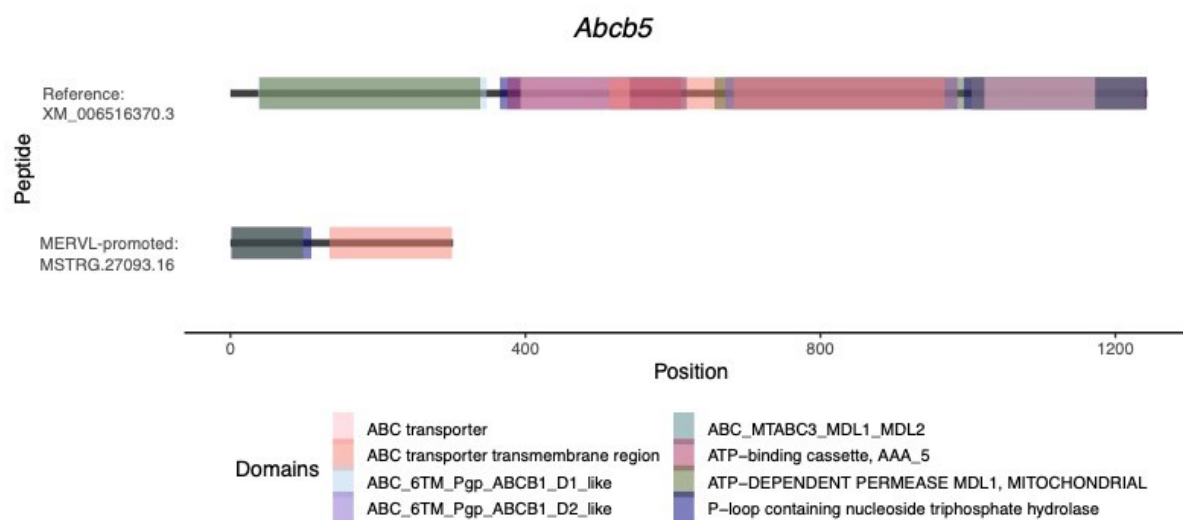
e Kinase; domain	e Kinase; domain
1,pkinase_C_6,Protein	1,pkinase_C_6,Protein
kinase C terminal	kinase C terminal
domain,Protein kinase	domain,Protein kinase
domain,Protein kinase	domain,Protein kinase
domain profile.,Protein	domain profile.,Protein
kinase-like (PK-	kinase-like (PK-
like),Protein kinases	like),Protein kinases
ATP-binding region	ATP-binding region
signature.,RIBOSOMAL	signature.,PX
PROTEIN S6	domain,PX domain
KINASE,Serine/Threoni	profile.,PX domain
ne protein kinases	superfamily,PX_2,PX_
active-site	CISK,RIBOSOMAL
signature.,serkin_6,STK	PROTEIN S6
c_SGK3,Transferase(Ph	KINASE,Serine/Threoni

	osphotransferase)	ne protein kinases			
	domain 1	active-site			
		signature.,serkin_6,STK			
		c_SGK3,Transferase(Ph			
		osphotransferase)			
		domain 1			
Zfp560	DNA/RNA	bed5,beta-beta-alpha	8	DNA/RNA	bed5,beta-beta-
	polymerases,dUTPase,d	zinc fingers,C2H2-type		polymerases,dUTPase,	alpha zinc
	UTPase-like,dUTPase-	zinc		dUTPase-	fingers,C2H2-type
	like superfamily,HIV	finger,c2h2final6,Classic		like,dUTPase-like	zinc
	Type 1 Reverse	Zinc Finger,Zinc finger		superfamily,HIV Type	finger,c2h2final6,C
	Transcriptase, subunit	C2H2 type domain		1 Reverse	lassic Zinc
	A, domain 1,Integrase	profile.,Zinc finger		Transcriptase, subunit	Finger,Zinc finger
	catalytic domain	C2H2 type domain		A, domain 1,Integrase	C2H2 type domain
	profile.,Integrase core	signature.,ZINC		catalytic domain	profile.,Zinc finger
	domain,POL	FINGER PROTEIN,Zinc		profile.,Integrase core	C2H2 type domain

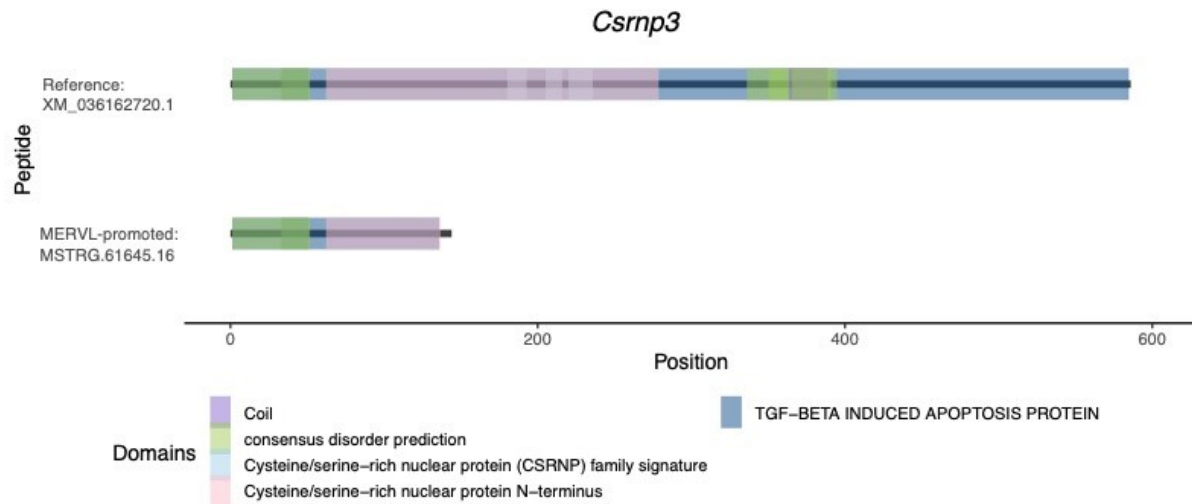
PROTEIN,Reverse	finger, C2H2 type,Zinc-	domain,POL	signature.,ZINC
transcriptase (RNA-	finger double domain	PROTEIN,Reverse	FINGER
dependent DNA		transcriptase (RNA-	PROTEIN,Zinc
polymerase),Reverse		dependent DNA	finger, C2H2
transcriptase (RT)		polymerase),Reverse	type,Zinc-finger
catalytic domain		transcriptase (RT)	double domain
profile.,Reverse		catalytic domain	
transcriptase/Diguanyla		profile.,Reverse	
te cyclase		transcriptase/Diguanyl	
domain,Ribonuclease H		ate cyclase	
superfamily,Ribonuclea		domain,Ribonuclease	
se H-like,RNase		H	
H,RNase H type-1		superfamily,Ribonucle	
domain profile.,RNase		ase H-like,RNase	
H-like domain found in		H,RNase H type-1	
reverse		domain profile.,RNase	

transcriptase,RT_ZFRE	H-like domain found
V_like,trimeric_dUTPas	in reverse
e	transcriptase,RT_ZFR
	EV_like,trimeric_dUT
	Pase

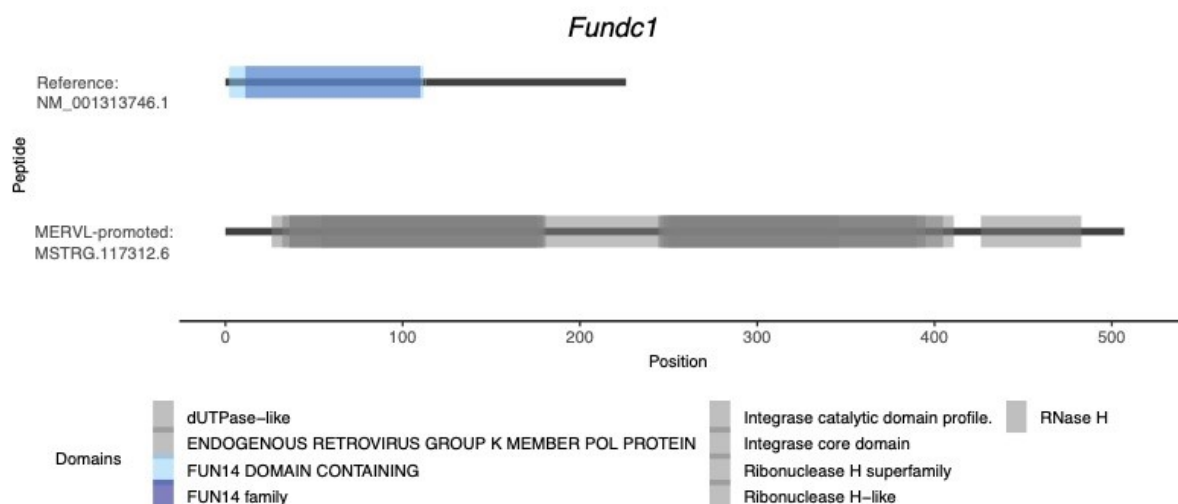
Table 5: Net changes in protein domains between the reference ORF and MERVL-promoted ORF with the highest coding potentials per 2C-peaking gene. Net change is positive if the MERVL-promoted isoform gains domains, negative if it loses domains.



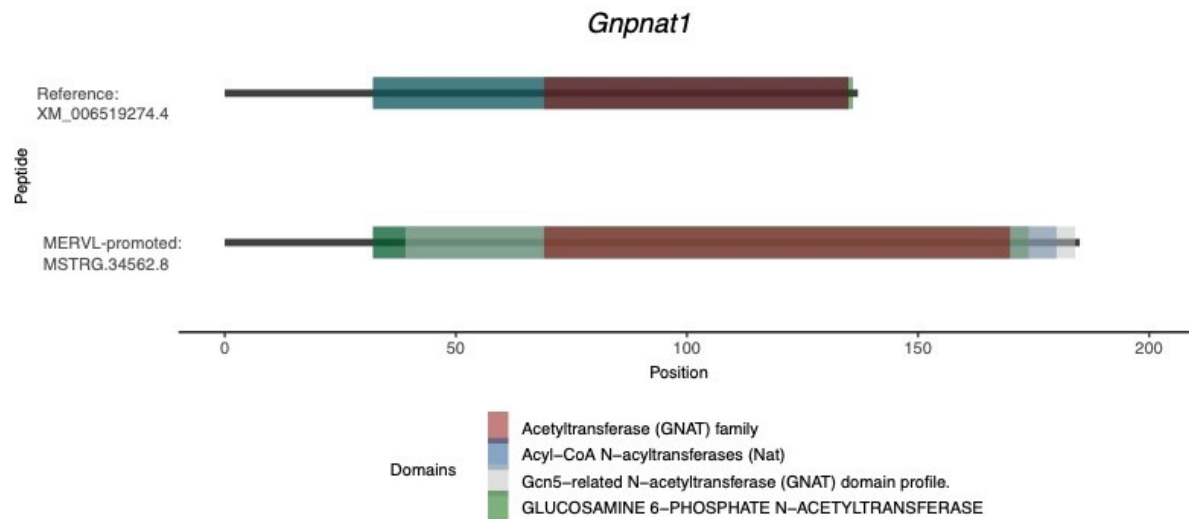
3.32: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Abcb5*. Reference isoform plus transcript code shown above, and MERVL-promoted isoform plus transcript code shown below, Black line indicates peptide length, coloured boxes (overlaps present) indicate protein domain homologies.



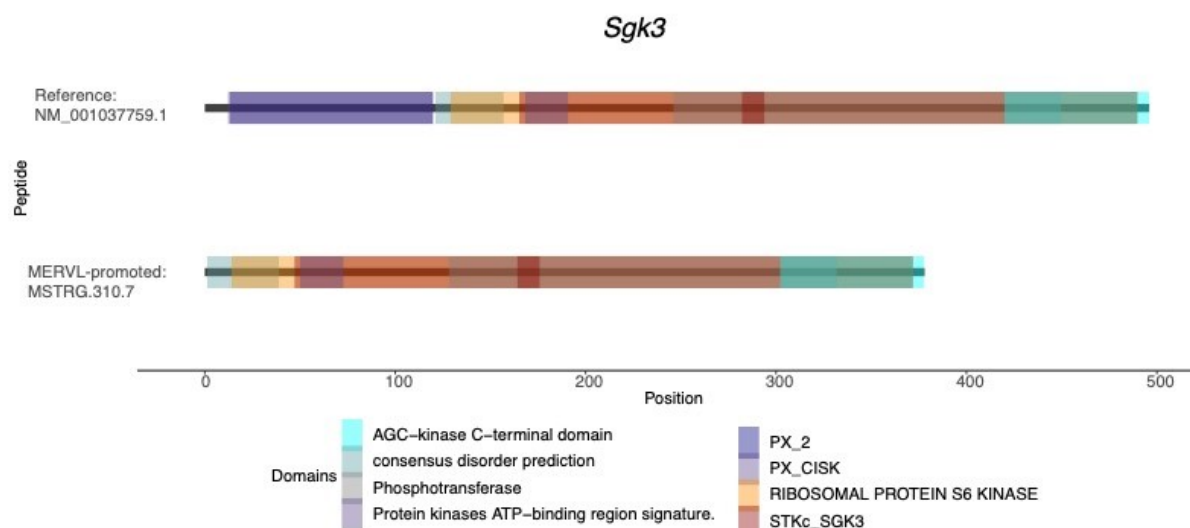
3.33: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Csrnp3*. Reference isoform plus transcript code shown above, and MERVL-promoted isoform plus transcript code shown below, Black line indicates peptide length, coloured boxes (overlaps present) indicate protein domain homologies.



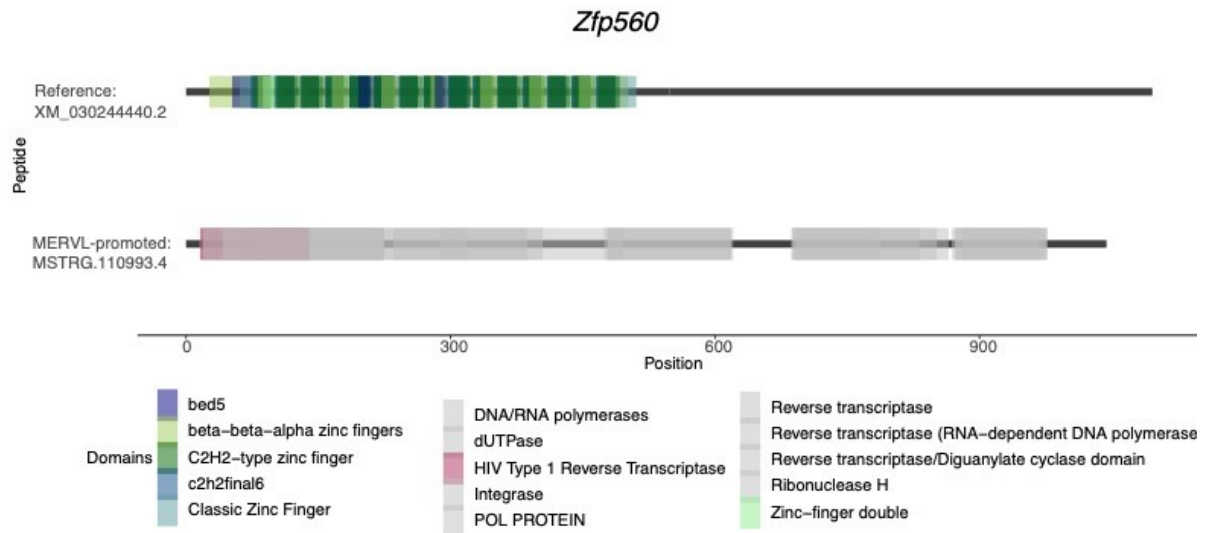
3.34: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Fundc1*. Reference isoform plus transcript code shown above, and MERVL-promoted isoform plus transcript code shown below, Black line indicates peptide length, coloured boxes (overlaps present) indicate protein domain homologies.



3.35: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Gnpnat1*. Reference isoform plus transcript code shown above, and MERVL-promoted isoform plus transcript code shown below, Black line indicates peptide length, coloured boxes (overlaps present) indicate protein domain homologies.



3.36: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Sgk3*. Reference isoform plus transcript code shown above, and MERVL-promoted isoform plus transcript code shown below, Black line indicates peptide length, coloured boxes (overlaps present) indicate protein domain homologies.



3.37: Comparison of manually filtered protein domains encoded in the reference isoform and MERVL-promoted isoform with the highest coding potentials for *Zfp560*. Reference isoform plus transcript code shown above, and MERVL-promoted isoform plus transcript code shown below, Black line indicates peptide length, coloured boxes (overlaps present) indicate protein domain homologies.

4. Discussion and future perspectives.

In this study, I aim to address an outstanding question; what are the dynamics of MERVL elements during ZGA, in promoter activity, transcription, and potentially, the function of transcript isoforms driven by MERVL alternative promoters? Here, utilising RNA-seq data, I also examine the expression of MERVL elements in 2C embryos and 2CLCs and show similarities in downstream expression from MERVL elements. With SLIC-CAGE data, I confirm that transcription is indeed initiating within the MERVL LTRs, and potential eRNA transcription occurs from MERVL elements in oocyte maturation. Through the use of long-read sequencing of 2C embryos, with support from short-read sequencing of both 2C embryos and an established *in vitro* model, 2CLCs, this project makes substantial progress in defining high-confidence transcript isoforms driven by MERVL promoters during ZGA. Finally, I examine the expression patterns of MERVL-promoted transcripts across preimplantation development using long-read sequencing data and examine the coding potential of these MERVL-promoted transcripts, to begin understanding whether or not these transcripts might have a cellular function during ZGA. This opens doors to the examination of protein domains changed in MERVL-promoted isoforms over the reference, known isoforms, and further experimental validation for future studies.

4.1 Contextualising results.

4.1.1 Defining MERVL elements.

The basis of this work depends on a list of MERVL elements. These have been previously defined by this lab (Kruse et al., 2019). Mouse embryonic stem cells, and thus also 2CLCs, are most commonly derived from 129P2/Ola strain mouse blastocysts (from the ICMs of these), and *in vivo* datasets (2C embryos) used in this study are all crosses containing at least one parent of the C57Bl6J strain. Thus, MERVL promoters to be studied must be fixed in both these strains. This logic was used in the original construction of the MERVL list of 973 total elements (604 solo LTRs, and 369 complete MERVL elements). That work was, however, completed using the mm10 mouse reference genome, which has since been revised. This study uses the mm39 reference genome, and therefore, I used UCSC's LiftOver tool (Kent et al., 2002) to lift the same coordinates of MERVL elements over to the new reference genome. During this process, some MERVL elements were lost (Figure 3.1-3.2); perhaps re-construction of the list would have been more appropriate, especially since the initial list by Kruse et al., itself was lifted over from the mm9 mouse assembly, one before mm10. Additionally, more structural variant analysis has been completed since, using PacBio long-read sequencing and the mm39 reference genome (Ferraï et al., 2023). Their work finds that 28% of structural variants between their chosen mouse strains result from ERV elements (including ERVK, ERV1, and MaLR elements). However, their study lacks the 129P2/Ola mouse strain, and thus cannot be used in this study to define fixed MERVL elements between my chosen *in vivo* and *in vitro* datasets. Furthermore, it would be worthwhile to expand on this list by including non-canonical MERVL element sequences. In this study, thus far only MT2_Mm LTR elements (and full MERVLs

composed of these) have been used, partially for simplicity and partially because their sequences are confirmed. However, prior examples of MERVL-promoted transcripts, specifically the example presented by Modzelewski et al., in 2023 of *CDK2AP1*, were driven by a non-canonical MT2B2 MERVL solo LTR. In fact, examining this specific MT2 element in the new mm39 RepeatMaster track on the UCSC genome browser (Kent et al., 2002), it appears to have been re-classified as an MT2C element. Honda et al.,’s most recent example of a novel *PRMT6* isoform was promoted by an MT2B2 element.

4.1.2 Transcriptional signal at MERVL loci.

Transcriptional signal is known to be present in MERVL elements and has been confirmed to read through the end of the MERVL elements (Kruse et al., 2019, using the RNA-seq data for 2CLCs and mESCs from Hendrickson et al., 2017). In this study, I re-assessed this with the latest reference genome and expanded this to other datasets (Berrens et al., 2021; Fan et al., 2020; Ishiuchi et al., 2015; Sakashita et al., 2023). In general, these all show a similar pattern, of transcriptional signal being highest in MERVL sequences’ LTRs (or solo LTRs only), and mainly only present in 2C embryos and 2CLCs (Figure 3.4). However, some faint signal at MERVL loci is visible in the Hendrickson et al., (2017) mESCs. The experimental design in Hendrickson et al.,’s study is to tag a MERVL element with GFP, and thus, cells expressing MERVLs also express GFP. Additionally, this study focuses on the role of *Dux*, therefore, there is a doxycycline-inducible *Dux* overexpression construct in this cell line. In mESC culture, only the 2CLC population expresses MERVL elements and therefore can be isolated via FACS sorting. I selected the cells which were treated with doxycycline to express a *Dux* transgene and GFP positive as 2CLCs. For the mESC control, I selected those *not* treated with doxycycline (therefore not

overexpressing *Dux*), and *also* GFP-negative (therefore, definitely not 2CLCs, and only mESCs). However, it appears that there has been some 2CLC contamination in the mESC faction, possibly 2CLCs which reverted after sorting but before RNA extraction. However, this cannot be proved.

The Caf-1 knockdown 2CLCs and mESCs (Ishiuchi et al., 2015) show a much more expected pattern of expression, with MERVL expression absent entirely in 2CLCs (Figure 3.4). The average TPM here is lower than the TPM for the Hendrickson et al., data, but this may be attributable to two things: 1. The overall lower sequencing depth, and 2. The shorter read length (Table 3, Appendix 1.1). This may be tested by downsampling the Hendrickson et al., dataset to the same depth as the Ishiuchi et al., data, and/or artificially trimming the reads from the 150bp length of the Hendrickson dataset to the 50bp length of the Ishiuchi dataset. The Sakashita et al., 2C embryo RNA-seq also recapitulates the expected pattern, but this shows a lower downstream transcriptional signal. Potentially this might suggest that read-through transcription is more prevalent in 2CLCs than 2C embryos; however, with only one 2C short-read dataset, this cannot be confidently stated and thus is only speculation. Analysing another 2C embryo short-read dataset would confirm or deny this.

Interestingly, both of the long-read RNA-seq datasets show a very sharp drop-off at the 3' terminus of the MERVLs (full and solo) (Figure 3.5). While there is inconsistent and speckled signal downstream of the MERVL elements, there is a distinct lack of signal upstream, suggesting that read-through transcription initiating in the MERVL loci is definitely happening in these cells and that the long-read sequencing techniques are indeed capturing this. Moreover, an assessment of average exon sizes and intron sizes of MERVL-promoted transcripts (Figure 3.17) suggests that this speckled signature is showing coverage specifically of exons only.

Finally, we see a difference in signal between the Berrens et al., dataset (CELLOseq) and the Fan et al. dataset (SCANseq). The library preparation techniques (discussed below) cause differences in the cleanliness of transcripts sequenced. This is examined in section 3.2, where we see more unconfirmed and artefactual transcripts assembled with Stringtie. This provides some insight into why the downstream signal is more scattered in the SCAN-seq expression heatmap.

4.1.3 Transcriptional initiation in MERVL LTRs.

Examination of CAGE-seq data (specifically SLIC-CAGE) shows that, as we would expect, transcription initiation occurs in the LTRs, where there is a known promoter sequence (Benachenhou et al., 2009, 2013; Choi et al., 2017; Hendrickson et al., 2017; De Iaco et al., 2020; Kruse et al., 2019; Maksakova et al., 2013; Peaston, 2008; Schoorlemmer et al., 2014) (Figure 3.6-3.7). CAGE-seq data, however, has very short reads (discussed below in 4.2.3), therefore, I chose to show both uniquely mapping reads only, and also, the single-best alignment (SBA) for each multimapping read. An example of why this may be a better option is the P14 For the 2C and 4C embryos, In the SBA plots, for full MERVL elements, CAGE-seq peaks are clearly seen in the LTRs. In solo LTRs, these are even seen in the uniquely mapping reads plots. As expected, by the mESC stage these peaks are lost.

The initial rationale behind assessing CAGE-seq data from oocyte maturation was twofold. Firstly, the P6 and P14 timepoint primordial oocyte CAGE data was processed because of the unexpected CAGE-seq signal in transcriptionally dormant MII oocytes. If mRNA in MII oocytes is present, it cannot have been actively transcribed at the MII stage and must have been transcribed in a prior stage of meiosis or of oogenesis (e.g., at transcriptionally active stages, which P6 and P14 stage oocytes are (Schultz et al., 2018)) and persisted to this stage. Therefore, I

expected to see a CAGE-seq signal at MERVL elements in primordial oocytes, which I did. Secondly, this analysis was completed because of the unexpected abundance of MERVL-promoted isoforms in the MII oocytes (section 3.3.2). Future work should examine the expression of MERVL-promoted transcripts during oocyte development, to confirm that the same transcripts are being expressed and that this is not artefactual.

Next, I examined the stranded expression of CAGE-seq reads. A faint signal in the SBA plots for both full MERVLs and solo LTRs shows minor bidirectional transcription initiating within the LTRs (Figure 3.8-3.9). Potentially, this indicates enhancer RNA transcription; recently, enhancer function has been evidenced for MERVL elements (albeit in HEK293 cells), which may warrant further examination.

Finally, one strange signal is consistently observed in all time points where full MERVLs are expressed: CAGE-seq transcriptional signal within the MERVL-int segment. To assess this further, isolation of the MERVL-int segments for the conserved C57Bl6J-129P2/Ola MERVL list, and examination of the CAGE-seq signal in depth, to confirm whether this is not a colour scaling issue. Furthermore, it would be important to assess whether there is *Dux* binding in the MERVL-int segment. Prior work by Kruse et al., in 2019 has confirmed that MERVL elements are only expressed if they harbour *Dux* binding sites, therefore, if the binding site is absent in these MERVL-int segments, then any CAGE-seq transcriptional signal in the MERVL-int is likely artefactual. Nonetheless, this has no bearing on the transcription of the complete MERVL element, since it has been confirmed that the LTRs contain *Dux* binding sites by Kruse et al., (2019).

4.1.4 Definition of a high-confidence list of ZGA-specific MERVL-promoted isoforms.

Next, I defined MERVL-promoted transcripts by overlapping the first exon of all transcripts with the selected MERVL loci, with a threshold of 10bp (Figure 3.10). This was done with the R package `GenomicRanges`. The default overlap for the `subsetByOverlaps` function is 1bp only, thus, I extended this to 10bp, for a little more certainty that the MERVL element was indeed part of the 1st exon. Other such analyses have kept it at 1bp (Modzelewski et al., 2021; Shah et al., 2023). In theory, this could have been extended to 150bp, the size of the smallest solo LTR (Figure 3.3), however, the promoter sequence in LTRs are known to be some way into the LTR (Benachenhou et al., 2009, 2013), therefore the overlap threshold would have to be less than that. It might be interesting to consider how many transcripts are gained or lost from the status of being MERVL-promoted if this overlap threshold is toggled. The approach followed to define genes with MERVL-promoted transcripts is the same as has been followed by Modzelewski et al., in 2021.

While it might initially strike as odd that not many parent genes are shared between different datasets (Figure 3.13), considering that other studies suffer this same symptom only highlights this is a shortcoming of all similar studies. Possibly, this result can be chalked up to batch effects (discussed below). Furthermore, compared to all other studies evaluated, my work identifies the highest number of genes with MERVL-promoted transcripts, and given the usage of long-read sequencing data to define this list, which other studies so far lack, these may be considered reliable.

4.1.5 Expression and translational capacity of transcripts.

When aggregating the expression of the transcript level to the gene level, there is a change in the overall result. Parent genes appear to match the expected pattern of expression peaking at the 2C stage, with a subset being maternally inherited genes. However, in the initial examination, where transcripts themselves are considered, there is a much more scattered expression pattern. Given that splicing complexity is crucial at ZGA (Abe et al., 2022; Lee et al., 2024; Svoboda et al., 2015; Tian et al., 2020; Wyatt et al., 2022), it is crucial to consider transcript levels separately to overall gene expression to fully understand the transcriptional landscape of ZGA.

On selecting genes with ZGA specific transcripts, I imposed a threshold of $\log_2FC > 1$ for late 2C stage over Oocyte, and over Morula both. The logic here was that by the morula stage, MERVL elements expression should be strongly decreased (Sakashita et al., 2023). However, it appears that several MERVL-promoted transcripts (e.g., those belonging to *Gnpnat1*, *Farp1*, *Vmn2r94*, *Zfp560*, *Ppig*, or *Csrnp3*) have a resurgence in expression AFTER the morula stage. It's possible that the MERVL promoter is activated (versus de-repressed) at these time points. This hypothesis may be explored using TE expression tools (e.g., Tetranscripts (Jin et al., 2015)), to examine expression of specific MERVL loci known to promote the aforementioned genes.

CPAT (Wang et al., 2013) was used to calculate the coding potential (i.e., the likelihood of translation) for open reading frames determined from transcripts of interest. This tool utilises information from coding and non-coding sequences, a benefit compared to other similar tools such as CPC2, which (Kang et al., 2017).

While CPAT draws out an abundance of data from only transcripts, drawing an exact correlation between the reference and MERVL-promoted coding potentials

is difficult. This is because each gene has several reference transcripts and several MERVL-promoted transcripts, and each transcript has several open reading frames predicted. Therefore, I selected the ‘best ORF’. For each gene, I selected the ORF with the highest coding potential, from the transcript with the highest average TPM in the SCAN-seq dataset (Fan et al., (2020), late 2C embryos). This was done for each gene’s MERVL-promoted transcript and each gene’s reference transcript and is available as a table in Appendix 4 (143 genes total, after removing those transcripts with no expression and those transcripts with no coding ORFs). The drawback here is that TPMs are limited to transcripts only, and while an ORF might be predicted to be the most likely to be translated, this gives no insight into protein abundance, or whether or not the protein will be decayed for other reasons after translation.

4.1.6 Specific cases

This study isolates several possible candidate genes with MERVL-promoted transcript isoforms, which might warrant further examination (potential analyses and experiments for these are discussed below). Each of these transcript structures and splicing patterns for long-read data is shown in Appendix 3.

In this study, I find that ~60% of genes with MERVL-promoted transcripts have a MERVL upstream of the gene. Of these, the known genes with MERVL-promoted transcripts predicted to have coding ORFs are *Fundc1*, *Gnpnat1*, *P4ha2*, *Plaat3*, and *Zfp560*. The remaining ~40% of genes have a MERVL inserted into an intron, resulting in a definite change to coding sequence. Of these genes with coding ORFs predicted, *Abcb5*, *Csrnp3*, *Ppig*, and *Sgk3* all have intronic MERVL elements. The start codon for *Ppig* is included in the MERVL-promoted isoform, despite being promoted from within the gene body, but *Abcb5*, *Csrnp3*, and *Sgk3* and have their canonical start codon skipped when the MERVL promoter drives transcription.

The MERVL-promoted isoforms of several of these genes have been recorded previously, or associated with ZGA separately. *P4ha2* is known to be an MT2_Mm-promoted (Yang et al., 2020). In this study, I find that 10 of the 15 total assembled isoforms in this study are MERVL-promoted (with 3 others being promoted by a non-repetitive intergenic region). One of these prematurely truncates, but all of these skip the first exon of this gene (Appendix 3, figure 9). Each of the MERVL-promoted isoforms is predicted to be coding. Whether this MERVL-promoted isoform retains the open reading frame of the reference isoform is to be confirmed. *P4ha2*'s start codon is in exon 3. In the work by Yang et al., no such in-depth analysis is completed, only evidence that the transcript is MERVL-promoted. Novel isoforms for *Sgk3* have been found using another long-read dataset, this time with PacBio data (Qiao et al., 2020), expressed at the 2C, 4C, and 8C stages (corroborating my own findings), but no association with MERVL elements was completed in Qiao et al.'s study, despite this isoform also being 5' truncated. In my work, I find 9 MERVL-promoted isoforms, out of the total 16 *Sgk3* isoforms assembled (4 of which are novel and not MERVL-promoted). Interestingly, *Sgk3* is also associated with germline mutations, if a MusD/ETn class II ERV is inserted into exon 6 (Gagnier et al., 2019). *Abcb5* is a known 2C-specific gene (Choi et al., 2017; Hackett et al., 2017; Zhang et al., 2019); in this study, I find 12/16 isoforms to be MERVL-promoted (1 is novel and not MERVL-promoted). Finally, *Zfp560* is known to be *Dux*-induced (Hendrickson et al., (2017), in line with the fact that MERVL expression relies on *Dux* binding (Kruse et al., 2019). In this study, I find 2 isoforms to be specifically MERVL-promoted, out of 7 isoforms assembled. Notably, 2 more of these isoforms initiate in an upstream unannotated promoter with a strong CAGE-seq signal at the 2C stage. These exclude the MERVL entirely (MERVLs do not contain splice acceptor sites (Franke et al., (2017))). While the MERVL-promoted transcripts truncate prematurely compared to

the reference transcripts, other novel transcripts carry on until the end of the gene and include full-length polyA tails.

On considering the protein domains for the MERVL-promoted transcripts for these genes, 6 out of the 9 known genes have any changes. These are *Abcb5*, *Csrpn3*, *Fundc1*, *Gnpnat1*, *Sgk3*, and *Zfp560*. The MERVL-promoted protein isoforms of *Fundc1* and *Zfp560* both have no domains in common with the reference isoforms (Figures 3.34, 3.37), only MERVL-int encoded domains. These are the only two genes with complete MERVL promoters. It is interesting to consider that *Gnpnat1* gains a functional domain associated with the gene's own function, in the MERVL-promoted isoform (Figure 3.35). *Abcb5* shows a case where Interproscan finds sequence homology within the MERVL-promoted isoform for shared domains with the reference isoform but in a shuffled order (Figure 3.32). Whether *Gnpnat1* and *Abcb5* result in functional isoforms is to be confirmed, possibly using structure prediction approaches (discussed below). *Csrpn3* provides a key example of the link between promoter choice and transcriptional termination, as discussed by Alfonso-Gonzalez et al., in 2023 since the MERVL-promoted peptide (resultant of the transcript) is C-terminal truncated relative to the reference (Figure 3.33). Finally, *Sgk3* shows a pattern following the initial hypothesis, that MERVL-promoted isoforms will skip coding sequences earlier on in the protein, and retain domains found later (Figure 3.36). It would be interesting to examine the structures of these, and potential functions during ZGA.

4.2 Technical considerations.

While short-read sequencing technology has been the gold standard for RNA sequencing due to deep sequencing capabilities, difficulties remain in resolving transcript isoform structures accurately, due to PCR biases in either the 5' or 3' end

of RNA reads, and since transcript lengths very frequently exceed lengths of short-reads (Amarasinghe et al., 2020; Goodwin et al., 2016). Using long-read sequencing offsets this difficulty since it can capture the entire length of the mRNA molecule. However, long-read sequencing via Oxford Nanopore Technology is relatively shallowly sequenced, especially in the datasets used here which are single-cell-based optimisations (Berrens et al., 2021; Fan et al., 2020); therefore, combining deeply sequenced short-read RNA-seq of 2C embryos and 2CLCs with more precise long-read sequencing can identify novel transcripts with more certainty than with either of these approaches alone.

4.2.1 Transcript Structure Resolution.

Certain limitations are present in the approaches taken in this study to isolate MERVL-promoted transcript isoforms' exact structures and sequences. The transcriptomes for both long-and short-read RNA-seq are assembled with Stringtie (Pertea et al., 2015, 2016), a multi-purpose assembler. The usage of Stringtie for isoform assembly is undertaken by several studies (Fan et al., 2020; Fuentes et al., 2018; Modzelewski et al., 2022; Qiao et al., 2020; Zhang et al., 2023), including in the development of other tools to extract TE-promoted transcripts such as TEProf2 (Shah et al., 2023). However, two potential issues arise from this: (1) assessment of the first 10bp of each transcript's first exon, the approach taken to define a MERVL-promoted transcript, therefore relies heavily on Stringtie's thresholds of transcript structure identification and exact exon starts and ends, and (2) when merging transcriptomes, an unclear threshold is imposed on isoforms which are less abundant than others, and are removed from the final merged GTF (Appendix 2, figure 2) (Pertea et al., 2016).

4.2.2 Batch effects.

The two 2CLC/mESC short-read datasets (Hendrickson et al., (2017) and Ishiuchi et al., (2015)) utilise two different library preparation methods (polyA selection versus rRNA depletion) and two vastly different read lengths (150bp versus 50bp, though both paired-end sequencing). As a result, despite the Ishiuchi dataset being more deeply sequenced, more of the initially sequenced reads are lost during alignment and multimapper filtering. This may be why the overall transcriptional signal (e.g., Figure 3.4) is lower. The two long-read datasets follow a similar path, where the Berrens et al., dataset (CELLO-seq) is more stringently cleaned pre-sequencing, and has fewer rounds of PCR during the cDNA-PCR library preparation step, than the Fan et al., dataset (SCAN-seq). While both of these are sequenced on a MinION flow cell with the same chemistry nanopore (MinION R9.4.1 flow cell), the Fan et al., dataset is 3x more deeply sequenced (18 single cells versus 6 single cells). As a result, there are differences in the expression of MERVL elements (Figure 3.5) and in the transcript assembly (Figure 3.11), in particular in the abundance of transcripts sequenced (Table 3). However, it appears that roughly the alignment rate for sequenced reads for both long-read datasets is high (Appendix 1, figure 2).

The Fan et al., (2020) dataset (SCAN-seq long-read RNA-seq) is the major outlier in both reference and non-reference transcript retrieval, post transcriptome assembly. Examining the “class codes” defined by GFFcompare (Pertea et al., 2020) shows that this dataset has almost ten times as many “i” and “p” transcripts (denoting intronic and polymerase-run on false positive transcripts), twice as many “s” transcripts (mapping error transcripts), four times as many “o” transcripts (opposite strand transcripts, also likely mapping errors), and seven times as many “u” transcripts (undefinable transcripts) (table 4), as the Berrens et al., (2021)

dataset (CELLO-seq). This supports the aforementioned notion that the library preparation method in the SCAN-seq protocol retains fragments of RNA which should be cleaned. However, this method does retain many more possible true positives, hence this dataset is retained for further analysis (e.g., almost twice as many “j” and “k” class transcripts, which denote matching exons with extended transcripts, which may be alternative start sites).

Initially, I attempted to collate a list of MERVL-promoted isoforms regardless of their parent genes by comparing transcript coordinates only, but this did not account for alternative splicing and polyA tail length complexity between isoforms (data not shown). Due to the different sequencing technologies used (long- versus short-read sequencing) and differences in library preparation and sequencing metrics within sequencing technologies (library preparation and read length), transcript assembly was variable between datasets, where to the eye, one transcript may look identical, but vary by 1 base due to a minor difference in coverage. Therefore, I followed the method that Modzelewski et al., use, by comparing parent genes. Potentially, tools like the ENCODE4 project’s Cerberus (Reese et al., 2023) may be beneficial here, by assigning each alternative isoform a triplet code according to promoter choice, alternative splicing pattern, and polyA tail complexity. However, if this is assigned per dataset, it will result in a similar issue, where isoforms cannot be compared between datasets by identifier alone. A recent study by Torre et al., 2023 appears to address this issue by first assembling long-read transcriptomic data into a novel transcriptome with TALON, a long-read specific assembler (Wyman et al., 2020), and then aligning short-read data to the novel transcriptome. Potentially, this approach is more accurate in isoform assembly, and would still utilise support of short-read RNA-seq datasets.

4.2.3 CAGE-seq read length and mappability.

The sequencing length of the SLIC-CAGE reads is 50bp, sequenced with single-end sequencing. As a result, there is a very clear mappability issue, since MERVL elements vastly exceed this length, and so reads will often be aligned entirely within repetitive regions. While this wouldn't cause issues on a family-specific analysis, i.e., whether a read comes from one TE family or another, in this case, where it is important to know exactly which TE locus a read originates from, this causes issues. Newer protocols such as long-read CAGE-seq (Maeng et al., 2023) would be very beneficial here if the data for 2C embryos or 2CLCs were to be produced. In this study, I have mainly considered the results of single-best alignment approaches in read alignment (`Bowtie option -k 1`). This randomly aligns multi-mapping reads to 1 location at most. Since with non-cap-trapped RNA-seq (short-read RNA-seq for 2C embryos and 2CLCs, and long-read 2C data), it is clear that all MERVLs have transcriptional signal, I can assume that the lack of CAGE-seq signal is *not* due to transcription *not* initiating in MERVLs (i.e., transcription *is* initiating in MERVL elements), but technical limitations only.

4.3 Future perspectives.

4.3.1 How robust are genes to alternative promoters?

It would be meaningful to understand the impact of MERVL promoters on a genome-wide scale. Are all genes evolved to be robust to transposon-derived alternative promoters being inserted and functional, or are they plastic enough that alternative promoters only increase expression, and do not alter function? Are the known cases of TE-derived alternative promoters driving validated protein isoforms the rule, or are they the exception?

MERVL elements insert near genes associated with ZGA (Macfarlan et al., 2011), and given they have only a 5bp target site (Kruse et al., 2019), this is likely due to the changing chromatin organisation during ZGA, especially given that MERVL elements are accessible to transcription factor binding during ZGA (Choi et al., 2017; Hendrickson et al., 2017; De Iaco et al., 2020; Kruse et al., 2019; Maksakova et al., 2013; Schoorlemmer et al., 2014). Given that MERVL elements are still relatively young, and in the case of full MERVL elements, still replication-competent (Bénit et al., 1999; Costas, 2003), if an element were to insert near another gene, would it act as an alternative promoter driving a functional alternative protein isoform?

To address this, it would be valuable to examine how often the first exon of a gene has a splice acceptor site. This would impact both whether the first exon would be skipped when an alternative promoter is active. In theory, there is no reason for a first exon to contain a splice acceptor site, since there is no exon preceding the first exon which needs splicing *from*. MERVL elements contain splice donor sites (Franke et al., 2017). In the case of an alternative promoter, however, if there is no splice acceptor, then the first exon will be spliced out (Figure 3.14, top). This formed the initial hypothesis for this study, as is the case for *Cdk2ap1*, and later, *Prmt6*, with MT2B2 alternative promoters. Later, I found that transcription may initiate internally to a gene if the MERVL is inserted into an intron (Figure 3.14, bottom). In this case, the following exon is retained since it contains a splice acceptor site.

While the above gives insight into the transcript, to understand the protein, it would be important to then assess the Methionine start codon position, relative to exons, across the protein-coding genome. This would explain if having a MERVL alternative promoter upstream would have any impact on the peptide ultimately translated. If the ATG codon is not in the first exon, and neither is a splice acceptor site, then it is unlikely that an alternative promoter would change the ultimate

protein structure. The caveat here is then classical alternative splicing of downstream exons, where promoter choice *does* have an impact on exon choice and polyA tail length (Alfonso-Gonzalez et al., 2023; Pecci et al., 2001). This gives insight more into whether an alternative promoter results in changes to protein structure, and thus, function.

Finally, there may be cases when, if the MERVL is upstream of the gene, the intergenic region between the MERVL element and the gene may contain a splice acceptor site, by chance, resulting in an extended exon fused to the canonical first exon. If this region contains a start codon also, this might result in a frameshift mutation OR an in-frame elongated protein. Such hypothetical cases here are numerous.

4.3.2 What is the impact of MERVL-promoted protein isoforms?

While examination of the gene-level and transcript-level of MERVL-promoted isoforms is important, whether these isoforms have any crucial molecular function would depend on the protein output. Only through proving whether the MERVL-drive isoform is indispensable would the co-option of the MERVL promoter in the context of ZGA be proved. There is substantial evidence that transcription of MERVL elements is needed for correct ZGA (Rosa et al., 2024; Sakashita et al., 2023) and important for 2CLC cycling (Choi et al., 2017). Recently, there has also been proof that MERVL alternative promoters drive transcript isoforms involved in later embryogenesis (Honda et al., 2024; Modzelewski et al., 2021). To this end, further analysis of this study's isoform's structures and functions would be valuable.

4.3.2.1 *Computational Analysis.*

Recent studies have shown that at the protein level, maternally inherited proteins are stable until the blastocyst stage, with minor shifts at major ZGA and the

end of the 4C stage. This is much less variable than at the RNA level, where transcripts are much more variable (Zhang et al., 2023). To address whether MERVL-promoted transcripts are translated into mature proteins, and protein isoform-level analysis of publicly available data would be worthwhile, specifically, of the proteomic data in studies by Wang et al., 2010 and Zhang et al., 2023. Here, gene-level protein abundance data is easily available. To assess exact isoform level abundance, peptide sequences would need to be compared.

Further to this, tools like AlphaFold (Jumper et al., 2021) would be very informative, in helping compare protein structures, when certain amino acids are changed, whether or not this results in changes to protein domains. These could then be compared to publicly available protein structures on Uniprot (The Uniprot Consortium, 2019) or PDB (Berman et al., 2000) (both of which are either AlphaFold predicted or experimentally confirmed). Subsequently, tools like AlphaFoldMultimer (Liu et al., 2023) and MULTICOM (Liu et al., 2023) can be used to predict potential binding partners for the MERVL-driven isoforms, to see if these differ in any way from the reference isoform. Most recently, AlphaFold3 has added this utility, including predictions of nucleic acid, ligand, and fatty acid binding and association (Abramson et al., 2024) For example, if functional domains are changed then a change in binding partners would be expected. However, a change may also be predicted if a change in signalling peptide is noted.

4.3.2.2 Experimental validation.

Post-selection of candidate MERVL-promoted chimeric isoforms, initial validation could be completed via qPCR analysis of the splice junctions of the transcripts, to ensure that the exact transcript isoform structure is present in 2CLCs. Western blots could initially confirm the presence of the MERVL-promoted isoform

at the protein level. This will be completed in one of two ways; either antibodies may already exist for the candidate protein, or the protein may require tagging. If this is the case, the MERVL-promoted isoform would need to be tagged. This is possible with CRISPR-cas9 modification of the gene locus and the addition of a FLAG-tag to the transcript (resulting in a FLAG-tagged C-terminus of the protein), to select for the MERVL-promoted isoform (Xiang et al., 2019). The MERVL-promoted isoform will only be expressed and translated in 2CLCs since the MERVL promoter will be active only in 2CLCs and not the control mESCs. This would also allow me to perform co-immunoprecipitation experiments to examine the binding partners of the protein isoform. Immunofluorescence experiments would allow for the examination of the localisation of the protein in the cell.

Subsequent experiments would define the exact function of the alternative isoform using the FLAG-tagging protocol. This may be done through several experiments, such as:

- If the candidate is a transcription factor: the expectation is that the DNA-binding domain is modified/absent, or the effector domain is modified/absent. In the former, a lack of binding of the isoform will be noted in 2CLCs. If the effector domain is modified, then binding to cofactors will be modified. In both cases, this may be assessed with chromatin fractionation and ChIP-qPCR.
- If the candidate is a transmembrane protein: the expectation is that in any 2CLCs still reverted to, localisation of protein to membrane is decreased and measurable with cell fractionation experiments and immunofluorescence. Validation of protein function is dependent on the specific TM protein domain.
- If the candidate is a protein involved in aerobic respiration: the expectation is that in any 2CLCs still reverted to, oxygen usage is altered, whether this is

(Rodriguez-Terrones et al., 2020). This may be measured using Seahorse experiments. Metabolic changes are known to occur during early embryogenesis (Rodriguez-Terrones et al., 2020; Wang et al., 2021a; Zhao et al., 2021)

- If the candidate is a splicing-related protein: the lack of a functional splicing factor in 2CLCs may result in a decreased ability to revert to mESCs since it is known that splicing control is more tightly regulated post-ZGA (Revil et al., 2010; Tian et al., 2021; Wyatt et al., 2022). Measurement of the overall splicing efficiency of genes expressed at the exit from totipotency may be assessed by RT-qPCR.
- If the candidate is a kinase protein: co-IP will indicate binding targets for this protein. These targets will then not be phosphorylated. Therefore, the effects of any proposed signalling pathway should be absent or decreased. These may be assessed using western blots of the target protein, or kinase assays measuring ATP to ADP catalysis rate. This may also lead to measurable downstream effects on the cell, but the exact effects would depend on the pathway.

Knocking out the MERVL element of interest in the genome, using CRISPR-Cas9, would be appropriate here too. In this case, the specific MERVL would be absent and thus could not act as an alternative promoter for the candidate gene, resulting in no MERVL-driven alternative transcript. This may influence 2CLC cycling efficiency, as *Dux* overexpression and *Caf-1* knockdown do.

4.3.3 How does the MERVL LTR promoter function at the molecular level?

Though this analysis begins addressing the role of MERVLs as alternative promoters for genes in preimplantation mouse development, the question of the molecular biology of LTR promoters remains outstanding. Each LTR is (in theory)

identical for the purposes of ERV transposition (Benachenhou et al., 2009, 2013; Stocking and Kozak, 2008; Wells and Feschotte, 2020), and MT2_Mm LTR sequences are both strongly conserved and relatively newly expanded within the mouse genome (Elmer and Ferguson-Smith, 2020). Therefore, multiple sequence analyses of the LTRs and base-pair resolution dissection of the transcriptional start sites within MERVL elements using CAGE-seq data (Cvetesic et al., 2018, 2020b) are required, especially to compare signals between MERVLs which do not act as promoters to those that do. The expected result will be strong similarity between elements. It is known that there have been two expansions of MERVL elements within the mouse genome (Stocking and Kozak, 2008), so it is worth keeping in mind that solo LTR sequences may diverge from full MERVL LTRs slightly. Furthermore, clustering CAGE-seq tags to identify TSS exact locations should give insight into promoter grammar during both oocyte development and ZGA, with the data presented by Cvetesic et al., in 2018 and 2020; this might provide insight into the promoter dynamics of TE-derived promoters, and if MERVL promoters mimic native/host promoters during oocyte maturation and ZGA. This analysis may be extended to ALL MERVL elements (MT2A/B/C), since they are active at this time, not only MT2_Mms, though this will require re-curation of these elements in the C57Bl6/J and 129P2/Ola genomes.

Additionally, keystone studies from the late 1980s suggest a link between DNase hypersensitive sites and accessibility differences between the LTRs of avian retroviruses (Coffin, 1979; Coffin et al., 1997; Guntaka, 1993; Swain and Coffin, 1993). While ATAC-seq data produced more recently suggests little-to-no difference between LTRs (Kruse et al., 2019), analysis of nucleosome level studies at the LTRs of a complete MERVL may illuminate a distinct promoter choice between LTRs, and

by extension, why an LTR promoter is chosen over a ‘native’ promoter in the mouse genome at ZGA.

4.3.4 A note on MERVL LTRs as enhancers.

Finally, a very recent study has evidenced MERVL LTR enhancer function, using a reporter assay in HEK293 cells (Li et al., 2023). Aside from this, no other work to my knowledge has assessed this function, especially not in early embryogenesis, and even Li et al., did not discuss the implication of any enhancer-specific role of the MERVL LTR in totipotency. Furthermore, in my work, a faint CAGE-seq signal in both directions is visible during oocyte maturation and ZGA, which is a key characteristic of enhancer RNA transcription. It would be *incredibly* interesting to first confirm that the enhancer in a MERVL LTR is active during the oocyte-to-embryo transition and subsequently zygotic genome activation, and consider which genes are epigenetically regulated by MERVL enhancers, and the role of this further in ZGA.

5. References

- Abe, K., Yamamoto, R., Franke, V., Cao, M., Suzuki, Y., Suzuki, M.G., Vlahovicek, K., Svoboda, P., Schultz, R.M., and Aoki, F. (2015). The first murine zygotic transcription is promiscuous and uncoupled from splicing and 3' processing. *EMBO J.* **34**, 1523–1537.
- Abe, K., Schauer, T., and Torres-Padilla, M.E. (2022). Distinct patterns of RNA polymerase II and transcriptional elongation characterize mammalian genome activation. *Cell Rep.* **41**, 111865.
- Abe, K. ichiro, Funaya, S., Tsukioka, D., Kawamura, M., Suzuki, Y., Suzuki, M.G., Schultz, R.M., and Aoki, F. (2018). Minor zygotic gene activation is essential for mouse preimplantation development. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6780–E6788.
- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*.
- Alfonso-Gonzalez, C., Legnini, I., Holec, S., Arrigoni, L., Ozbulut, H.C., Mateos, F., Koppstein, D., Rybak-Wolf, A., Bönisch, U., Rajewsky, N., et al. (2023). Sites of transcription initiation drive mRNA isoform selection. *Cell* **186**, 1–18.
- Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020** *211* **21**, 1–16.
- Andrews, S. (2020). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.
- Arkhipova, I.R. (2017). Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob.*

DNA 2017 81 8, 1–14.

Arkhipova, I.R. (2018). Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. *Mol. Biol. Evol.* 35, 1332–1337.

Asami, M., Lam, B.Y.H., Hoffmann, M., Suzuki, T., Lu, X., Yoshida, N., Ma, M.K., Rainbow, K., Gužvić, M., VerMilyea, M.D., et al. (2023). A program of successive gene expression in mouse one-cell embryos. *Cell Rep.* 42.

Aw, J.G.A., Lim, S.W., Wang, J.X., Lambert, F.R.P., Tan, W.T., Shen, Y., Zhang, Y., Kaewsapsak, P., Li, C., Ng, S.B., et al. (2020). Determination of isoform-specific RNA structure with nanopore long reads. *Nat. Biotechnol.* 2020 393 39, 336–346.

Baker, C.L., and Pera, M.F. (2018). Capturing Totipotent Stem Cells. *Cell Stem Cell* 22, 25–34.

Banville, D., and Boie, Y. (1989). Retroviral long terminal repeat is the promoter of the gene encoding the tumor-associated calcium-binding protein oncomodulin in the rat. *J. Mol. Biol.* 207, 481–490.

Benachenhou, F., Jern, P., Oja, M., Sperber, G., Blikstad, V., Somervuo, P., Kaski, S., and Blomberg, J. (2009). Evolutionary Conservation of Orthoretroviral Long Terminal Repeats (LTRs) and ab initio Detection of Single LTRs in Genomic Data. *PLoS One* 4, e5179.

Benachenhou, F., Sperber, G.O., Bongcam-Rudloff, E., Andersson, G., Boeke, J.D., and Blomberg, J. (2013). Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). *Mob. DNA* 4, 5.

Bénit, L., Lallemand, J.-B., Casella, J.-F., Philippe, H., and Heidmann, T. (1999). ERV-L Elements: a Family of Endogenous Retrovirus-Like Elements Active throughout the Evolution of Mammals. *J. Virol.* 73, 3301–3308.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids*

Res. 28, 235–242.

Berrens, R. V., Yang, A., Laumer, C.E., Lun, A.T.L.L., Bieberich, F., Law, C.-T.T., Lan, G., Imaz, M., Bowness, J.S., Brockdorff, N., et al. (2021). Locus-specific expression of transposable elements in single cells with CELLO-seq. *Nat. Biotechnol.* 40, 546–554.

Biasini, A., and Marques, A.C. (2020). A Protocol for Transcriptome-Wide Inference of RNA Metabolic Rates in Mouse Embryonic Stem Cells. *Front. Cell Dev. Biol.* 8, 97.

Bonnet-Garnier, A., Feuerstein, P., Chebrou, M., Fleurot, R., Jan, H.U., Debey, P., and Beaujean, N. (2012). Genome organization and epigenetic marks in mouse germinal vesicle oocytes. *Int. J. Dev. Biol.* 56, 877–887.

Bouniol, C., Nguyen, E., and Debey, P. (1995). Endogenous Transcription Occurs at the 1-Cell Stage in the Mouse Embryo. *Exp. Cell Res.* 218, 57–62.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. *Genome Biol.* 2018 19, 1–12.

Brady, T., Lee, Y.N., Ronen, K., Malani, N., Berry, C.C., Bieniasz, P.D., and Bushman, F.D. (2009). Integration target site selection by a resurrected human endogenous retrovirus. 633–642.

Britten, R.J. (1997). Mobile elements inserted in the distant past have taken on important functions. *Gene* 205, 177–182.

Chen, F., Zhang, W., Xie, D., Gao, T., Dong, Z., and Lu, X. (2020). Histone chaperone FACT represses retrotransposon MERVL and MERVL-derived cryptic promoters. *Nucleic Acids Res.* 48, 10211–10225.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890.

Chen, Y., Davidson, N.M., Wan, Y.K., Patel, H., Yao, F., Low, H.M., Hendra, C., Watten, L., Sim, A., Sawyer, C., et al. (2021). A systematic benchmark of Nanopore

long read RNA sequencing for transcript level analysis in human cell lines. *BioRxiv* 1, 2021.04.21.440736.

Choi, Y.J., Lin, C.-P., Risso, D., Chen, S., Kim, T.A., Tan, M.H., Li, J.B., Wu, Y., Chen, C., Xuan, Z., et al. (2017). Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells. *Science* (80-.). 355, eaag1927.

Choudhary, M.N.K., Quaid, K., Xing, X., Schmidt, H., and Wang, T. (2023).

Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes. *Nat. Commun.* 2023 141 14, 1–12.

Cockburn, K., and Rossant, J. (2010). Making the blastocyst: Lessons from the mouse. *J. Clin. Invest.* 120, 995–1003.

Coffin, J.M. (1979). Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J. Gen. Virol.* 42, 1–26.

Coffin, J.M., Hughes, S.H., and Varmus, H.E. (1997). *Retroviruses*.

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

Cotto, K.C., Feng, Y.-Y., Ramu, A., Richters, M., Freshour, S.L., Skidmore, Z.L., Xia, H., McMichael, J.F., Kunisaki, J., Campbell, K.M., et al. (2023). Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat. Commun.* 2023 141 14, 1–18.

Cvetesic, N., Leitch, H.G., Borkowska, M., Müller, F., Carninci, P., Hajkova, P., and Lenhard, B. (2018). SLIC-CAGE: High-resolution transcription start site mapping using nanogram-levels of total RNA. *Genome Res.* 28, 1931–1942.

Cvetesic, N., Borkowska, M., Hatanaka, Y., Yu, C., Vincent, S.D., Müller, F., Tora, L., Leitch, H.G., Hajkova, P., and Lenhard, B. (2020a). Global regulatory transitions at core promoters demarcate the mammalian germline cycle. *BioRxiv* 2020.10.30.361865.

Cvetesic, N., Borkowska, M., Hatanaka, Y., Leitch, H.G., Müller, F., Yu, C., Vincent, S.D., Tora, L., Hajkova, P., Lenhard, B., et al. (2020b). Global regulatory transitions at core promoters demarcate the mammalian germline cycle. *BioRxiv* 2020.10.30.361865.

Denker, H.-W.W. (2016). Self-Organization of Stem Cell Colonies and of Early Mammalian Embryos: Recent Experiments Shed New Light on the Role of Autonomy vs. External Instructions in Basic Body Plan Development. *Cells* 5, 1–15.

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet.* 34, 666–681.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15.

Dyce, J., George, M., Goodall, H., and Fleming, T.P. (1987). Do trophectoderm and inner cell mass cells in the mouse blastocyst maintain discrete lineages. *Development* 100, 685–698.

Eckersley-Maslin, M., Alda-Catalinas, C., Blotenburg, M., Kreibich, E., Krueger, C., and Reik, W. (2019). Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes Dev.* 33, 194–208.

Eckersley-Maslin, M.A., Svensson, V., Krueger, C., Stubbs, T.M., Giehr, P., Krueger, F., Miragaia, R.J., Kyriakopoulos, C., Berrens, R. V., Milagre, I., et al. (2016). MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Rep.* 17, 179–192.

Eckersley-Maslin, M.A., Alda-Catalinas, C., and Reik, W. (2018). Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nat. Rev. Mol. Cell Biol.* 2018 197 19, 436–450.

Eickbush, T.H., and Jamburuthugoda, V.K. (2008). The diversity of retrotransposons

and the properties of their reverse transcriptases. *Virus Res.* **134**, 221–234.

Eleftheriou, K., Peter, A., Fedorenko, I., Schmidt, K., Wossidlo, M., and Arand, J. (2022). A transition phase in late mouse oogenesis impacts DNA methylation of the early embryo. *Commun. Biol.* **2022** 51 5, 1–11.

Elmer, J.L., and Ferguson-Smith, A.C. (2020). Strain-specific epigenetic regulation of endogenous retroviruses: The role of Trans-acting modifiers. *Viruses* **12**, 1–22.

Evsikov, A. V., De Vries, W.N., Peaston, A.E., Radford, E.E., Fancher, K.S., Chen, F.H., Blake, J.A., Bult, C.J., Latham, K.E., Solter, D., et al. (2004). Systems biology of the 2-cell mouse embryo. *Cytogenet. Genome Res.* **105**, 240–250.

Fan, X., Tang, D., Liao, Y., Li, P., Zhang, Y., Wang, M., Liang, F., Wang, X., Gao, Y., Wen, L., et al. (2020). Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. *PLOS Biol.* **18**, e3001017.

Ferraj, A., Audano, P.A., Balachandran, P., Czechanski, A., Flores, J.I., Radecki, A.A., Mosur, V., Gordon, D.S., Walawalkar, I.A., Eichler, E.E., et al. (2023). Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements. *Cell Genomics* **3**, 100291.

Feschotte, C., and Pritham, E.J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genomics Hum. Genet.* **41**, 331–368.

Fierro-González, J.C., White, M.D., Silva, J.C., and Plachta, N. (2013). Cadherin-dependent filopodia control preimplantation embryo compaction. *Nat. Cell Biol.* **15**, 1424–1433.

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.* **27**, 1384–1394.

Fu, X., Djekidel, M.N., and Zhang, Y. (2020). A transcriptional roadmap for 2C-like–

to-pluripotent state transition. *Sci. Adv.* 6, 5181–5210.

Fuentes, D.R., Swigut, T., and Wysocka, J. (2018). Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife* 7, 1–29.

Fueyo, R., Judd, J., Feschotte, C., and Wysocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. 23.

Fujimori, T. (2010). Preimplantation development of mouse: A view from cellular behavior. *Dev. Growth Differ.* 52, 253–262.

Gagnier, L., Belancio, V.P., and Mager, D.L. (2019). Mouse germ line mutations due to retrotransposon insertions. *Mob. DNA* 2019 101 10, 1–22.

Garland, W., Müller, I., Wu, M., Schmid, M., Imamura, K., Rib, L., Sandelin, A., Helin, K., and Jensen, T.H. (2022). Chromatin modifier HUSH co-operates with RNA decay factor NEXT to restrict transposable element expression. *Mol. Cell* 82, 1–17.

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* (80-.). 378, 1305–1315.

Goolam, M., Scialdone, A., Graham, S.J.L., MacAulay, I.C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J.C., and Zernicka-Goetz, M. (2016). Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* 165, 61–74.

Gould, S.J., and Vrba, E.S. (1982). Exaptation-A Missing Term in the Science of Form. *Paleobiology* 8, 4–15.

Guntaka, R. V. (1993). Transcription termination and polyadenylation in retroviruses. *Microbiol. Rev.* 57, 511–521.

Guo, Y., Kitano, T., Inoue, K., Murano, K., Hirose, M., Li, T.D., Sakashita, A., Ishizu, H., Ogonuki, N., Matoba, S., et al. (2022). Obox4 secures zygotic genome activation upon loss of Dux. *BioRxiv* 2022.07.04.498763.

Haberle, V., and Lenhard, B. (2016). Promoter architectures and developmental gene regulation. *Semin. Cell Dev. Biol.* 57, 11–23.

Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., Gehrig, J., Dong, X., Akalin, A., Suzuki, A.M., et al. (2014). Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* 507, 381–385.

Haberle, V., Forrest, A.R.R., Hayashizaki, Y., Carninci, P., and Lenhard, B. (2015). CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* 43, e51.

Hackett, J.A., Kobayashi, T., Dietmann, S., and Surani, M.A. (2017). Activation of Lineage Regulators and Transposable Elements across a Pluripotent Spectrum. *Stem Cell Reports* 8, 1645–1658.

Hendrickson, P.G., Doráis, J.A., Grow, E.J., Whiddon, J.L., Lim, J.W., Wike, C.L., Weaver, B.D., Pflueger, C., Emery, B.R., Wilcox, A.L., 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.

Herman, S.A., and Coffin, J.M. (1986). Differential transcription from the long terminal repeats of integrated avian leukosis virus DNA. *J. Virol.* 60, 497–505.

Honda, S., Hatamura, M., Kunimoto, Y., Ikeda, S., and Minami, N. (2023). Chimeric PRMT6 protein produced by an endogenous retrovirus promoter regulates cell fate decision in mouse preimplantation embryos. *BioRxiv* 2022.08.23.505038.

Honda, S., Hatamura, M., Kunimoto, Y., Ikeda, S., and Minami, N. (2024). Chimeric PRMT6 protein produced by an endogenous retrovirus promoter regulates cell fate decision in mouse preimplantation embryos. *Biol. Reprod.* 1–13.

Hu, Z., Tan, D.E.K., Chia, G., Tan, H., Leong, H.F., Chen, B.J., Lau, M.S., Tan, K.Y.S., Bi, X., Yang, D., et al. (2020). Maternal factor NELFA drives a 2C-like state in mouse embryonic stem cells. *Nat. Cell Biol.* 2020 222 22, 175–186.

Huang, Y., Kim, J.K., Do, D.V., Lee, C., Penfold, C.A., Zylitz, J.J., Marioni, J.C., Hackett, J.A., and Surani, M.A. (2017). Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *Elife* 6.

De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Publ. Gr.* 49.

De Iaco, A., Coudray, A., Duc, J., Trono, D., Iaco, A. De, Coudray, A., Duc, J., and Trono, D. (2019). DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO Rep.* 20, 1–10.

De Iaco, A., Verp, S., Offner, S., Grun, D., and Trono, D. (2020). DUX is a non-essential synchronizer of zygotic genome activation. *Dev.* 147, 1–6.

Imbeault, M., Helleboid, P.Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nat.* 2017 5437646 543, 550–554.

Ing-Simmons, E., Rigau, M., and Vaquerizas, J.M. (2022). Emerging mechanisms and dynamics of three-dimensional genome organisation at zygotic genome activation. *Curr. Opin. Cell Biol.* 74, 37–46.

Iouranova, A., Grun, D., Rossy, T., Duc, J., Coudray, A., Imbeault, M., de Tribolet-Hardy, J., Turelli, P., Persat, A., and Trono, D. (2022). KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. *Mob. DNA* 13, 4.

Ishiuchi, T., Enriquez-Gasca, R., Mizutani, E., Boškoviä, A., Ziegler-Birling, C., Rodriguez-Terrones, D., Wakayama, T., Vaquerizas, J.M., and Torres-Padilla, M.E.

(2015). Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat. Struct. Mol. Biol.* **22**, 662–671.

Iturbide, A., and Torres-Padilla, M.-E.E. (2020). A cell in hand is worth two in the embryo: recent advances in 2-cell like cell reprogramming. *Curr. Opin. Genet. Dev.* **64**, 26–30.

Iturbide, A., Ruiz Tejeda Segura, M.L., Noll, C., Schorpp, K., Rothenaigner, I., Ruiz-Morales, E.R., Lubatti, G., Agami, A., Hadian, K., Scialdone, A., et al. (2021). Retinoic acid signaling is critical during the totipotency window in early mammalian development. *Nat. Struct. Mol. Biol.* **2021 286 28**, 521–532.

Jachowicz, J.W., Bing, X., Pontabry, J., Bošković, A., Rando, O.J., and Torres-Padilla, M.E. (2017). LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat. Genet.* **2017 4910 49**, 1502–1510.

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.

Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TETranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599.

Johnson, J.M., Verkerke, A.R.P., Maschek, J.A., Ferrara, P.J., Lin, C. Te, Kew, K.A., Neuffer, P.D., Lodhi, I.J., Cox, J.E., and Funai, K. (2020). Alternative splicing of UCP1 by non-cell-autonomous action of PEMT. *Mol. Metab.* **31**, 55–66.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly

accurate protein structure prediction with AlphaFold. *Nat.* 2021 5967873 596, 583–589.

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45.

Karlic, R., Ganesh, S., Franke, V., Svobodova, E., Urbanova, J., Suzuki, Y., Aoki, F., Vlahovicek, K., and Svoboda, P. (2017). Long non-coding RNA exchange during the oocyte-to-embryo transition in mice. *DNA Res.* 24, 129–141.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nat.* 2011 4777364 477, 289–294.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Res.* 12, 996–1006.

Kigami, D., Minami, N., Takayama, H., and Imai, H. (2003). MuERV-L Is One of the Earliest Transcribed Genes in Mouse One-Cell Embryos¹. *Biol. Reprod.* 68, 651–654.

Kojima, Y., Tam, O.H., and Tam, P.P.L. (2014). Timing of developmental events in the early mouse embryo. *Semin. Cell Dev. Biol.* 34, 65–75.

Krause, M., Niazi, A.M., Labun, K., Torres Cleuren, Y.N., Müller, F.S., and Valen, E. (2019). tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. *RNA* 25, 1229–1241.

Kruse, K., Díaz, N., Enriquez-Gasca, R., Gaume, X., Torres-Padilla, M.-E.E., and Vaquerizas, J.M. (2019). Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *BioRxiv* 523712.

Kuo, R.I., Cheng, Y., Zhang, R., Brown, J.W.S., Smith, J., Archibald, A.L., and Burt, D.W. (2020). Illuminating the dark side of the human transcriptome with long read

transcript sequencing. *BMC Genomics* 21, 1–22.

van de Lagemaat, L.N., Landry, J.-R., Mager, D.L., and Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *TRENDS Genet.* 19, 530–537.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.

Lee, K., Cho, K., Morey, R., and Cook-Andersen, H. (2024). An extended wave of global mRNA deadenylation sets up a switch in translation regulation across the mammalian oocyte-to-embryo transition. *Cell Rep.* 43, 113710.

Lee, S., Liu, B., Lee, S., Huang, S.X., Shen, B., and Qian, S.B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* 109, E2424–E2432.

Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G., and Smith, D.B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* 46, D708–D717.

Leger, A., Amaral, P.P., Pandolfini, L., Capitanchik, C., Capraro, F., Miano, V., Migliori, V., Toolan-Kerr, P., Sideri, T., Enright, A.J., et al. (2021). RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.* 2021 121 12, 1–17.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and

SAMtools. *Bioinformatics* 25, 2078–2079.

Li, Z., Xu, H., Li, J., Xu, X., Wang, J., Wu, D., Zhang, J., Liu, J., Xue, Z., Zhan, G., et al. (2023). Selective binding of retrotransposons by ZFP352 facilitates the timely dissolution of totipotency network. *Nat. Commun.* 14, 3646.

Liu, J., Guo, Z., Wu, T., Roy, R.S., Quadir, F., Chen, C., and Cheng, J. (2023). Enhancing alphafold-multimer-based protein complex structure prediction with MULTICOM in CASP15. *Commun. Biol.* 6, 1–14.

Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Grüning, B., Ramírez, F., and Manke, T. (2021). pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* 37, 422–423.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., et al. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes Dev.* 25, 594–607.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* 487, 57–63.

Macia, A., Muñoz-Lopez, M., Cortes, J.L., Hastings, R.K., Morell, S., Lucena-Aguilar, G., Marchal, J.A., Badge, R.M., and Garcia-Perez, J.L. (2011). Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells. *Mol. Cell. Biol.* 31, 300–316.

Maeng, J.H., Jang, H.J., Du, A.Y., Tzeng, S.-C., and Wang, T. (2023). Using long-read CAGE sequencing to profile cryptic-promoter-derived transcripts and their contribution to the immunopeptidome. *Genome Res.* 33, 2143–2155.

Magiorkinis, G., Gifford, R.J., Katzourakis, A., De Ranter, J., and Belshaw, R. (2012). Env-less endogenous retroviruses are genomic superspreaders. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 7385–7390.

Maksakova, I.A., Thompson, P.J., Goyal, P., Jones, S.J.M., Singh, P.B., Karimi, M.M., and Lorincz, M.C. (2013). Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics and Chromatin* *6*, 1–16.

Malik, H.S., and Eickbush, T.H. (2001). Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses. *Genome Res.* *11*, 1187–1197.

Margolin, G., Khil, P.P., Kim, J., Bellani, M.A., and Camerini-Otero, R.D. (2014). Integrated transcriptome analysis of mouse spermatogenesis. *BMC Genomics* *15*, 1–19.

Martens, J.H.A., O’Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J.* *24*, 800–812.

McCarthy, E.M., and McDonald, J.F. (2004). Long terminal repeat retrotransposons of *Mus musculus*. *Genome Biol.* *5*, R14.

McClintock, B. (1950). THE ORIGIN AND BEHAVIOR OF MUTABLE LOCI IN MAIZE. *Proc. Natl. Acad. Sci. USA* *36*, 344–355.

Medstrand, P., Van De Lagemaat, L.N., and Mager, D.L. (2002). Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes. *Genome Res.* *12*, 1483–1495.

Modzelewski, A.J., 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.

Modzelewski, A.J., Gan Chong, J., Wang, T., and He, L. (2022). Mammalian genome innovation through transposon domestication. *Nat. Cell Biol.* 2022 249 24, 1332–1340.

Nellåker, C., Keane, T.M., Yalcin, B., Wong, K., Agam, A., Belgard, T.G., Flint, J., Adams, D.J., Frankel, W.N., and Ponting, C.P. (2012). The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol.* 13.

Nit, L.B., De Parseval, N., Casella, J.-F.O., Callebaut, I., Agne`, A., Cordonnier, A., And, †, and Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J. Virol.* 71, 5652–5657.

Olbrich, T., Vega-Sendino, M., Tillo, D., Wu, W., Zolnerowich, N., Pavani, R., Tran, A.D., Domingo, C.N., Franco, M., Markiewicz-Potoczny, M., et al. (2021). CTCF is a barrier for 2C-like reprogramming. *Nat. Commun.* 12, 1–12.

Olsen, H.E., Brooks, A.N., Akeson, M., and Timp, W. (2016). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* 16, 1297–1305.

Ondiřová, M., Oakey, R.J., and Walsh, C.P. (2020). Is imprinting the result of “friendly fire” by the host defense system? *PLOS Genet.* 16, e1008599.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 2017 144 14, 417–419.

Peaston, A.E. (2008). Retrotransposons of Vertebrates. *Encycl. Virol.* 436–445.

Peaston, A.E., Evsikov, A. V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* 7, 597–606.

Pecci, A., Viegas, L.R., Baraňao, J.L., and Beato, M. (2001). Promoter Choice

Influences Alternative Splicing and Determines the Balance of Isoforms Expressed from the Mouse *bcl-X* Gene *. *J. Biol. Chem.* 276, 21062–21069.

Percharde, M., Lin, C.J., Yin, Y., Guan, J., Peixoto, G.A., Bulut-Karslioglu, A., Biechele, S., Huang, B., Shen, X., and Ramalho-Santos, M. (2018). A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* 174, 391-405.e19.

Perry, A.C.F., Asami, M., Lam, B.Y.H., and Yeo, G.S.H. (2023). The initiation of mammalian embryonic transcription: to begin at the beginning. *Trends Cell Biol.* 33, 365–373.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 2015 333 33, 290–295.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 2016 119 11, 1650–1667.

Pertea, M., Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research* 9, 304.

Puri, D., Koschorz, B., Engist, B., Onishi-Seebacher, M., Ryan, D., Soujanya, M., and Montavon, T. (2021). Foxd3 controls heterochromatin-mediated repression of repeat elements and 2-cell state transcription. *EMBO Rep.* 22.

Qiao, Y., Ren, C., Huang, S., Yuan, J., Liu, X., Fan, J., Lin, J., Wu, S., Chen, Q., Bo, X., et al. (2020). High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. *Nat. Commun.* 2020 111 11, 1–13.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for

deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165.

Razo-Mejia, M., Barnes, S.L., Belliveau, N.M., Chure, G., Einav, T., Lewis, M., and Phillips, R. (2018). Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. *Cell Syst.* **6**, 456-469.e10.

Reese, F., Williams, B., Balderrama-Gutierrez, G., Wyman, D., Çelik, M.H., Rebboah, E., Rezaie, N., Trout, D., Razavi-Mohseni, M., Jiang, Y., et al. (2023). The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *BioRxiv* **18**, 2023.05.15.540865.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nat.* **2007** 4477143 **447**, 425–432.

Revil, T., Gaffney, D., Dias, C., Majewski, J., and Jerome-Majewska, L.A. (2010). Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* **11**, 1–17.

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

Rodriguez-Terrones, D., Gaume, X., Ishiuchi, T., Weiss, A., Kopp, A., Kruse, K., Penning, A., Vaquerizas, J.M., Brino, L., and 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., Hartleben, G., Gaume, X., Eid, A., Guthmann, M., Iturbide, A., and Torres-Padilla, M. (2020). A distinct metabolic state arises during the emergence of 2-cell-like cells. *EMBO Rep.* **21**.

Rosa, S. de la, Rigual, M. del M., Vargiu, P., Ortega, S., and Djouder, N. (2024). Endogenous retroviruses shape pluripotency specification in mouse embryos. *Sci. Adv.* **10**.

Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P. V., Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. *Nat.* 2010 4637278 463, 237–240.

Sakashita, A., Kitano, T., Ishizu, H., Guo, Y., Masuda, H., Ariura, M., Murano, K., and Siomi, H. (2023). Transcription of MERVL retrotransposons is required for preimplantation embryo development. *Nat. Genet.* 1–12.

Sartorelli, V., and Lauberth, S.M. (2020). Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.* 27, 521–528.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell* 148, 335–348.

Schoorlemmer, J., Pérez-Palacios, R., Climent, M., Guallar, D., and Muniesa, P. (2014). Regulation of mouse retroelement MuERV-L/MERVL expression by REX1 and epigenetic control of stem cell potency. *Front. Oncol.* 4, 14.

Schultz, R.M., Letourneau, G.E., and Wassarman, P.M. (1979). Program of early development in the mammal: Changes in the patterns and absolute rates of tubulin and total protein synthesis during oocyte growth in the mouse. *Dev. Biol.* 73, 120–133.

Schultz, R.M., Stein, P., and Svoboda, P. (2018). The oocyte-to-embryo transition in mouse: past, present, and future. *Biol. Reprod.* 99, 160–174.

Shah, N.M., Jang, H.J., Liang, Y., Maeng, J.H., Tzeng, S.-C., Wu, A., Basri, N.L., Qu, X., Fan, C., Li, A., et al. (2023). Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nat. Genet.* 2023 1–9.

Shen, H., Yang, M., Li, S., Zhang, J., Peng, B., Wang, C., Chang, Z., Ong, J., and Du,

P. (2021). Mouse totipotent stem cells captured and maintained through spliceosomal repression. *Cell* *184*, 2843–2859.e20.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* *100*, 15776–15781.

De Souza, F.S.J., Franchini, L.F., and Rubinstein, M. (2013). Exaptation of Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always Strong? *Mol. Biol. Evol.* *30*, 1239.

Srinivasan, R., Nady, N., Arora, N., Hsieh, L.J., Swigut, T., Narlikar, G.J., Wossidlo, M., and Wysocka, J. (2020). Zscan4 binds nucleosomal microsatellite DNA and protects mouse two-cell embryos from DNA damage. *Sci. Adv.* *6*.

Sripathy, S.P., Stevens, J., and Schultz, D.C. (2006). The KAP1 Corepressor Functions To Coordinate the Assembly of De Novo HP1-Demarcated Microenvironments of Heterochromatin Required for KRAB Zinc Finger Protein-Mediated Transcriptional Repression. *Mol. Cell. Biol.* *26*, 8623–8638.

Stocking, C., and Kozak, C.A. (2008). Murine endogenous retroviruses. *Cell. Mol. Life Sci.* *65*, 3383.

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* *12*, 1–14.

Subrini, J., and Turner, J. (2021). Y chromosome functions in mammalian spermatogenesis. *Elife* *10*, 1–20.

Svoboda, P. (2018). Mammalian zygotic genome activation. *Semin. Cell Dev. Biol.* *84*, 118–126.

Svoboda, P., Franke, V., and Schultz, R.M. (2015). Sculpting the Transcriptome

During the Oocyte-to-Embryo Transition in Mouse. *Curr. Top. Dev. Biol.* *113*, 305–349.

Swain, A., and Coffin, J.M. (1993). Influence of sequences in the long terminal repeat and flanking cell DNA on polyadenylation of retroviral transcripts. *J. Virol.* *67*, 6265–6269.

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* *2012 73 7*, 542–561.

The Uniprot Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* *47*, D506–D515.

Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R., and Sandelin, A. (2019). CAGEfightR: Analysis of 5′-end data using R/Bioconductor. *BMC Bioinformatics* *20*, 1–13.

Tian, G.G., Li, J., and Wu, J. (2020). Alternative splicing signatures in preimplantation embryo development. *Cell Biosci.* *10*, 1–10.

Tian, L., Jabbari, J.S., Thijssen, R., Gouil, Q., Amarasinghe, S.L., Voogd, O., Kariyawasam, H., Du, M.R.M., Schuster, J., Wang, C., et al. (2021). Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol.* *22*, 1–24.

Torre, D., Francoeur, N.J., Kalma, Y., Gross Carmel, I., Melo, B.S., Deikus, G., Allette, K., Flohr, R., Fridrikh, M., Vlachos, K., et al. (2023). Isoform-resolved transcriptome of the human preimplantation embryo. *Nat. Commun.* *2023 141 14*, 1–23.

Torres-Padilla, M.E. (2020). On transposons and totipotency. *Philos. Trans. R. Soc. B* *375*.

Torres-Padilla, M.-E., Parfitt, D.-E., Kouzarides, T., and Zernicka-Goetz, M. (2007).

Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* **445**, 214–218.

Vastenhouw, N.L., Cao, W.X., and Lipshitz, H.D. (2019). The maternal-to-zygotic transition revisited. *Development* **146**.

Vuoristo, S., Bhagat, S., Hydén-Granskog, C., Yoshihara, M., Gawriyski, L., Jouhilahti, E.-M., Ranga, V., Tamirat, M., Huhtala, M., Kirjanov, I., et al. (2022). DUX4 is a multifunctional factor priming human embryonic genome activation. *IScience* **25**, 104137.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74.

Wang, S., Kou, Z., Jing, Z., Zhang, Y., Guo, X., Dong, M., Wilmut, I., and Gao, S. (2010). Proteome of mouse oocytes at different developmental stages. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 17639–17644.

Wang, Y., Na, Q., Li, X., Tee, W.-W., Wu, B., Bao, S., and State Key, T. (2021a). Retinoic acid induces NELFA-mediated 2C-like state of mouse embryonic stem cells associates with epigenetic modifications and metabolic processes in chemically defined media. *Cell Prolif.* **54**.

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K.F. (2021b). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **2021** 3911 **39**, 1348–1365.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2008** 101 **10**, 57–63.

Wells, J.N., and Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. <https://doi.org/10.1146/annurev-genet-040620-022145> **54**, 539–561.

Wick, R.R., Judd, L.M., and Holt, K.E. (2019). Performance of neural network

basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* 20, 1–10.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., 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.* 2007 812 8, 973–982.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov.* 2, 100141.

Wyatt, C.D.R., Pernaute, B., Gohr, A., Miret-Cuesta, M., Goyeneche, L., Rovira, Q., Salzer, M.C., Boke, E., Bogdanovic, O., Bonnal, S., et al. (2022). A developmentally programmed splicing failure contributes to DNA damage response attenuation during mammalian zygotic genome activation. *Sci. Adv.* 8, 4935.

Wyman, D., Balderrama-Gutierrez, G., Reese, F., Jiang, S., Rahmanian, S., Forner, S., Matheos, D., Zeng, W., Williams, B., Trout, D., et al. (2020). A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *BioRxiv* 672931.

Xiang, X., Li, C., Chen, X., Dou, H., Li, Y., Zhang, X., and Luo, Y. (2019). CRISPR/Cas9-Mediated Gene Tagging: A Step-by-Step Protocol. In *Proceedings of the National Academy of Sciences*, pp. 255–269.

Xie, S.Q., Leeke, B.J., Whilding, C., Wagner, R.T., Garcia-Llagostera, F., Low, Y., Chammas, P., Cheung, N.T.-F., Dormann, D., McManus, M.T., et al. (2022). Nucleolar-based Dux repression is essential for embryonic two-cell stage exit. *Genes Dev.* 36, 331–347.

Yang, F., Huang, X., Zang, R., Chen, J., Fidalgo, M., Sanchez-Priego, C., Yang, J., Caichen, A., Ma, F., Macfarlan, T., et al. (2020). DUX-miR-344-ZMYM2-Mediated Activation of MERVL LTRs Induces a Totipotent 2C-like State. *Cell Stem Cell* 26, 234-250.e7.

Yu, C., Ji, S.Y., Dang, Y.J., Sha, Q.Q., Yuan, Y.F., Zhou, J.J., Yan, L.Y., Qiao, J., Tang, F., and Fan, H.Y. (2016). Oocyte-expressed yes-associated protein is a key activator of the early zygotic genome in mouse. *Cell Res.* 2016 263 26, 275–287.

Yu, C., Cveticic, 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, H., Ji, S., Zhang, K., Chen, Y., Ming, J., Kong, F., Wang, L., Wang, S., Zou, Z., Xiong, Z., et al. (2023). Stable maternal proteins underlie distinct transcriptome, translatoe, and proteome reprogramming during mouse oocyte-to-embryo transition. *Genome Biol.* 2023 241 24, 1–25.

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

Zhao, J., Yao, K., Yu, H., Zhang, L., Xu, Y., Chen, L., Sun, Z., Zhu, Y., Zhang, C., Qian, Y., et al. (2021). Metabolic remodelling during early mouse embryo development. *Nat. Metab.* 2021 310 3, 1372–1384.

Zhou, L., and Dean, J. (2015). Reprogramming the genome to totipotency in mouse embryos. *Trends Cell Biol.* 25, 82–91.

Zhou, Q., Zhang, F., Kerbl-Knapp, J., Korbelius, M., Kuentzel, K.B., Vujić, N., Akhmetshina, A., Hörl, G., Paar, M., Steyrer, E., et al. (2022). Phosphatidylethanolamine N-Methyltransferase Knockout Modulates Metabolic Changes in Aging Mice. *Biomolecules* 12.

Appendix 1: RNA-seq QC

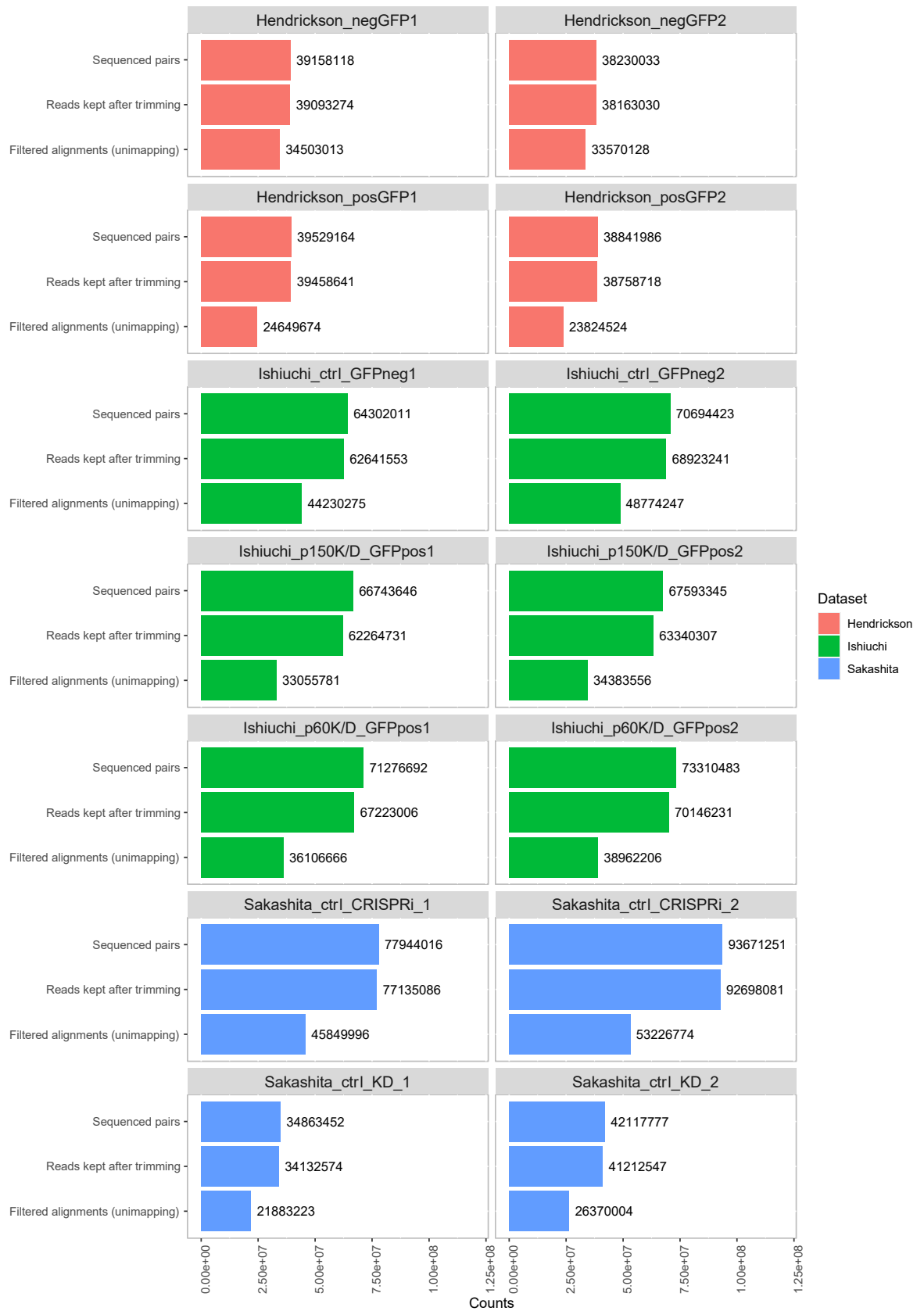


Figure 1: Quality control for short-read RNA-seq data. Total pairs sequenced, total reads aligned, and total reads post QC filtering and selecting for uniquely mapping reads counted. Data analysed from (Hendrickson et al., 2017; Ishiuchi et al., 2015; Sakashita et al., 2023).

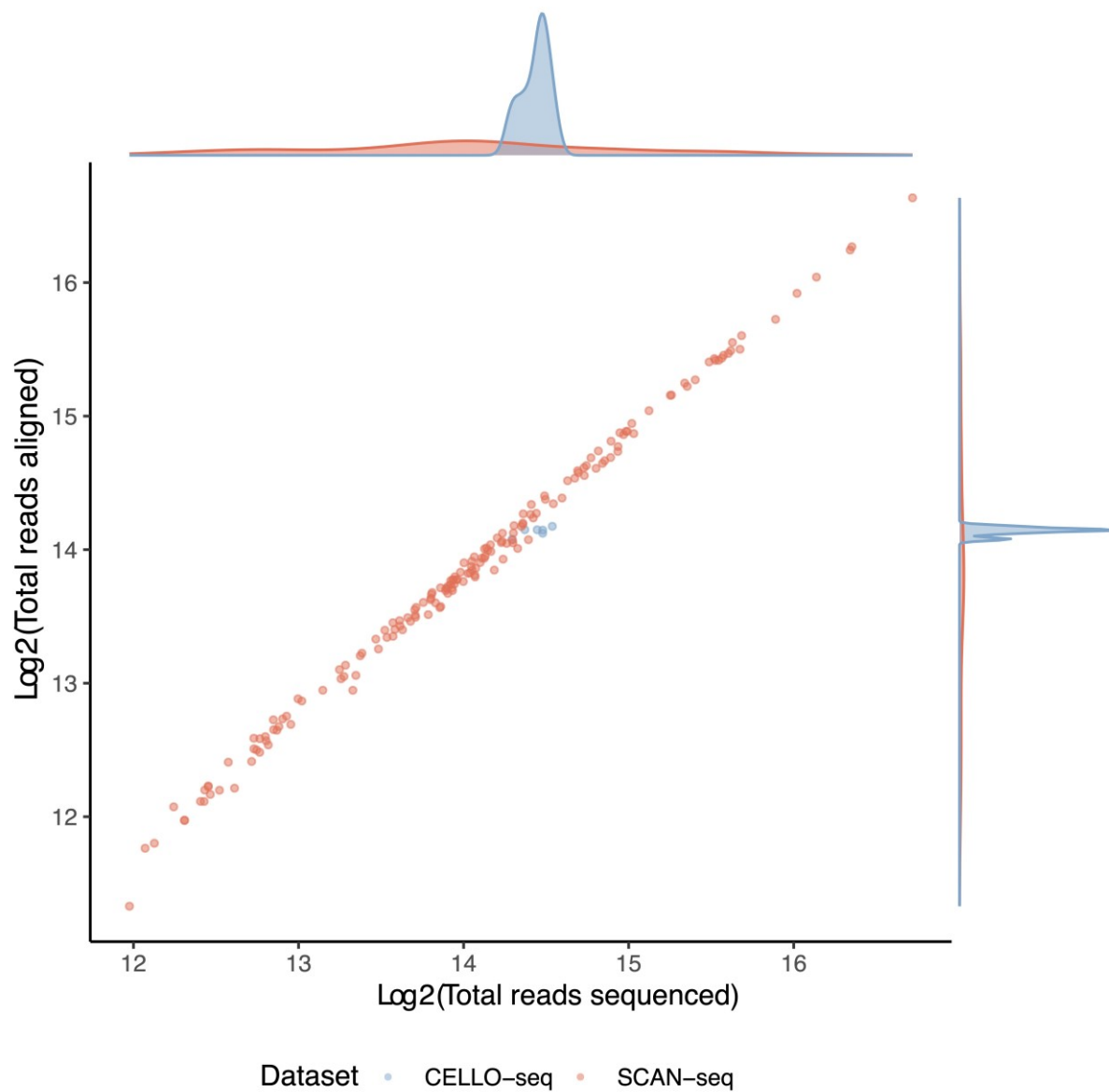


Figure 2: Quality control for long-read RNA-seq data. Both datasets are single-cell modifications to cDNA-PCR Oxford Nanopore sequencing. Log2 transformation shown for total reads sequenced versus reads aligned. Data analysed from (Berrens et al., 2021; Fan et al., 2020).

Appendix 2: Troubleshooting parent genes

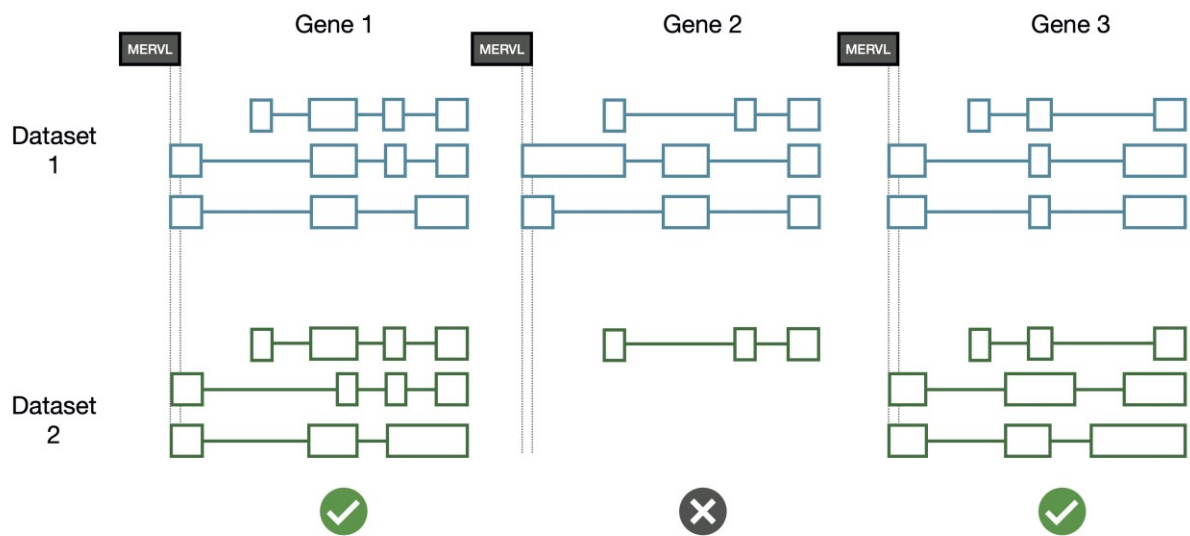


Figure 1: schematic defining how parent genes are selected. If a gene has a MERVL-promoted transcript in >1 dataset, it is considered a parent gene, and if it has a MERVL-promoted transcript in ≤ 1 dataset, it is discarded.

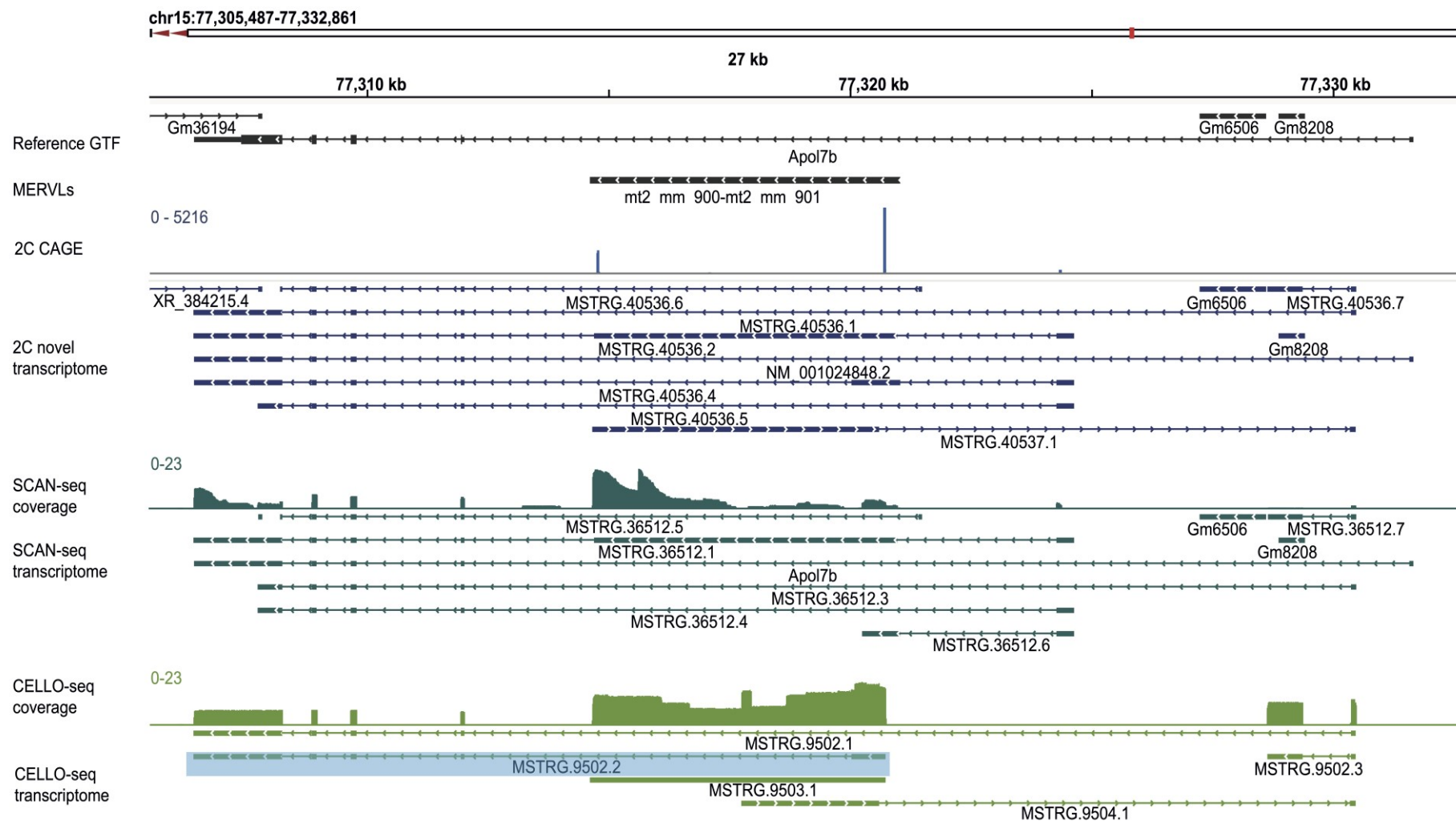


Figure 2: IGV view of example gene *Apol7b*, where one dataset has a MERVL-promoted transcript (Berrens et al., (2021), CELLO-seq) and one does not (Fan et al., (2020), SCAN-seq), highlighted with a light blue rectangle.

Tracks in order: Chromosome location and window size (27Kb), mm39 reference transcript models, conserved MERVL elements, 2C embryo CAGE-seq coverage (single best alignment), 2C novel transcriptome transcript models, coverage from Fan et al., (2020) SCAN-seq dataset, coverage from Berrens et al., (2021) CELLO-seq dataset. All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

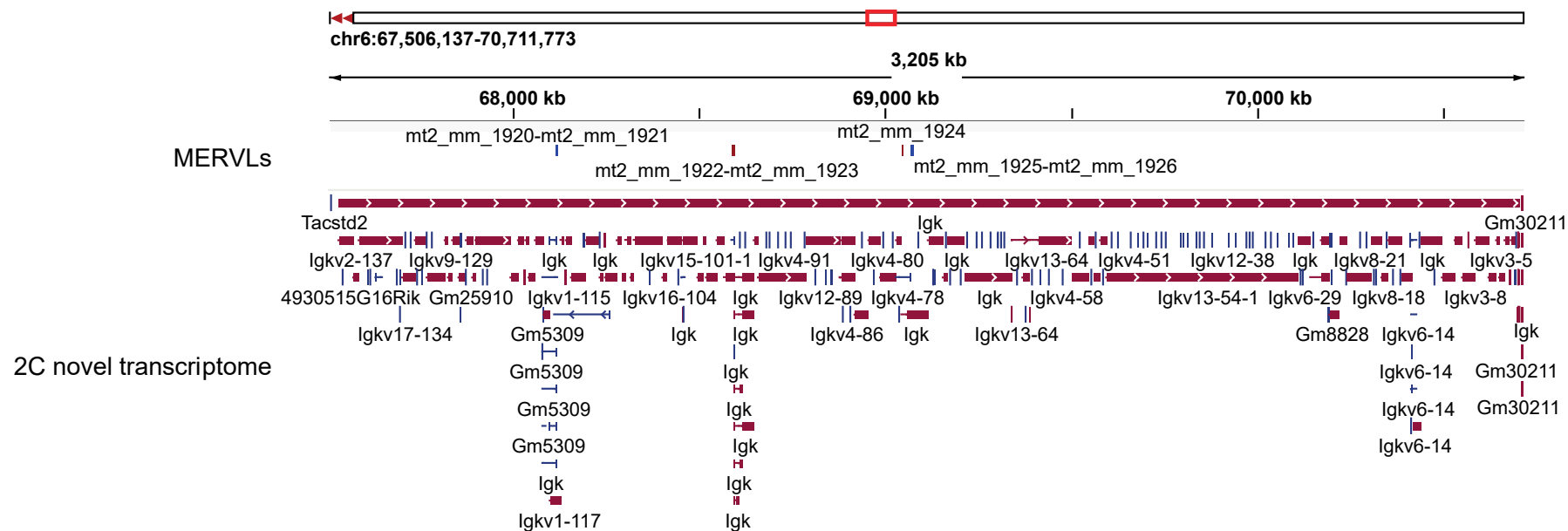


Figure 3: IGV view of *Igk* gene locus, removed from analysis due to complexity. Tracks in order: Chromosome location and window size (3205 Kb), conserved MERVL elements, and 2C novel transcriptome transcript models. Data analysed from Berrens et al., (2021) and Fan et al., (2020).

Appendix 3: Transcript isoform structures

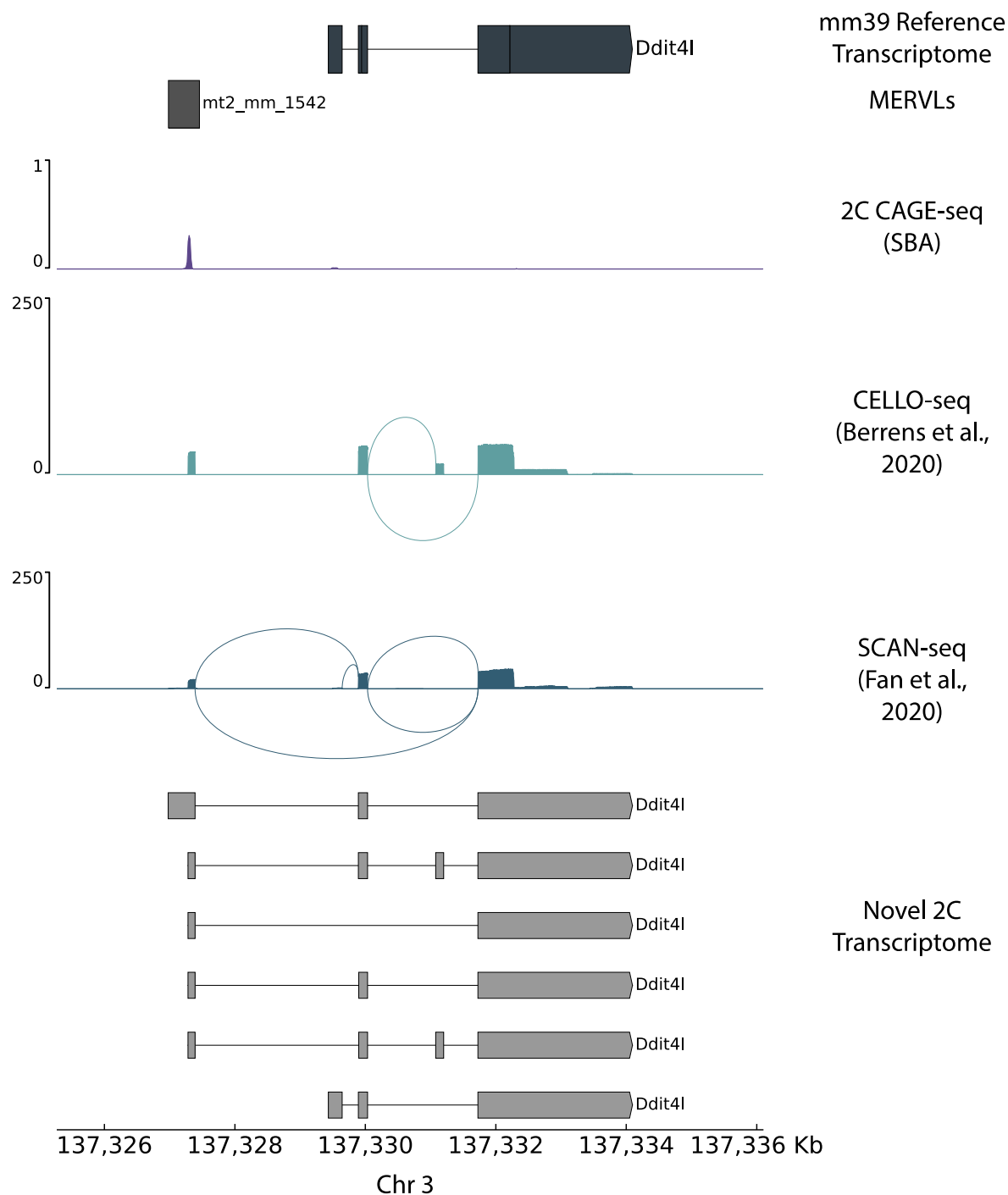


Figure 1: Sashimi plot of *Ddit4l* locus (Chr3:137325270-137336094).

Tracks in order: mm39 reference annotated transcript models, conserved MERV elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020) SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data. All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

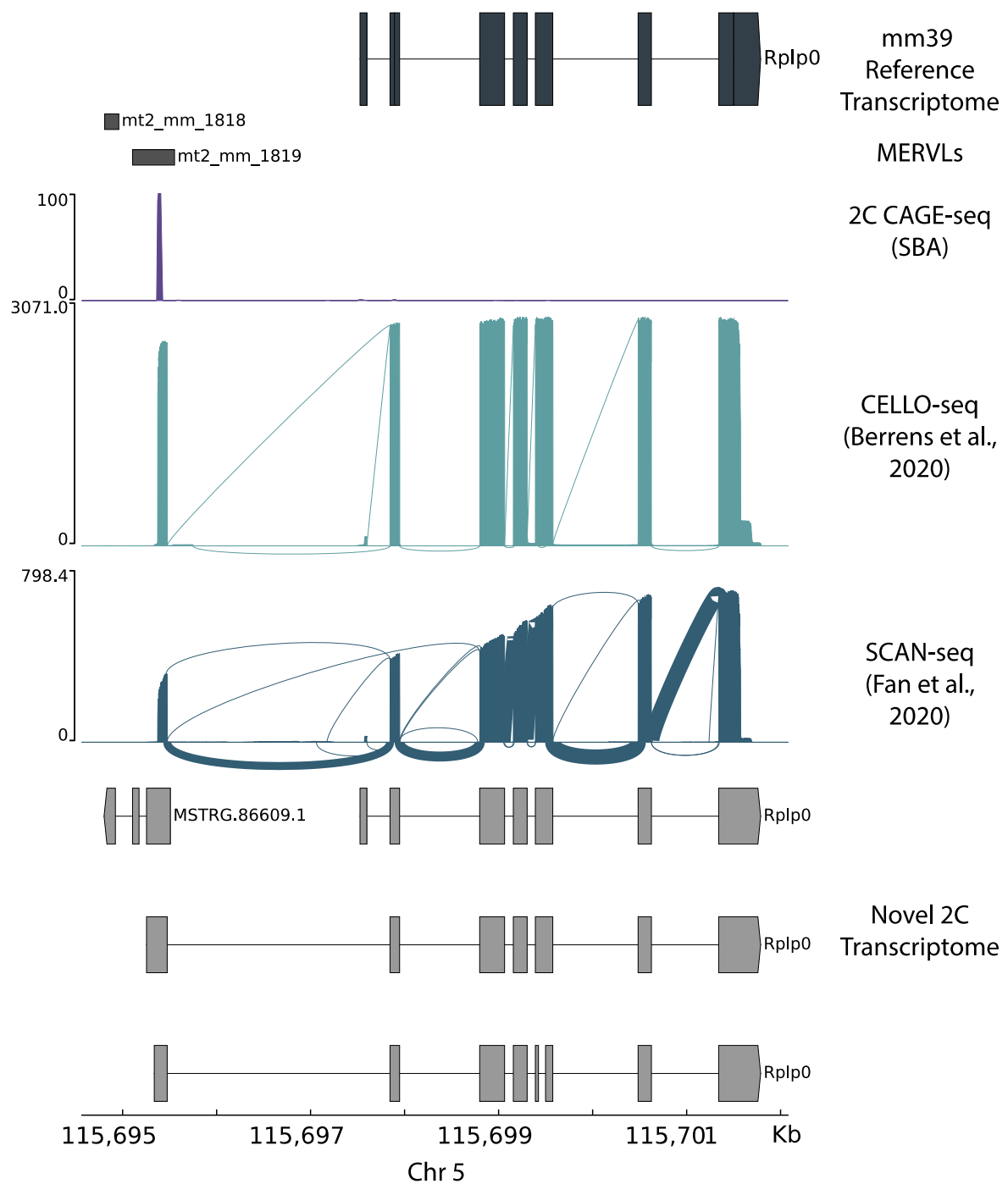


Figure 2: Sashimi plot of *Rplp0* locus (Chr5:115694567-115702078).

Tracks in order: mm39 reference annotated transcript models merged, conserved MERVLS elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi

plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020) SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data. All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

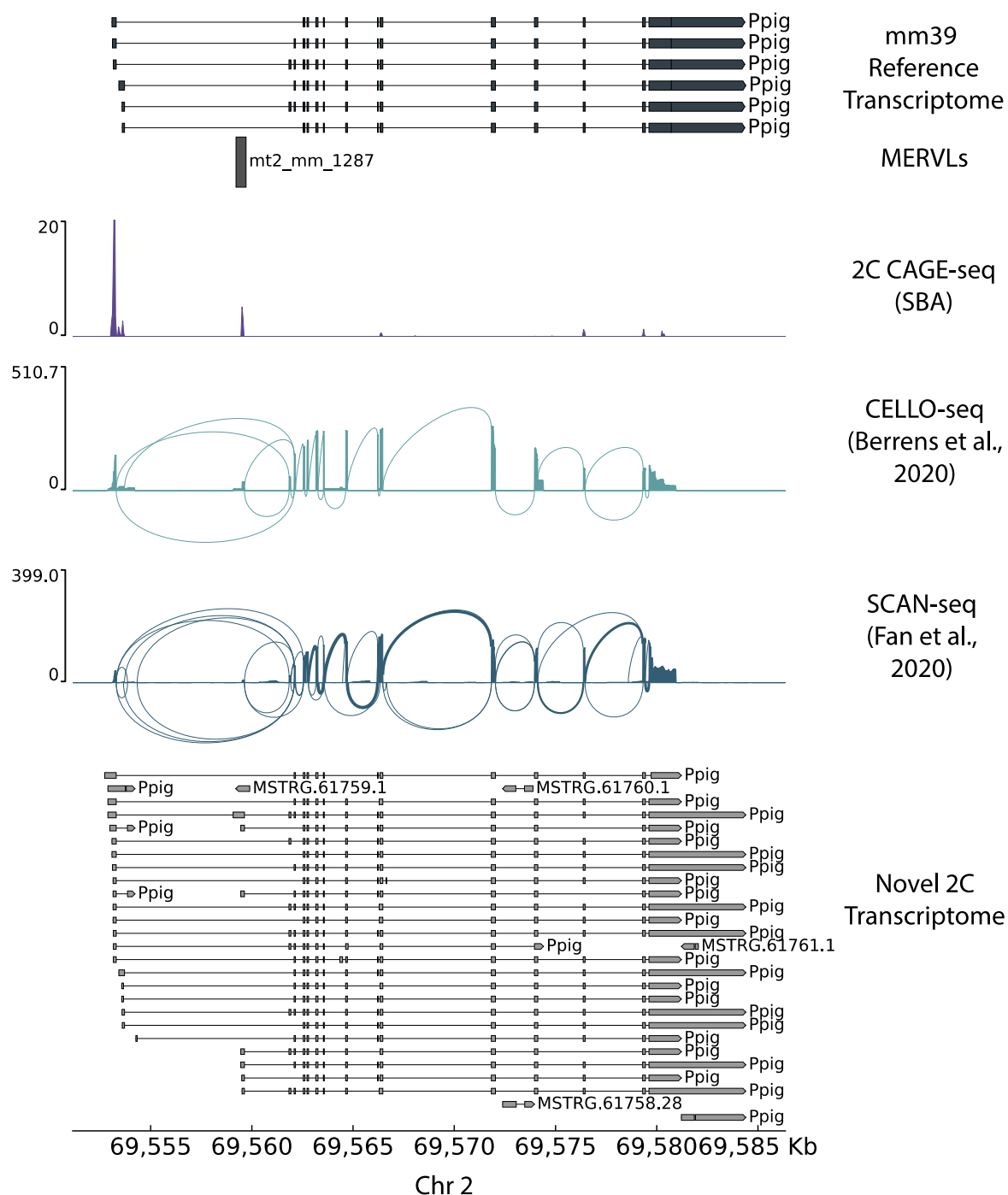


Figure 3: Sashimi plot of *Ppig* locus (Chr2:69551152-69586356). Tracks in order: mm39 reference annotated transcript models merged, conserved MERVLS elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq dataset,

sashimi plot of coverage and splice junctions from Fan et al., (2020) SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data. All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

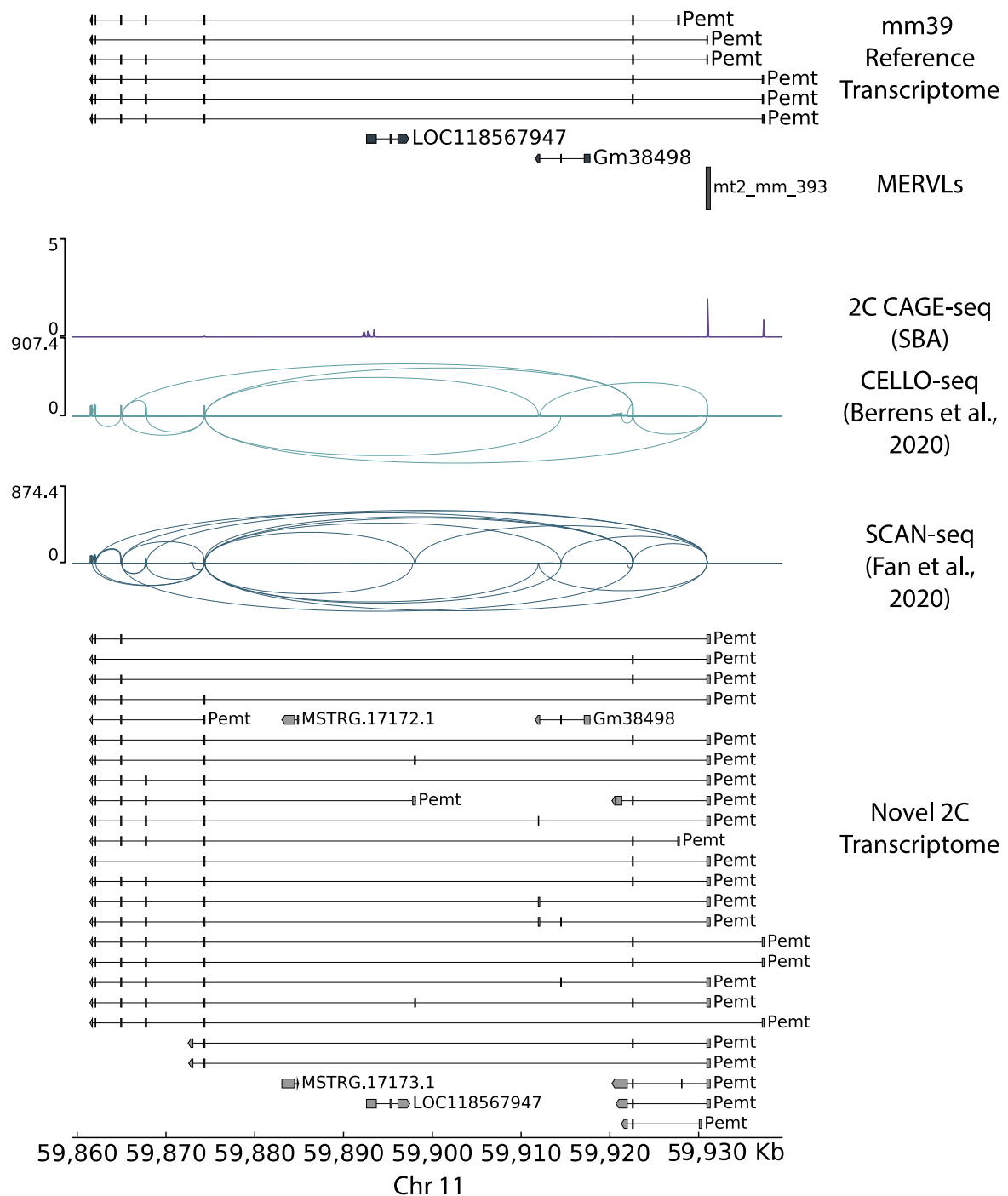


Figure 4: Sashimi plot of *Pemt* locus (Chr11:59859446-59939394):

Tracks in order: mm39 reference annotated transcript models merged, conserved MERV elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq

dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020)
SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data.
All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

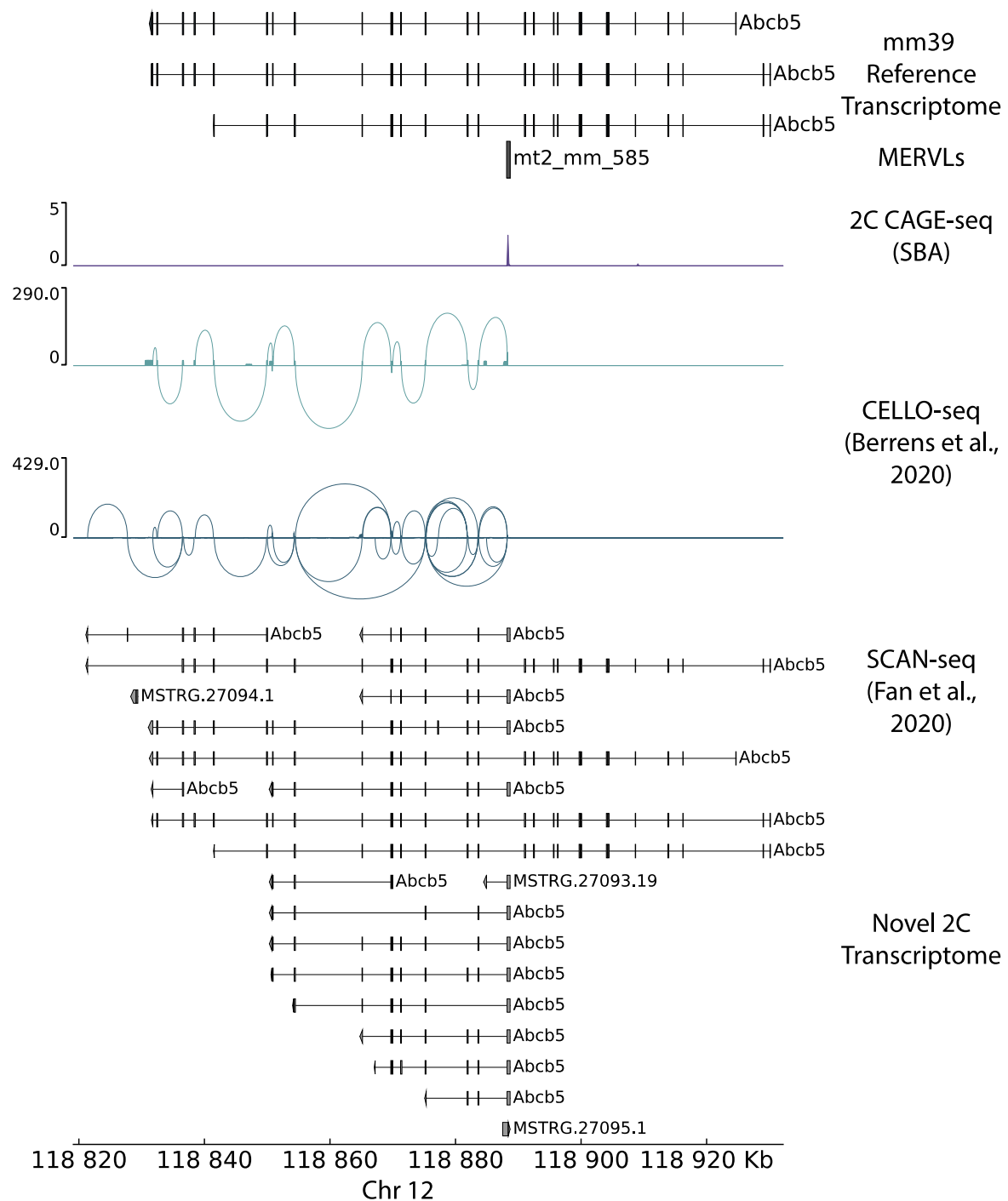


Figure 5: Sashimi plot of *Abcb5* locus (Chr12:118819122-118932169).

Tracks in order: mm39 reference annotated transcript models merged, conserved MERV elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi

plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020) SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data. All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

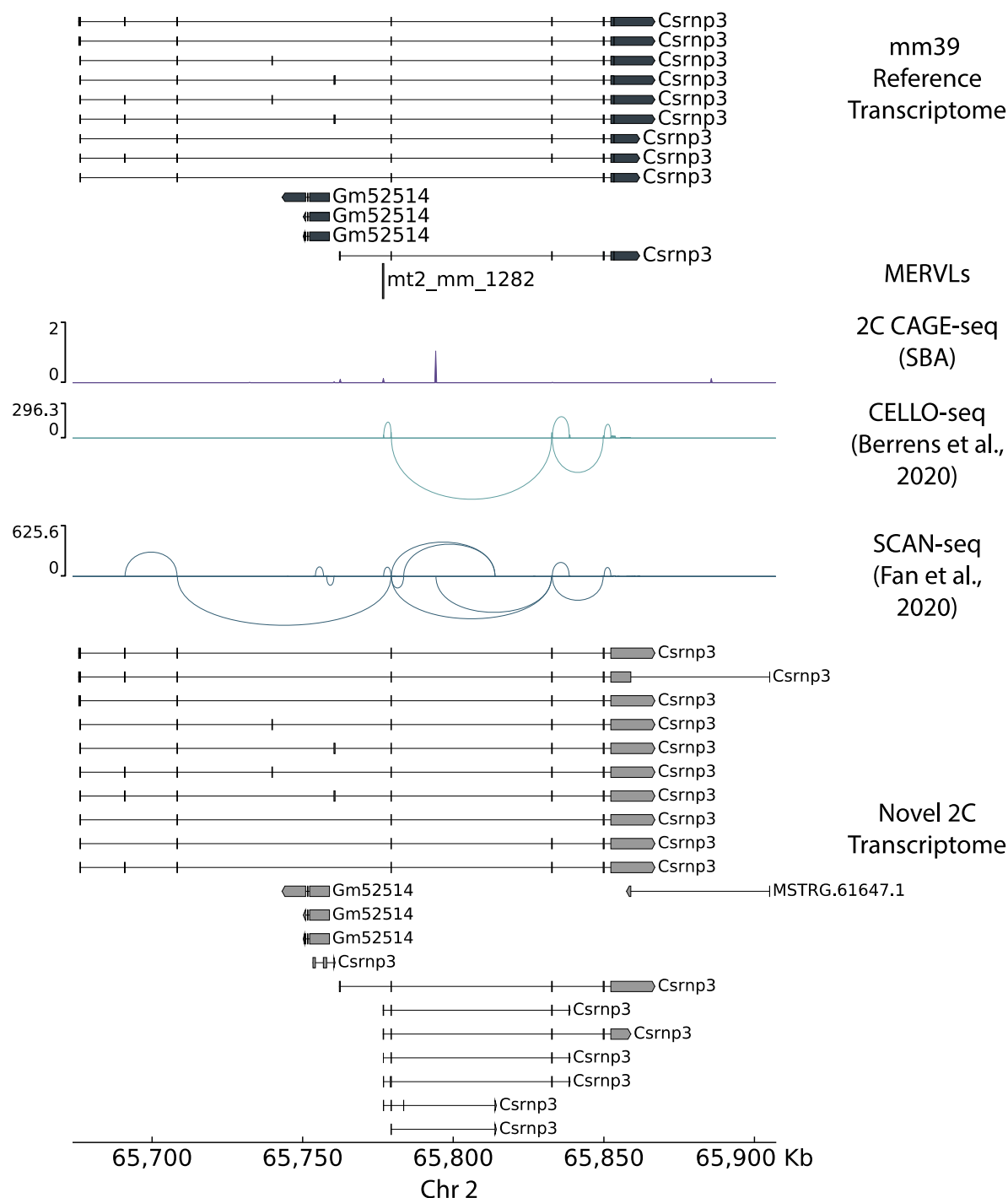


Figure 6: Sashimi plot of *Csrnp3* locus (Chr2:65673628-65907061).

Tracks in order: mm39 reference annotated transcript models merged, conserved MERVL elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq

dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020)
SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data.
All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

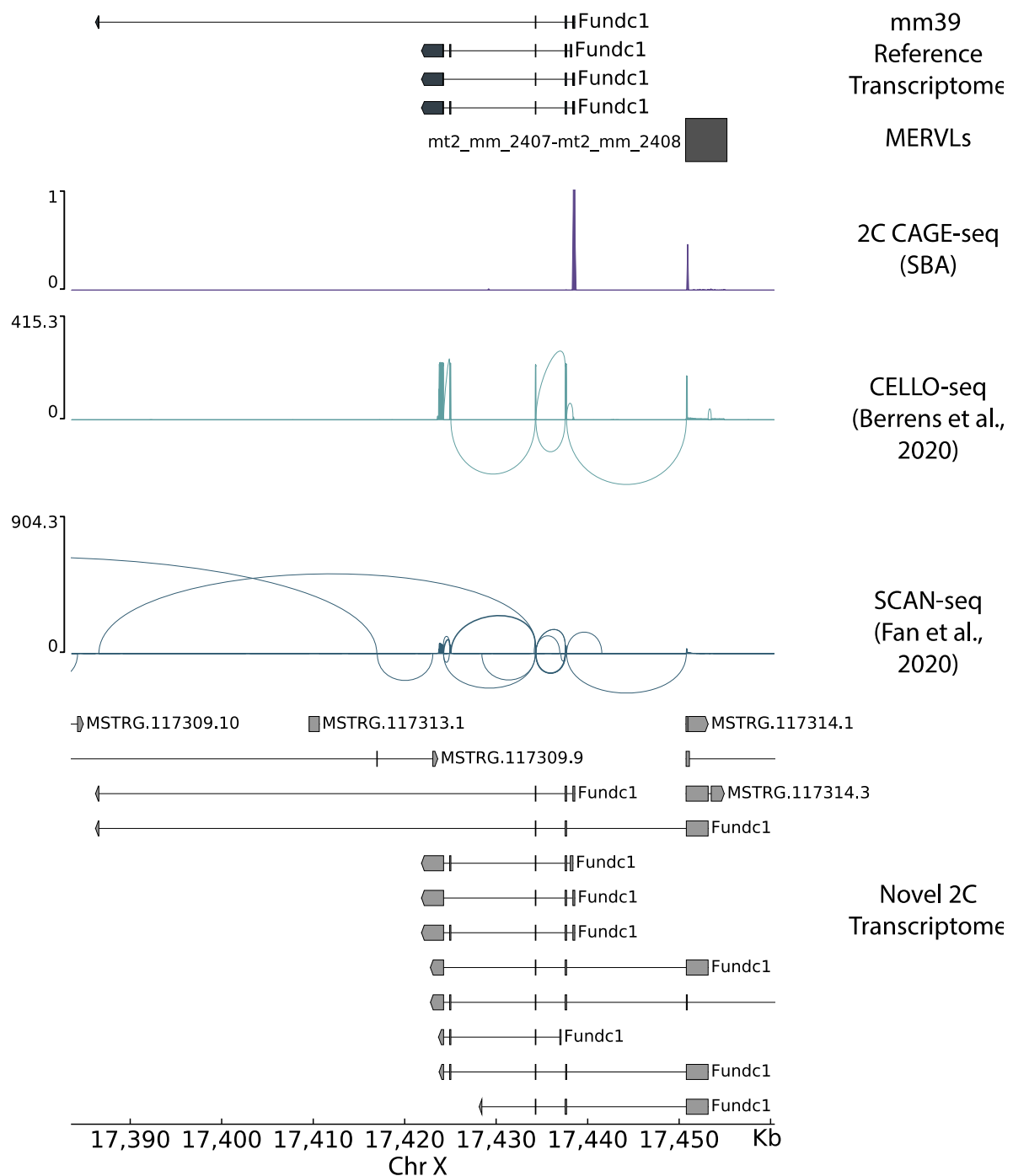


Figure 7: Sashimi plot of *Fundc1* locus (ChrX:17383598-17460387).

Tracks in order: mm39 reference annotated transcript models merged, conserved MERVL elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq

dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020)
SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data.
All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020)
SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data.
All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

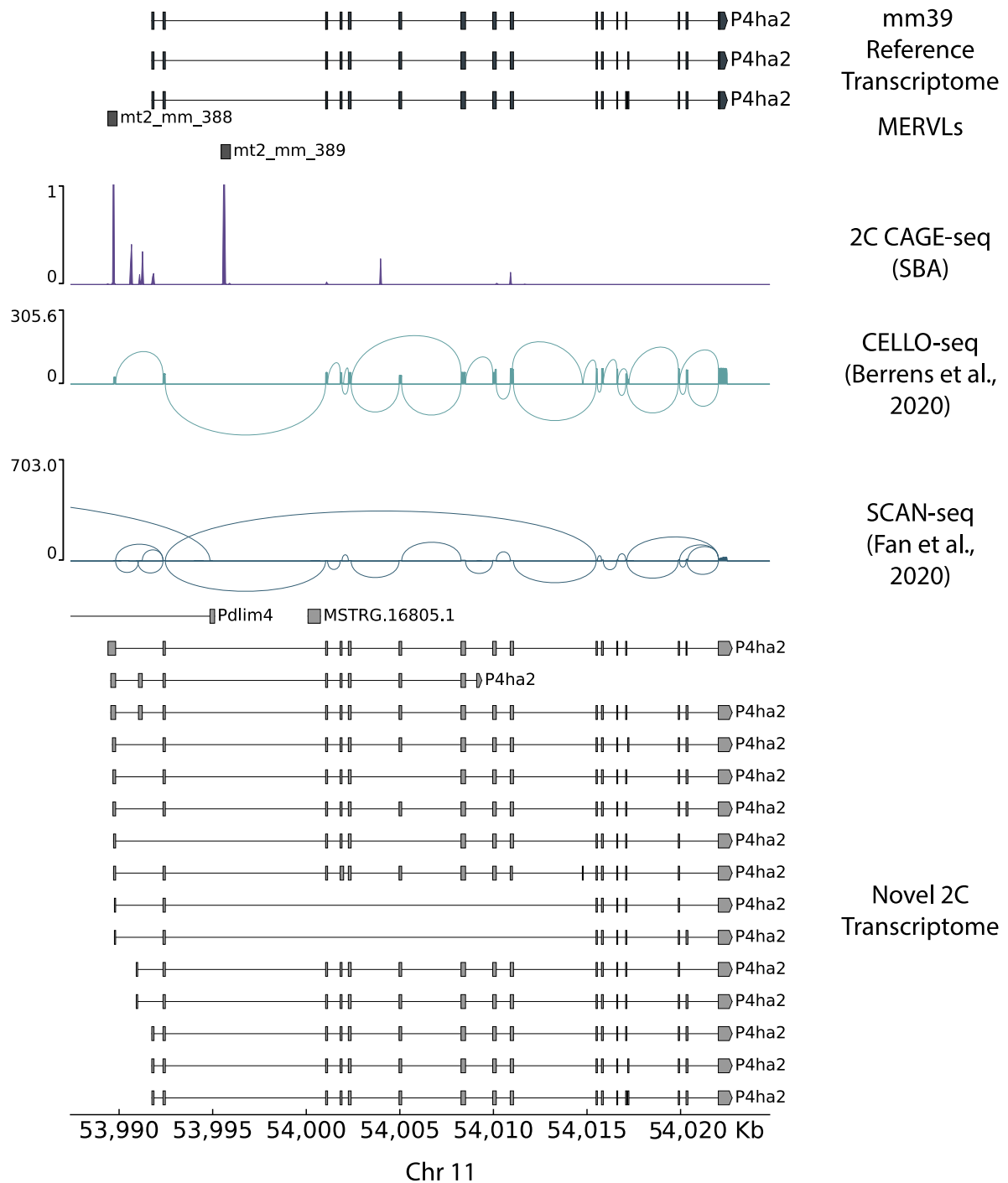


Figure 9: Sashimi plot of *P4ha2* locus (Chr11:53987410-54024752).

Tracks in order: mm39 reference annotated transcript models merged, conserved MERVL elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq

dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020)
SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data.
All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020) SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data. All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

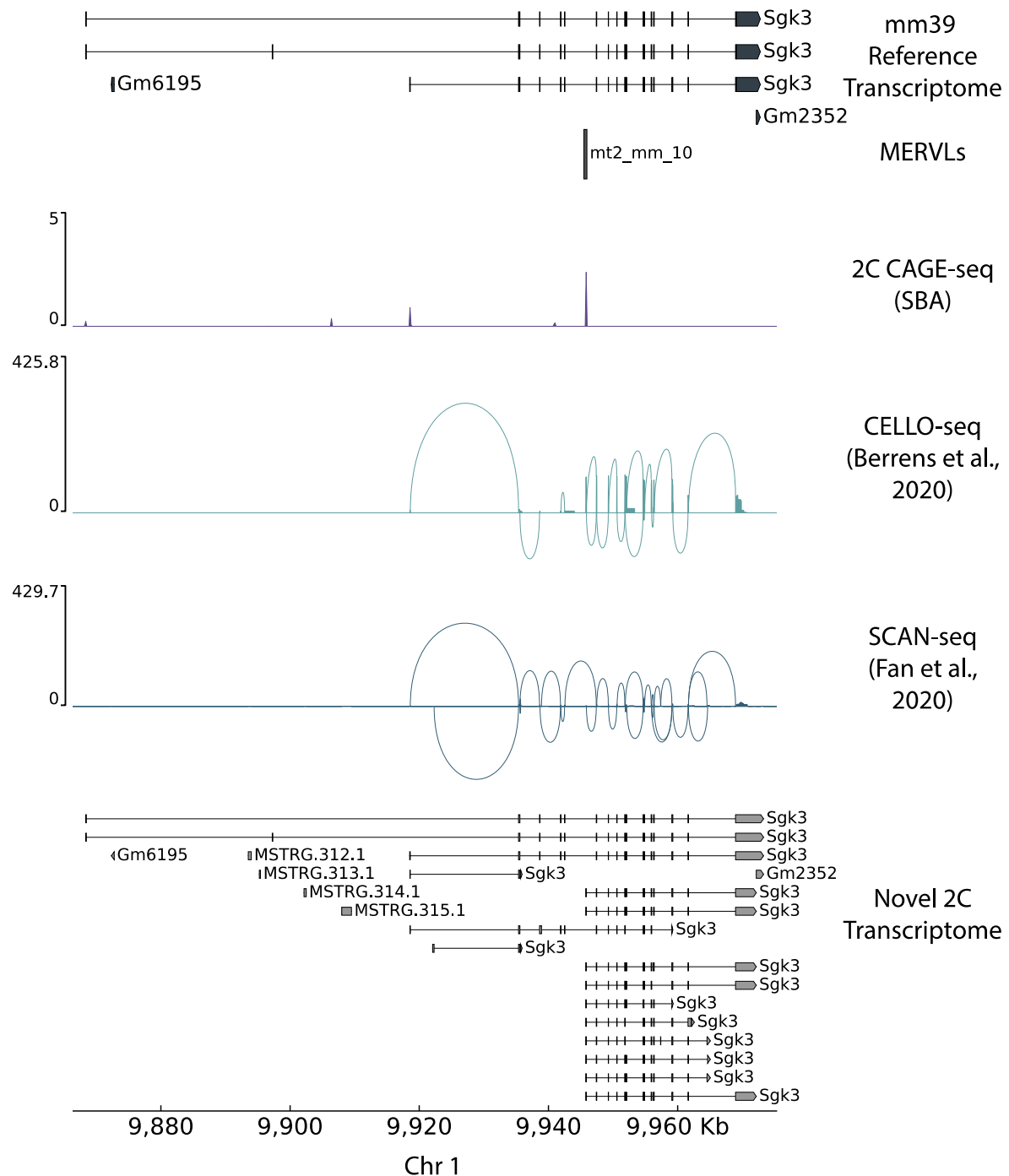


Figure 11: Sashimi plot of *Sgk3* locus (Chr1:9866355-9975296). Tracks in order: mm39 reference annotated transcript models merged, conserved MERVLS elements, 2C embryo CAGE-seq coverage (single best alignment), sashimi plot of coverage and splice junctions from Berrens et al., (2021) CELLO-seq dataset,

sashimi plot of coverage and splice junctions from Fan et al., (2020) SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data. All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

dataset, sashimi plot of coverage and splice junctions from Fan et al., (2020)
SCAN-seq dataset, and novel transcriptome models of 2C embryo long-read data.
All coverage is shown in TPM. Data analysed from Berrens et al., 2021; Cvetesic et al., 2020; Fan et al., 2020.

Appendix 4: ‘best’ ORF per gene (MERVL-promoted versus reference)

	Reference ORFs						MERVL-promoted ORFs					
Gene Name	Transcript ID and ORF	mR NA leng th	OR F len gth	Fick ett scor e	Codin g proba bility	meanTP M at 2C stage	Transcript ID and ORF	mR NA leng th	OR F len gth	Fick ett scor e	Codin g proba bility	meanTP M at 2C stage
Sgk3	NM_001037759.1_ORF_1	5899	1491	0.8838	0.99938649	4.952160944444445	MSTRG.4331130.7_ORF_1	433	1137	0.8914	0.99238382	4.451278666666667
Dst	XM_006495686.5_ORF_1	12317	7923	1.2279	1	19.691121444444445	MSTRG.1183.36_ORF_1	14417	7923	1.2279	1	20.82737861111111
9130227L01Rik	NR_045837.1_ORF_4	1917	144	1.305	0.25967932	0.18975516666666664	MSTRG.2190.3_ORF_3	522	78	0.9144	0.03786976	12.508691
Satb2	NM_139146.2_ORF_1	6747	2202	1.2667	0.99999943	2.3918452222222222	MSTRG.2211.8_ORF_1	2101	846	1.1422	0.98001911	0
Gm28403	XR_878428.3_ORF_1	1182	114	0.9814	0.0491771	0	MSTRG.4331.12_ORF_1	5631	2166	0.9849	0.99999439	14.455802444444444
Lmx1a	XM_006496611.3_ORF_1	4198	1149	1.0631	0.99863211	6.244253333333333	MSTRG.6905.7_ORF_1	3566	1641	1.0395	0.99980193	5.571217611111111
Ddr2	XM_006496696.5_ORF_1	8625	2565	1.1117	0.99999993	8.348922222222223	MSTRG.6963.14_ORF_1	2962	216	0.8532	0.12008414	27.13474777777778
Gm31904	NR_166113.1_ORF_2	907	114	0.7221	0.02293965	0	MSTRG.8648.2_ORF_3	1701	120	0.9203	0.0585577	0
Gm10145	XM_036156152.1_ORF_1	699	465	1.122	0.79508501	0.4188860555555556	MSTRG.9374.4_ORF_1	788	294	1.0069	0.3867112	1.353401111111111

Mcm 9	XM_01731 4098.3_O RF_1	7141 	234 0	0.79 43	0.999 99883	6.471348 5555555 55	MSTRG.1 0498.6_O RF_1	7263 	313 5	1.03 25	1 	1.742621 3333333 334
Micu1	XM_0361 55734.1_O RF_1	834 8	744 	1.30 09	0.983 95456	6.851560 611111111	MSTRG.1 0721.11_O RF_3	2536 	243 	1.27 18	0.662 28987	1.050141 444444 4443
LOC1 15487 393	XR_0039 48878.1_O RF_1	266 2	321 	0.54 54	0.1076 7277	7.177696 3888888 88	MSTRG.1 0862.1_O RF_1	350 3	321 	0.54 54	0.1057 9934	7.73624 9277777 778
Bicc1	NM_0013 47189.1_O RF_1	5564 	285 6	1.09 73	0.999 99999	6.97545 	MSTRG.11 307.4_OR F_1	1883 	852 	1.24 13	0.988 62983	4.15683 4777777 777
R3hd m2	XM_0065 14153.2_O RF_1	4131 	313 5	0.8 494	1 	2.779419 611111111	MSTRG.1 4068.1_O RF_1	438 6	313 5	0.8 494	1 	1.87559 444444 44443
Olfr7 87	NM_0010 11822.1_O RF_1	939 	939 	0.99 01	0.9712 8231	0.06497 439	MSTRG.1 4285.1_O RF_1	1412 	939 	0.99 01	0.970 97244	1.02308 094444 44444
Olfr8 15	NM_1466 70.1_ORF _1	958 	951 	0.8 964	0.933 74852	0 	MSTRG.1 4326.1_O RF_1	3312 	951 	0.8 964	0.930 26381	0.61359 0277777 7777
Thoc5	XM_0065 14448.2_O RF_1	234 2	208 8	1.29 34	0.999 99901	10.47223 811111111	MSTRG.1 4552.14_ ORF_1	262 9	205 2	1.31 32	0.999 99878	12.6322 9422222 2223
1700 093K 21Rik	XM_0065 14789.3_O RF_1	228 7	591 	1.01 04	0.832 73163	2.292532 8888888 887	MSTRG.15 494.1_OR F_1	1586 	591 	1.01 04	0.835 00059	3.47877 3833333 3333
57305 22Eo 2Rik	NR_02797 3.1_ORF_ 1	1520 	315 	0.54 05	0.070 59292	0.325615 9999999 9996	MSTRG.15 576.9_OR F_1	952 	207 	0.6 042	0.070 43241	2.51697 266666 66667
P4ha 2	XM_006 532452.2 _ORF_1	259 8	167 4	1.10 49	0.999 9678 6	2.32419 694444 44446	MSTRG. 16804.6 _ORF_1	242 8	147 0	1.11 21	0.999 8664 4	2.6502 62888 88888 9
Pemt	NM_0012 90013.1_O RF_1	1195 	600 	1.18 86	0.922 55388	23.09777 2722222 22	MSTRG.17 171.18_OR F_1	1291 	378 	1.15 88	0.696 60454	6.50199 3722222 222

Gm4 0019	XR_87992 4.1_ORF_ 5	1516	105	0.66 27	0.063 52366	0	MSTRG.17 420.1_OR F_2	240 8	180	0.58 95	0.061 23684	0.08717 767
Clec1 oa	NM_0012 04252.1_O RF_1	1873	918	1.14 48	0.992 94028	1.798226 5	MSTRG.17 819.1_OR F_1	807 3	313 5	1.03 25	1	6.08675 866666 6666
Olfr3 76	NM_1469 22.2_ORF _1	1193	954	1.23 81	0.992 45451	1.542345 8333333 333	MSTRG.1 8089.1_O RF_1	1314	963	1.23 81	0.992 98535	1.119848 666666 6667
Gm3 8407	NR_11048 7.1_ORF_ 2	3113	246	1.00 01	0.2195 3351	3.693919 9444444 446	MSTRG.2 2107.2_O RF_1	1773	246	1.00 01	0.224 94497	2.54905 688888 8889
Gm31 624	XR_87255 2.3_ORF_ 1	236 8	273	0.61 83	0.063 95466	1.595375 7777777 77	MSTRG.2 2712.9_O RF_1	1635	270	0.4 818	0.056 75985	1.80780 9611111 11
Gm59 54	XM_01731 5315.1_OR F_1	2109	501	1.05 78	0.862 4824	0.08081 372	MSTRG.2 2886.3_O RF_2	1402	216	1.31 14	0.661 65987	1.53041 044444 44444
Tpo	XR_00395 0045.1_O RF_1	3427	239 7	0.85 72	0.999 99959	0.237627 9444444 4446	MSTRG.2 3132.8_O RF_1	208 0	261	0.94 8	0.1167 6814	7.65976 7277777 777
Rbm2 5	XM_0302 46878.2_ ORF_1	539 0	251 7	1.16 47	0.999 99989	30.9092 4766666 6666	MSTRG.2 5384.6_O RF_1	3557	189 3	1.09 71	0.999 99097	14.72071 3722222 223
Abcb 5	XM_006 516370.3 _ORF_1	413 5	373 2	0.9 97	1	0.2909 450555 5555555	MSTRG. 27093.16 _ORF_1	167 8	90 9	1.01 06	0.974 8988 8	17.1484 927777 7778
Pou6f 2	XM_0361 57980.1_O RF_1	626 0	207 6	1.10 02	0.999 99703	2.926661 5555555 555	MSTRG.2 7825.13_ ORF_1	1616	405	1.08 36	0.560 78889	4.58936 35
Gm4 0910	XR_87323 0.2_ORF_ 1	852	171	0.8 491	0.1321 7244	1.591365 3333333 333	MSTRG.2 8730.4_O RF_1	805 7	313 5	1.03 25	1	4.86164 7444444 444
Ecm2	NM_0010 12324.3_O RF_1	359 6	201 3	0.8 682	0.999 98493	1.348366 5555555 555	MSTRG.2 9373.1_O RF_1	1361	483	0.67 19	0.306 71576	0.29809 9111111 1115

Gm4 0954	XR_87342 7.1_ORF_ 1	481	276	0.59 88	0.109 02297	0.482184 8333333 3334	MSTRG.2 9625.1_O RF_1	1435	237	0.65 12	0.102 39058	4.361651 2777777 78
Gm19 792	XR_00395 0606.1_O RF_2	233 9	120	0.96 92	0.048 16844	2.544087	MSTRG.3 0152.5_O RF_2	1594	282	1.06 36	0.224 57677	0.45832 26111111 111
Fam1 72a	XM_0361 58144.1_O RF_1	389 2	960	1.00 42	0.987 07646	2.203437 0555555 56	MSTRG.3 1025.2_O RF_1	7156	313 5	1.03 25	1	5.758119 8333333 34
Spz1	NM_0302 37.3_ORF _1	1602	113 7	1.16 17	0.997 58903	31.71949 3666666 665	MSTRG.3 1425.1_O RF_1	1579	717	1.25 98	0.965 89811	18.5160 41611111 11
Tent2	NM_0013 61536.1_O RF_1	5277	144 3	0.73 76	0.998 45568	2.841609 8333333 335	MSTRG.3 1452.11_O RF_1	2237	144 3	0.73 24	0.998 52518	0.89297 266666 66666
Scam p1	XM_0302 47117.1_O RF_1	3719	861	0.92 84	0.965 05657	13.23994 9222222 222	MSTRG.3 1520.9_O RF_1	9267	296 1	1.00 45	0.999 99998	39.4480 56611111 11
Gm2 0767	XM_0065 17734.2_O RF_1	910	510	0.79 46	0.637 30865	5.715938 3333333 34	MSTRG.3 2653.4_O RF_1	5385	202 5	0.87 29	0.999 9851	17.71298 694444 4445
Gm3 8418	XM_0361 58945.1_O RF_1	920	543	1.04 27	0.786 52832	0	MSTRG.3 3205.5_O RF_1	895	141	1.22 92	0.233 2498	1.44944 1555555 5555
Gnp nat1	XM_006 519274.4 _ORF_1	210 0	411	0.9 782	0.399 3694 8	5.55898 8	MSTRG. 34562.8 _ORF_1	199 6	555	0.7 356	0.521 4063 8	6.6319 5411111 1111
Pck2	XM_0302 48084.2_ ORF_1	3217	107 4	0.92 47	0.996 31296	1.999268 77777777 78	MSTRG.3 5179.1_O RF_1	2181	154 8	0.8 447	0.999 85683	1.54818 46111111 11
Phfi1 a	NM_1726 03.3_ORF _1	1355	882	0.74 15	0.924 0137	0.76049 32777777 778	MSTRG.3 5425.4_O RF_1	1154	447	0.47 03	0.2117 3469	0.67413
Tpt1	NM_0094 29.3_ORF _1	911	519	1.28 97	0.925 84398	37.32783 861111111	MSTRG.3 6249.6_O RF_1	1326	519	1.28 97	0.9251 755	44.5006 08111111 11

Gm4 6481	XR_00178 1421.2_OR F_1	472	345	0.65 57	0.2551 6286	0	MSTRG.3 6945.5_O RF_4	1299	135	0.62 19	0.062 80393	3.20694 8333333 3335
Farp1	XM_0065 18915.4_O RF_1	528 5	311 7	1.30 09	1	4.018450 7222222 22	MSTRG.3 7625.6_O RF_1	577	210	0.57 17	0.0371 9476	0.26382 7722222 2222
Gm41 258	XR_87500 8.3_ORF_ 3	339 9	147	1.08 6	0.105 07324	0.08298 367	MSTRG.3 7910.2_O RF_1	300 5	195 0	1.01 82	0.999 97799	9.58027 3166666 666
Pd2d 2	NM_0010 81064.2_ ORF_1	1155 6	839 1	0.94 87	1	4.99558 8722222 222	MSTRG.3 8238.28_ ORF_1	1474	480	1.23 74	0.922 61729	0.011581 39
Zfp64 7	XM_0361 59399.1_O RF_1	282 4	160 8	0.96 12	0.999 94139	0.49043 9777777 7775	MSTRG.4 0499.15_ ORF_1	7825	313 5	1.03 25	1	3.66395 0555555 5556
Gm10 865	NR_04574 6.1_ORF_ 1	713	300	0.75 53	0.186 26747	0	MSTRG.4 0629.2_O RF_1	1314	192	0.9 063	0.1263 1632	0.32439 894444 444445
Tfcp2	XM_0361 59312.1_O RF_1	362 0	915	1.08 57	0.989 47094	4.345095 2777777 7	MSTRG.4 2024.18_ ORF_1	439 9	218 4	0.9 849	0.999 99548	6.38408 6222222 222
Gm35 974	XR_38461 9.1_ORF_ 1	234 3	288	0.8 067	0.067 32901	0.362217 2222222 222	MSTRG.4 2340.2_O RF_5	1976	177	0.8 917	0.1158 0878	5.21469
Gm36 195	XR_38462 2.5_ORF_ 1	760 8	444	0.6 009	0.1369 5259	2.84634 25	MSTRG.4 2360.4_O RF_1	6714	313 5	1.03 25	1	9.75623 883333 3333
Ercc4	NM_0157 69.2_ORF _1	670 3	275 4	1.12 37	0.999 99998	5.80634 5055555 555	MSTRG.4 2891.2_O RF_1	1092 9	313 8	1.03 25	1	11.94491 4111111 1
Mpv1 71	XM_0361 60190.1_O RF_2	4770	273	0.97 76	0.3175 3909	1.733662 2222222 222	MSTRG.4 2948.3_O RF_1	1011	366	0.97 7	0.196 64062	69.5907 2511111 11
Gtf2e 1	NM_0013 55727.1_O RF_1	3514	132 3	1.18 68	0.999 60888	1.626277 5555555 555	MSTRG.4 4358.5_O RF_1	200 2	132 3	1.18 68	0.999 62245	0.53140 138888 88889

Robo 1	XM_0065 22961.4_O RF_1	848 8	490 2	0.8 924	1 924	1.851559 9444444 444	MSTRG.4 5454.17_ ORF_1	6955	294 0	0.96 81	0.999 99998	8.90236 2333333 333
D17E rtd64 8e	NR_0458 o8.1_ORF _2	891	168	0.95 85	0.1547 1254	2.585725 3333333 334	MSTRG.4 6896.2_O RF_2	715	120	0.8 678	0.099 62868	0.55338 36111111 111
Vmn2 r94	NM_0011 04543.2_ ORF_1	2515	245 7	0.77 5	0.999 99887	0	MSTRG.4 7285.7_O RF_1	1134	348	0.8 056	0.094 60577	1.89319 638888 8889
Vmn1 r227	NM_13419 5.3_ORF_ 1	957	921	0.75 93	0.8716 6099	0.258819 9444444 4445	MSTRG.4 7394.7_O RF_1	1017	345	0.82 2	0.086 73333	0.12288 1333333 33334
Snrpc	NM_01143 2.2_ORF_ 1	763	480	1.15 93	0.7767 3521	12.68334 011111111	MSTRG.4 8117.2_O RF_1	1311	480	1.15 93	0.7745 0694	167.432 3737777 7777
Olfr11 8	NM_21372 1.2_ORF_ 1	1054	966	0.82 07	0.965 31941	0.188123 27777777 776	MSTRG.4 9122.2_O RF_1	748	441	0.8 643	0.486 71256	0.60867 438888 88889
Esp2 4	NM_0012 56050.1_O RF_1	552	207	0.56 95	0.036 53444	4.78486 6944444 4445	MSTRG.4 9224.2_O RF_2	1601	153	0.93 25	0.046 82723	3.961851 7222222 223
Supt3	NM_1786 52.3_ORF _1	1498	112 5	0.9 096	0.996 15296	1.235285 611111111 2	MSTRG.4 9409.6_O RF_1	1170	207	1.05 5	0.106 52697	1.241055 0555555 555
Camk mt	XM_0065 24993.5_ ORF_1	1094 7	600	0.92 3	0.805 19207	2.464518 1666666 667	MSTRG.51 230.9_OR F_1	1678 4	313 5	1.03 25	0.999 99999	24.5199 6105555 5555
LOC1 18568 292	XR_0049 39620.1_O RF_4	554	78	1.05 01	0.067 00921	0.06497 439	MSTRG.51 326.1_OR F_1	605	336	0.55 68	0.1349 9712	0.86703 1666666 6666
Gm2 0939	NM_0013 79445.1_O RF_1	470 0	216 3	1.14 2	0.999 99407	4.872892 7222222 22	MSTRG.51 583.3_OR F_1	495 0	206 7	1.14 2	0.999 98709	4.86357 1833333 333
Fgf1	XM_0065 25642.5_ ORF_2	3918	531	0.92 51	0.820 82186	2.153357 27777777 8	MSTRG.5 3201.19_ ORF_2	684	180	0.51 97	0.055 91832	1.32031 650000 00001

Csnk1	XM_0065	4719	136	0.71	0.997	7.755660	MSTRG.5	366	909	0.66	0.940	7.83640
g3	26231.2_O		8	85	69714	8888888	3711.11_O	6		56	41963	4222222
	RF_1					89	RF_1					222
Dcc	NM_0078	1048	434	0.91	1	8.84495	MSTRG.5	7296	313	1.03	1	4.04406
	31.3_ORF	4	4	87		5166666	4486.5_O		5	25		7944444
	_1					667	RF_1					4445
Plaat	NM_139	352	48	1.2	0.89	20.3503	MSTRG.	370	48	1.2	0.89	11.5484
3	269.2_O	3	9	424	6489	2727777	55724.4_	2	9	424	6100	862777
	RF_1				03	778	ORF_1				29	77778
Slc22	XM_01124	202	155	0.8	0.999	0.725012	MSTRG.5	1272	486	0.8	0.322	6.63257
a28	7277.2_O	8	1	063	31674	5000000	5758.4_O			869	60896	266666
	RF_1					001	RF_1					6667
Ankr	NM_0242	1754	576	1.13	0.829	3.723634	MSTRG.5	7190	313	1.03	1	7.78268
d22	04.6_ORF			27	06201	1666666	7142.2_O		5	25		6222222
	_1					67	RF_1					222
Xpnp	XM_01731	230	165	1.28	0.999	4.138099	MSTRG.5	270	187	1.25	0.999	5.68998
ep1	8075.3_O	2	9	68	97662	0555555	8250.7_O	6	2	98	99494	2833333
	RF_1					55	RF_1					3335
Echdc	NM_0242	1813	903	1.06	0.9871	2.20232	MSTRG.5	3585	155	1.00	0.999	5.05369
3	08.4_ORF			34	66	4166666	8934.5_O		7	47	58596	388888
	_1					6666	RF_1					8889
LOC1	LOC11856	7810	266	0.96	0.999	4.597788	MSTRG.5	713	180	0.6	0.074	0.12994
18568	8501_ORF		7	68	9999	9444444	9536.2_O			817	57424	8722222
501	_1					45	RF_1					22223
Gm35	XR_37416	1139	141	1.04	0.096	0	MSTRG.5	1853	141	1.04	0.095	0.28539
689	4.3_ORF_			99	59807		9878.1_O			99	15201	0111111
	3						RF_4					111
Med2	NM_0268	1272	936	1.27	0.995	12.70848	MSTRG.6	1758	936	1.27	0.995	14.56491
7	96.5_ORF			46	44885	5333333	0127.1_O			46	39711	3277777
	_1					334	RF_1					778
Gm52	XR_0049	639	210	0.8	0.1451	0	MSTRG.6	695	313	1.03	1	10.4273
552	40892.1_			932	9076		0616.1_O	4	5	25		3516666
	ORF_1						RF_1					6667
Mbd5	XM_0302	8641	398	0.91	1	0.840951	MSTRG.6	842	449	0.77	1	0.53350
	46638.2_		4	65		6666666	1072.23_	4	7	66		1666666
	ORF_1					667	ORF_1					6667

Csrn	XM_036	161	175	1.14	0.999	2.68532	MSTRG.	956	43	0.7	0.694	2.5380
p3	162720.1 _ORF_1	46	8	58	9747 9	31111111 113	61645.16 _ORF_1		2	988	4241 3	037777 77778
Ppig	XM_006	625	22	0.7	0.999	42.4165	MSTRG.	620	22	0.7	0.999	32.709
	499185.5 _ORF_1	5	59	697	9964 6	7105555 556	61758.22 _ORF_1	5	59	697	9964 7	863055 55556
Zswi	XM_0361	1185	750	0.70	0.878	0.913719	MSTRG.6	638	313	1.03	1	3.20465
m2	62578.1_O RF_1			54	00747	3888888 888	2373.12_ ORF_1	4	5	25		4833333 333
Aqr	NM_0097	4973	444	1.25	1	6.669215	MSTRG.6	852	444	1.25	1	9.77773
	02.3_ORF _1		6	87		8333333 33	3824.1_O RF_1	3	6	87		438888 8888
Fam9	XM_0065	1858	120	0.77	0.994	8.09398	MSTRG.6	1021	540	0.8	0.6571	6.27437
8b	00096.5_ ORF_1		9	35	28755	9777777 77	3913.9_O RF_1			848	0829	3722222 222
Zc3h	XM_01123	439	276	1.04	0.999	4.629619	MSTRG.6	1333	489	0.94	0.645	9.16854
6	9851.1_OR F_1	5	3	39	99995	7777777 8	4581.2_O RF_1			72	87515	5944444 444
AU01	NR_0334	243	294	0.94	0.1781	7.109934	MSTRG.6	869	180	0.52	0.058	5.74655
9990	69.1_ORF _1	4		85	0825		4849.4_O RF_2			69	79743	988888 8889
Csnk	NM_0077	4277	117	0.99	0.997	39.77172	MSTRG.6	2947	504	0.81	0.695	6.39525
2a1	88.3_ORF _1		6	39	06442	1222222 226	5544.19_ ORF_1			17	0657	7333333 333
Rpn2	XM_0361	360	188	1.28	0.999	7.609645	MSTRG.6	309	124	1.28	0.999	4.59106
	59778.1_O RF_1	3	7	05	99499	7777777 8	5961.1_O RF_1	0	2	05	46993	7944444 444
Gm3	XR_86726	2341	249	1.06	0.1841	0.30956	MSTRG.6	919	81	0.93	0.025	1.12696
8537	4.4_ORF_ 1			54	825	9277777 7776	9266.1_O RF_4			81	13237	8277777 7779
Arhge	NM_0010	520	261	1.02	0.999	11.97540	MSTRG.6	250	387	1.05	0.344	3.45870
f26	81295.1_O RF_1	8	0	74	99994	8388888 889	9685.6_O RF_1	2		07	56438	2611111 11
Gm33	XR_00395	1653	228	0.77	0.049	0.94434	MSTRG.6	8691	313	1.03	1	4.16635
923	4490.1_O RF_1			89	97306	2944444 4444	9686.3_O RF_1		5	25		9111111 11

Sprr2 e	NM_01147 1.2_ORF_ 1	648	231	1.31 77	0.276 64471	0	MSTRG.71 128.1_OR F_1	1023	231	1.31 77	0.274 89481	0.49104 6222222 2222
6530 403H o2Rik	NR_13017 0.1_ORF_ 4	3914	201	0.96 69	0.1173 1014	0	MSTRG.7 2639.4_O RF_1	935	180	0.58 26	0.059 70461	1.868717 944444 4444
Tacr3	NM_0213 82.6_ORF _1	3764	135 9	1.20 25	0.999 64007	3.183705 77777777 8	MSTRG.7 3207.2_O RF_1	1173	501	1.04 17	0.808 51638	2.58280 694444 44446
Ddit4 l	NM_0301 43.4_ORF _1	2723	582	0.8 681	0.802 2584	5.297784 3888888 89	MSTRG.7 3298.1_O RF_1	292 4	582	0.8 681	0.801 51232	15.6560 02611111 111
Gm32 823	XR_00395 5364.1_O RF_2	4914	240	0.76 61	0.144 45613	1.754351 0555555 555	MSTRG.7 4166.9_O RF_2	100 4	150	0.92 35	0.105 45752	4.60505 988888 88885
Mob3 b	XM_0361 63953.1_O RF_1	6077	651	1.32 04	0.972 06296	0.89844 3055555 5555	MSTRG.7 5253.7_O RF_1	297 8	651	1.32 04	0.973 9635	0.67775 5722222 2222
4930 578G 10Rik	NR_15545 6.1_ORF_ 2	634	150	1.02 68	0.1172 0644	0	MSTRG.7 5645.2_O RF_1	716	207	0.69 35	0.084 2481	4.47098 894444 4445
Gm11 232	XR_0049 42280.1_ ORF_1	1579	906	0.65 35	0.947 84853	0.06497 439	MSTRG.7 6951.1_O RF_1	266 5	906	0.65 35	0.946 5799	3.48620 41111111 113
Gm32 658	XR_39067 0.3_ORF_ 4	1097	159	0.87 14	0.079 09388	0.669167 9444444 445	MSTRG.7 7184.4_O RF_1	883	180	0.55 66	0.062 74543	0.99013 8222222 2223
Gm12 790	NM_0013 77350.1_O RF_1	5487	260 1	0.9 849	0.999 99977	11.25565 61111111 2	MSTRG.7 8039.2_O RF_1	603 7	313 5	0.99 61	1	18.14977 1
Gm52 673	XR_00395 5092.1_O RF_3	1104	237	0.81 85	0.1180 4812	0	MSTRG.7 8288.2_O RF_1	464 6	260 1	0.9 849	0.999 99978	0.8661
Trabd 2b	XR_0049 41946.1_O RF_1	7052	115 8	1.32 7	0.999 28233	1.039359 6666666 667	MSTRG.7 8594.7_O RF_1	582 4	639	1.10 09	0.922 45389	0.61146 56111111 111

Arap2	NM_1784 07.3_ORF _1	7717	511 2	0.79 02	1	76.56392 7944444 44	MSTRG.8 4028.15_ ORF_1	1769	921	0.54 66	0.9311 5682	3.923717 61111111 1
0610 040J o1Rik	XM_01124 0792.4_O RF_1	202 8	942	1.21 9	0.991 99498	2.28939 2	MSTRG.8 4043.6_O RF_1	394 4	155 7	1.05 23	0.999 61089	7.45406 0777777 777
Limc h1	XM_0361 65610.1_O RF_1	6119	266 1	0.87 78	0.999 99996	2.137611 9444444 446	MSTRG.8 4226.17_ ORF_1	789 2	313 5	1.03 25	1 788888 8889	2.63925 788888 8889
Rplp o	NM_0074 75.5_ORF _1	1360	954	1.31 42	0.997 85786	251.1529 2344444 444	MSTRG.8 6610.1_O RF_1	1507	954	1.31 42	0.997 85051	400.891 23111111 11
Ccdc6 o	NM_0013 60004.1_ ORF_1	259 9	163 5	1.31 74	0.999 98204	1.228123 8888888 888	MSTRG.8 6651.3_O RF_1	7316	313 5	1.03 25	1 31111111 113	3.98453 31111111 113
Gm4 2808	XR_00178 5007.3_O RF_3	270 3	228	0.76 58	0.1131 6834	2.184651 611111111	MSTRG.8 8184.1_O RF_1	4816	218 4	1.01 9	0.999 99551	11.3684 3944444 4444
Spdye 4b	NM_1778 80.4_ORF _1	2169	711	1.16 78	0.966 99156	1.737028 1666666 667	MSTRG.8 8345.5_O RF_3	2019	180	0.8 053	0.059 60685	5.01260 994444 4445
Gm3 8781	XR_00395 6419.2_O RF_1	1890	393	0.76 37	0.268 79093	0.457117 8333333 3333	MSTRG.9 0094.1_O RF_1	1719	393	0.76 37	0.269 57691	0.82973 6222222 2222
Sva	NM_0092 99.2_ORF _1	654	498	0.67 59	0.373 6826	0.950011 2222222 222	MSTRG.9 0563.5_O RF_1	242 0	129 3	1.09 85	0.997 69217	3.22199 5222222 222
Ccser 1	XM_0065 05955.3_ ORF_1	1934 6	222 0	0.72 89	0.999 99435	4.049316 8333333 33	MSTRG.9 1521.12_O RF_1	1584 0	218 4	1.00 47	0.999 99433	8.46651 2
Nat8f 2	NM_0530 96.3_ORF _1	3297	717	0.82 97	0.866 32531	5.691726 5555555 55	MSTRG.9 2526.1_O RF_1	1113 3	313 5	1.03 25	1 55	9.09781 55
Antxr 1	NM_0540 41.2_ORF _1	526 4	173 1	0.82 06	0.999 95609	25.31479 561111111 2	MSTRG.9 2633.2_O RF_1	1018 5	313 5	1.03 25	1 404444 44445	27.2039 404444 44445

Gm55	NR_0269	1474	180	0.82	0.088	6.419168	MSTRG.9	382	260	0.9	0.999	5.93480
77	90.1_ORF_2			25	54402	3888888	2690.3_ORF_1	9	1	849	99978	76111111
						89						11
Speer	NR_00158	775	132	0.78	0.1291	0	MSTRG.9	867	180	0.49	0.048	0.68630
9-ps1	3.3_ORF_3			15	7281		5548.1_ORF_1			8	95165	4055555
												5555
Dnajb	XM_0065	1175	936	1.22	0.995	0.503171	MSTRG.1	1136	120	1.29	0.493	0.64980
13	08184.3_ORF_1			61	3771	5555555	01410.5_ORF_2			42	61128	05
						556						
Usp17	NM_0010	1715	163	1.00	0.999	1.564283	MSTRG.1	5749	298	0.96	0.999	13.17837
ld	01559.2_ORF_1		8	37	90895	9444444	01622.4_ORF_1		2	81	99998	788888
						445						8889
Ric3	NM_1787	1927	687	0.8	0.9112	3.117276	MSTRG.1	7391	174	1.05	0.999	11.04110
	80.3_ORF_1			87	8822	4444444	02074.6_ORF_1		6	41	87443	988888
						443						8888
Ctf2	NM_1988	900	615	1.00	0.904	0.182533	MSTRG.1	242	495	0.99	0.795	8.34723
	58.1_ORF_1			36	87899		03046.1_ORF_1	3		67	53311	06111111
												12
Gm33	XR_37852	3676	159	1.24	0.213	5.899041	MSTRG.1	1233	180	0.64	0.076	1.924671
248	1.3_ORF_3			45	07817	2222222	03188.1_ORF_1			9	55697	7777777
						22						778
Gm33	XR_0049	4014	252	0.96	0.240	0.176408	MSTRG.1	1404	120	0.8	0.1421	9.94211
555	34334.1_ORF_3			35	36405	11111111	03207.22_ORF_2			492	9498	9833333
						13						333
Dhcr7	NM_0013	2742	843	1.05	0.987	1.237958	MSTRG.1	3148	141	1.01	0.999	0.76579
	82500.1_ORF_1			25	97886	5555555	03854.9_ORF_1		6	57	72684	1666666
						557						6666
Gm3	XR_00177	779	114	0.8	0.1519	0.74664	MSTRG.1	1449	114	0.8	0.149	5.68136
0220	8491.2_ORF_2			935	5357	6722222	03946.1_ORF_2			935	94726	694444
						2223						4444
Dlgap	XM_0065	249	316	0.8	1	2.211205	MSTRG.1	234	387	1.05	0.345	6.72136
2	08781.5_ORF_1	98	8	885		5	04552.6_ORF_1	3		07	86908	688888
												88884
Gm32	XR_37884	292	105	0.77	0.042	0	MSTRG.1	1043	135	0.66	0.066	2.37334
567	8.3_ORF_3			68	64603		05430.1_ORF_2			5	61711	4111111
												11

Lonp	NM_0258	290	255	1.07	0.999	11.92903	MSTRG.1	1948	387	1.05	0.347	7.06188
2	27.3_ORF_1	9	9	5	99991	95	07928.8_ORF_1			07	49351	7833333
Gm4	XR_87915	100	138	0.94	0.085	1.351892	MSTRG.1	1971	387	1.03	0.329	3.521574
2040	7.2_ORF_3	0		55	30173	8888888	08965.5_ORF_1			25	8136	5555555
2810	NM_0256	1405	846	1.04	0.983	31.55445	MSTRG.11	1016	313	0.99	0.999	112.6104
004N	15.2_ORF_1			51	52398	3555555	0100.1_ORF_1	8	5	61	99999	388888
23Rik						558						889
Endo	XM_01731	4552	104	1.20	0.995	9.634159	MSTRG.11	4124	260	0.9	0.999	12.3223
di	3601.2_ORF_1		4	7	408	8333333	0692.4_ORF_1		1	849	99978	295
Zfp5	XM_030	383	164	0.9	0.999	2.90021	MSTRG.	713	313	1.0	1	34.592
60	244440.2_ORF_1	3	7	174	7205	422222	110993.4_ORF_1	0	5	595		097777
					5	22225						77778
Olfr8	NM_14641	1142	945	1.09	0.976	2.616474	MSTRG.11	1034	291	1.02	0.2107	5.13825
81	8.2_ORF_1			36	42402	8888888	1835.1_ORF_1			61	5441	7555555
						89						556
Bud1	NM_1460	223	191	1.09	0.999	6.556454	MSTRG.11	248	177	1.05	0.999	7.519102
3	00.2_ORF_1	2	4	21	99107	2777777	2400.4_ORF_1	5	0	05	97114	7222222
						8						22
Slc9a	NM_1779	349	193	1.04	0.999	3.012804	MSTRG.11	6729	237	0.99	0.999	13.9450
9	09.5_ORF_1	2	5	41	99041		4756.2_ORF_1		9	91	99905	86
Alas1	NM_0205	270	192	1.13	0.999	8.961837	MSTRG.11	245	192	1.13	0.999	3.311457
	59.2_ORF_1	2	9	24	99474	2777777	5322.3_ORF_1	0	9	24	99477	5555555
						7						557
Ulk4	XM_0065	404	383	1.16	1	0.417162	MSTRG.11	859	231	1.30	0.658	4.34549
	12023.3_ORF_1	2	1	74		0555555	6212.1_ORF_2			54	5915	15
						556						
Gm2	NR_16854	2756	150	0.96	0.102	0.561406	MSTRG.11	820	93	0.76	0.031	1.937351
6797	7.1_ORF_1			98	69985	2222222	6288.2_ORF_1			22	32099	5555555
						223						556
Fund	NM_001	305	33	1.01	0.332	9.58361	MSTRG.	405	152	0.9	0.999	5.5380
c1	313746.1_ORF_1	8	9	66	3635	55	117312.6_ORF_1	9	1	237	3068	945
					6						5	

Rbmx	NM_17337	1761	981	0.8	0.987	22.7228	MSTRG.11	1962	981	0.8	0.987	25.4825
2	6.3_ORF_1			929	76575	0616666	8588.1_ORF_1			929	70887	75111111
						667						11
Sh3k	NM_0011	5168	199	0.92	0.999	1.811329	MSTRG.1	486	182	0.9	0.999	3.09663
bpi	35727.2_ORF_1		8	92	99338	8333333	23210.17_ORF_1	2	1	803	97727	4555555
						334						5552