|  |
| --- |
| City university of Hong Kong |
| Dissect the molecular mechanisms of MERVL mediated shaping of chromatin 3D structure in differentiate mouse embryonic stem cells (mESC) to neural progenitors |
| **Student Name: LIU Zhe 刘哲**  **Student No.: 55283822**  **Programme: BMS2/P**  **Report Type: Qualify Report** |
| **2020/7/14** |

|  |
| --- |
|  |



# A Survey of the Relevant Literature

Chromosome conformation capture technology has revealed important insights into chromatin interaction and the three-dimensional (3D) structure [1]. However, how chromatin remodels from mouse embryonic stem cells (ESC) to neural progenitor cells (NPC) was not fully studies, and how spatial genomic architecture is related to gene expression and cell fate was not clearly understood [2, 3]. Here, we have comprehensively mapped 3D chromatin organization during the differentiation from mESC to NPC *in vitro*, generating the highest-resolution Hi-C maps [4].

Transposable elements (TEs) are rich genetic components of eukaryotic genomes and have important regulatory characteristics that can affect transcription, splicing, and recombination. There are two types of MERVL TEs (**Figure 1**): (1) complete MERVL, consisting of a MERVL-int element and 2 MT2\_Mm long terminal repeats (LTR), and (ii) solo MT2\_Mm, which come from homologous recombination [5]. We focus on type I class of MERVL TEs because we are more interested in complete MERVL sequences. Here, we demonstrate that the mouse endogenous retroviral element (MERVL) family of transposable elements can drive 3D recombination of the genome in the differentiation process from mESC to NPC and effectively repress differentiating into neural cells by preventing lncRNA expression that represses NPC.

MERVL involved in the process of regulating chromatin 3D structure. In despite of their functional relevance, it is impossible to check the chromatin conformation changes associated with TEs because of the lack of dynamic technology [13, 14]. One research lab reported that MERVL plays an important role in embryonic development, and another paper reported that MERVL plays an important role in chromatin formation (via zinc fingers, MERV-L integrated in the first intron of the Zinc fingers) [6, 7]. TEs are often suppressed in mESC development with the regulation of ERVs. Zinc-finger proteins bind to specific sequences in ERVs to silence TEs [8-10]. The epigenetic system is regulating TEs is known by scientist. However, there are lots of distinct types of TE, for which the epigenetic regulation is unclear [11].

# Identification of A Specific Research Topic

1. Chromatin architecture undergoes reorganization during differentiation in human and mouse [15-17]. How chromatin is reorganized during this differentiation process remains to be fully elucidated in mouse?
2. HERV-H can shape cell type-specific chromatin architecture during human pluripotent stem cell (hPSC) differentiation [13, 16, 18, 19]. Both HERV-H and MERVL are long terminal repeat element (LTR) type of repeat element and they are ERV family members. So, whether MERVL involved in the process of differentiation from mESC to NPC? And if MERVL retroviral elements can delineate TAD boundaries?

# The Research Methodology

**ChIP-Seq data processing.**

***Read Alignment.*** In total, there are 45 TF or chromatin modification ChIP-Seq involved in our research. In which, 30 of them have pre-processed wig/bw/bedgraph format data with mm9 as reference genome. For the other 15 ChIP-seq data, we utilized bowtie2 for mapping. PCR duplicates were removed by their flag labels.

***ChIP-Seq peak calling.*** TF or chromatin modification binding sites were called using MACS2 technology. We set the parameters as follows: - -nomodel - -exsize 180 [19]. TF or chromation modification binding sites were predicted by peaks. For ChIP-seq data with time points, such as, CTCF and H3K27ac, we used mergeBed of bedtools to merge them into one peak.

***ChIP-Seq signal visualization.*** The deepTools is a frequently used method for generating aggregated figures. Here, we choose computeMatrix and plotProfile function of deepTools produce aggregated signal profiles.

**RNA-Seq data processing.**

***Read alignment and quantification.*** We utilized Tophat2 software to map sequence reads to mouse reference genome (mm9) [20]. And following assembling by using cufflinks tools [20].

***Differentiation gene expression analysis.*** We identified the differentially expressed genes if their fold changes are larger than 2.0 and adjusted p-values are smaller than 0.05.

**Hi-C data processing.**

At the very first, we should generate the restriction site file. We searched the experiment explanation from the NCBI GEO database. It was shown that the restriction enzyme is MboI in this research. The juicer software was chosen by us for analyzing Hi-C data. The reason is juicer is one of the easiest software for comprehensively and automatically analyzing Hi-C data [21]. The users do not need advanced and sophisticated programming skills during the entire data analysis process, they only need some basic operating specifications to analysis Hi-C data. To sum up the workflow: bwa alignment --> sort --> merge --> remove PCR duplicate --> generate hic format file. For visualization, we obtained contact matrix with 10 kb resolution, and then we normalized using the Knight-Ruiz method. It should be noted that the juicer software version is 1.5.6, and the BWA version is 0.7.17.

***DI***

In general, we used directionality index (DI) for TAD identification. It indicates whether a bin on a chromsome tends to interact with upstream or downstream. The value of each bin is obtained by comparing the number of read A in the upstream 2M range and the number of read B in the downstream 2M range with the average of the two. Here, we used Knight–Ruiz normalized contact matrix as input data of domain call software at 10 kb resolution.

**TADs.**

***TAD identification.*** We called TADs using arrowhead algorithms. Arrowhead was run with the “-m 2000 -r 10000” parameters at 10 kb resolution.

***Identification of non-redundant TAD boundaries.*** The first step of clustering TAD boundaries is to identify non-redundant boundaries across 3 time points. Firstly, we used quantile normalization method for calculating DI scores for each time point. Then, we computed the DI delta scores for each TAD boundary. The DI delta score is defined as the average difference of DI between the 4 bins upstream and the 4 bins downstream of the boundary. We merged them all TAD boundaries and sorted them by their DI delta scores in a descending order. Then, we picked one TAD boundary from the top of the list and removed any remaining boundary within 50 kb of the top TAD boundary. Then, we picked the next TAD boundary on the list and repeated the same process until the entire list was traversed. And we remove the TAD boundaries less than 2 calls. Finally, we only retain TAD boundaries that their delta DI scores are larger than 200.

***Dynamic TAD boundaries across stages.*** To identify ESC+ specific TAD boundaries, we use k-means algorithm for clustering.

***Overlaps of TAD and TAD boundaries with repeat elements and CTCF.*** We used Bedtools intersect function for counting the overlaps between CTCF binding sites and repeat elements with TAD boundaries. In which, we expended the TAD boundaries to upstream 50 kb and download 50 kb because the limitation of calling TAD boundaries. The sequences of repeat elements were downloaded from <http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/>chr\*\_rmsk.txt.gz; modified 22 Aug 2007). And we only kept repeats with more than 1 thousand copies. We used fold enrichment for testing which repeat elements were more enriched in ESC+ TAD boundaries than stable TAD boundaries. And proportion test was used for calculating the significance. The cutoff of fold enrichment is 1.5 and the cutoff of adjusted p-value is 0.05.

**MERVL related analysis.**

***ChIP-Seq enrichment analysis at TAD boundary-associating MERVL loci.*** For selecting which proteins are involved in the process of TAD boundaries formation, we downloaded mESC ChIP-Seq bigwig files and raw fastq files from NCBI GEO. For raw fastq data, we computed the average signals on MERVL sequences. Then, we computed the fold change between the signals at top 50 MERVL sequences and at 51-300 MERVL sequences.

# A Discussion on the Possible Outcome

**The chromatin structure is reorganized during the differentiation process.**

There is a researcher reported that chromatin structure is dynamic during the hESC development. In order to systematically study the chromatin structure during mESC development, we download WT mESCs Hi-C data with the accession number GSE74055 from NCBI GEO [22]. In which, the samples were treated with retinoic acid (RA) for generating NPC. These cells in 0 day, 1.5 day and 2.5 day were collected and sequenced for RNA-seq and Hi-C sequencing. We obtained about 3 billion raw read pairs for each sample (**Table 1**).

Consistent with previous research, we observed that the number of TADs is decreased when we use arrowhead call TAD algorithms (**Figure 2**). Interestingly, the Hi-C contract matrix shown that chromatin structure changed during mESC differentiation to NPC (**Figure 3**).

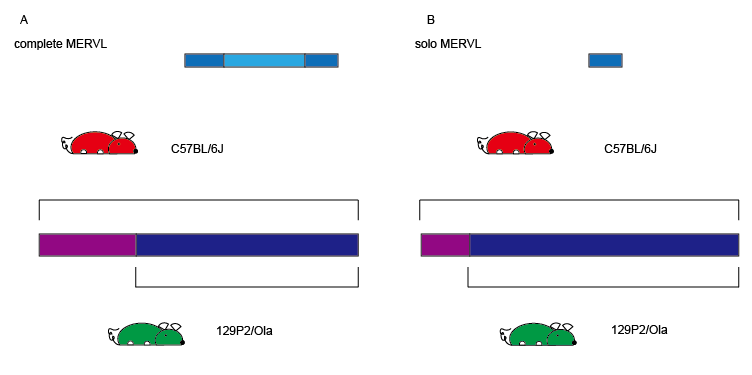
**MERVL and CTCF are involved in the regulation of losing TAD boundaries.**

To explore which repeat elements may involve in the process of chromatin reorganization during mESC differentiation to NPC, we calculate the fold enrichment and q-value for each repeat elements across ESC+ specific TAD boundaries over all stage TAD boundaries. Fortunately, we got 11 significant repeat elements with fold changes larger than 1.5 and q-value less than 0.05, they are (TAAAA)n, ORR1B1-int, RMER4A, RLTR21, MurSatRep1, RMER2, RMER17C-int, IAPEY3-int, RLTR27, U6, and MERVL-int (**Figure 5**). And we also witness a phenomenon that ESC+ specific TAD boundaries are related with RNA expression at the MERVL regions (**Figure 6**).

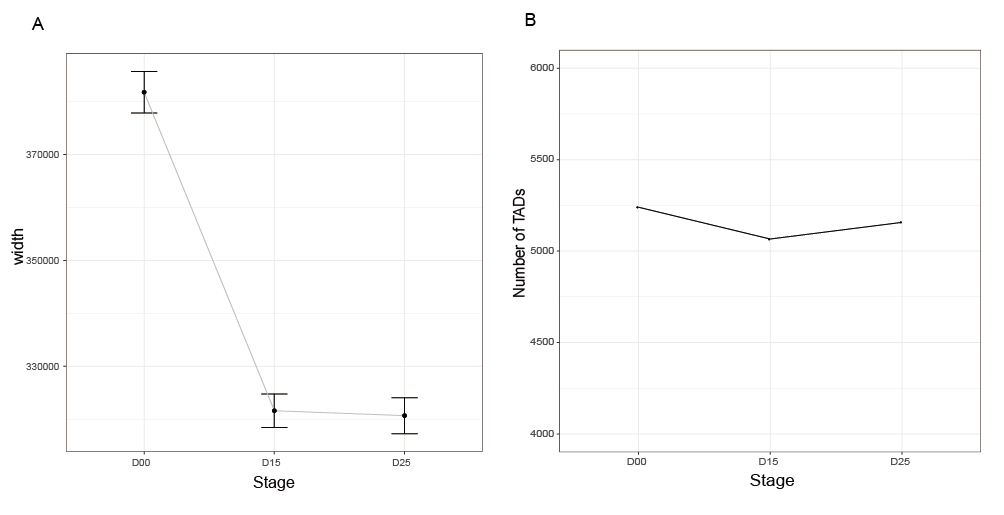
In addition, CTCF plays an important role in splitting TAD boundaries. Based on this result, we calculated the overlap between two types of TAD boundaries and 2 ChIP-seq peaks. For ESC+ specific TAD boundaries, there are 2,727 peaks overlapped with ESC+ boundaries in NPC, and there are 565 peaks overlapped with ESC+ boundaries in ESC. While for stable TAD boundaries, there are 10,695 peaks overlapped with stable TAD boundaries in NPC, and there are 2,271 peaks overlapped with stable TAD boundaries in ESC. Because CTCF can split TAD boundaries, so we explore the relationship between MERVL sequences and CTCF binding sequences. Interestingly, we found CTCF motifs in MERVL sequences (**Figure 7**).

We also download 45 TFs and chromatin modification ChIP-Seq data from NCBI GEO in mESCs. The figure shown that RNA polymerase II subunit A (POLR2A) and cohesion proteins (RAD21 and SMC3) were enriched at sites for the top 50 MERVL sequences (**Figure 8**).

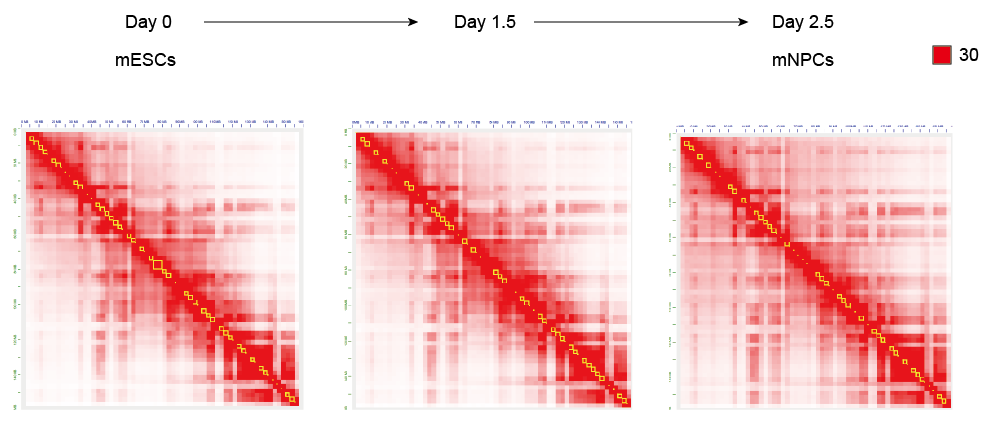
# Figure legends



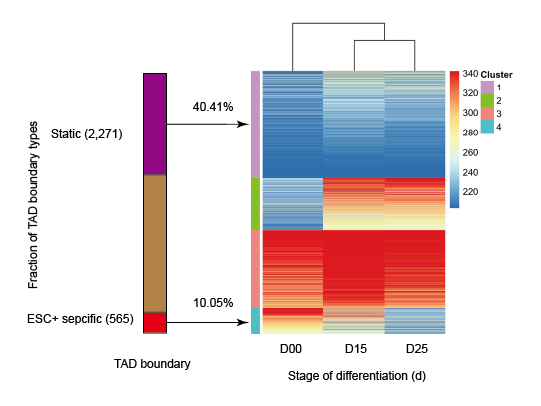
**Figure 1.** Schematic representation of complete MERVL and solo LTR. The (A) is for complete MERVL, and the (B) is for solo LTR.



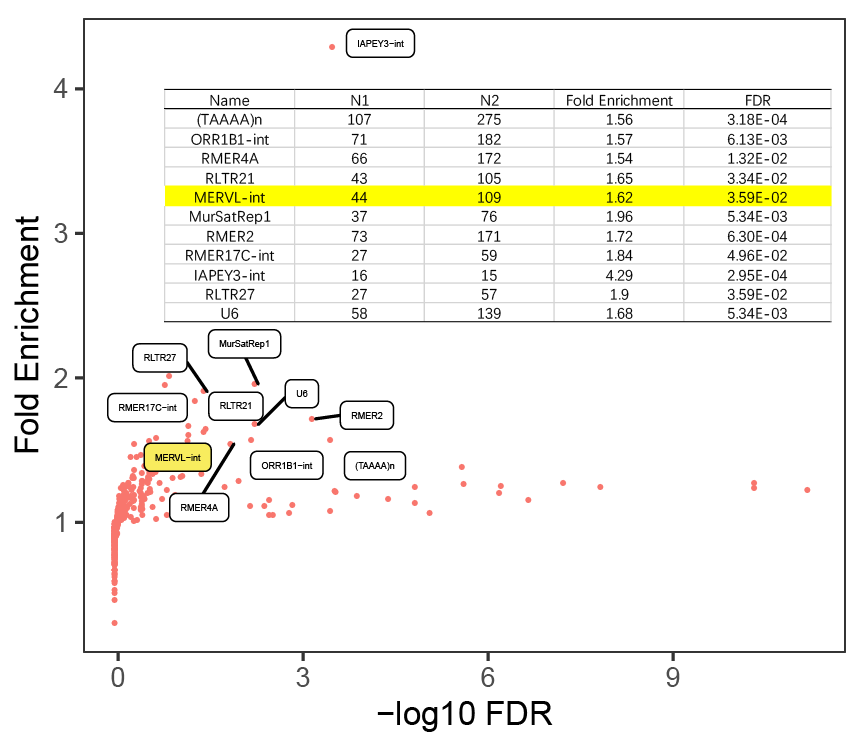
**Figure 2.** The dynamic changes of TADs during the process of NPC differentiation. (A) The sizes of TADs in D0, D1.5, and D2.5. Each group indicates each stage with the size of TADs. Here, the percentage range of the confidence interval is set to 95%. (B) The numbers of TADs in D0, D1.5, and D2.5. Each group represents each stage with the number of TADs.



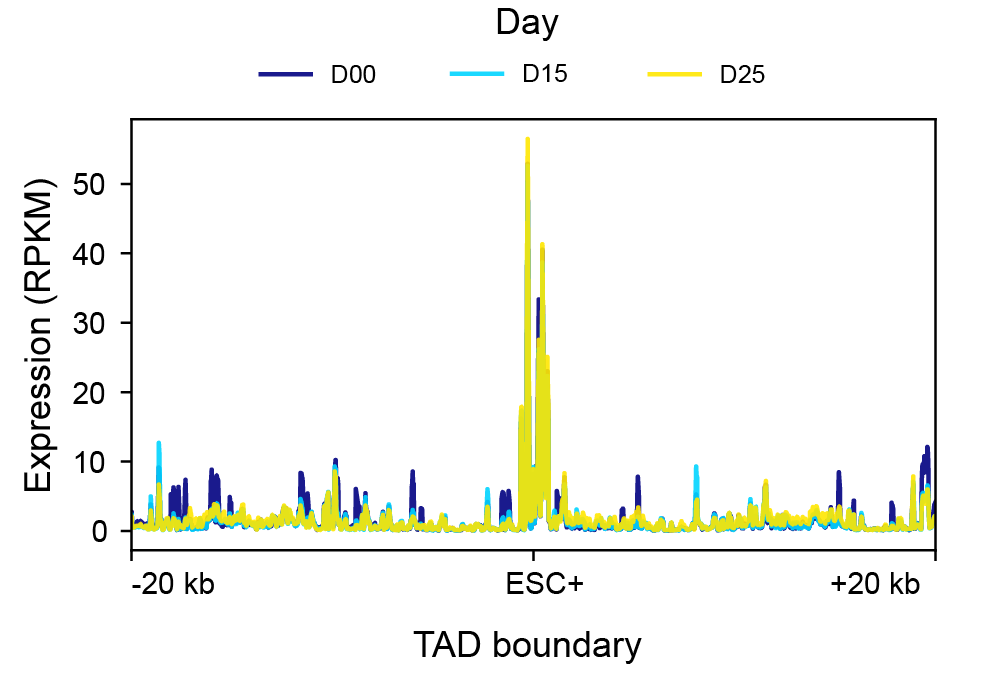
**Figure 3.** Hi-C contact matrix for NPC differentiation at 2.5 Mkb resolution in mouse chr4 chromosome. In which the red color indicates the quality of this contact larger than 30. The yellow represents TAD identified by arrowhead software.



**Figure 4.** Hi-C contact matrices for each stage of NPC differentiation at megabase resolution (2.5Mkb) in mouse chr4 chromosome. In which the red color indicates the quality of this contact larger than 30. The yellow represents TAD identified by arrowhead software.



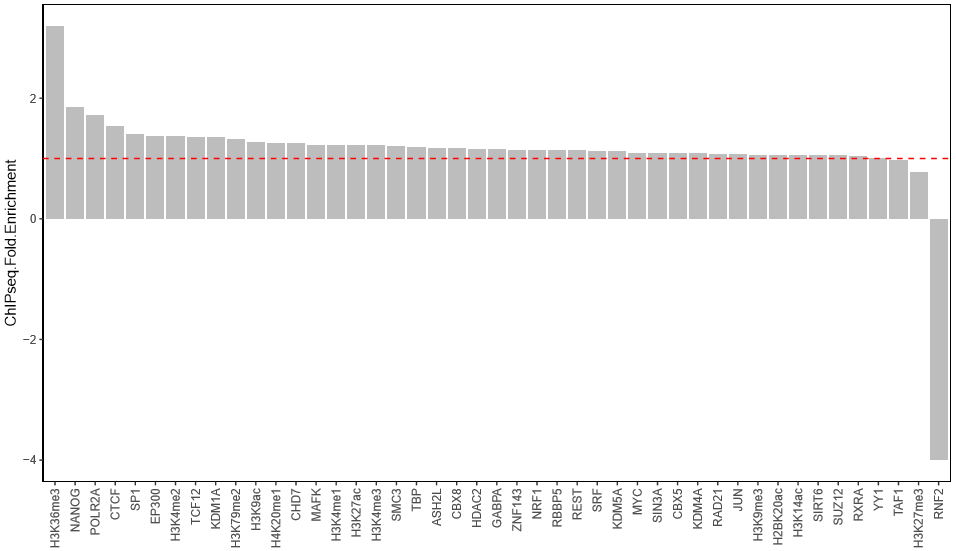
**Figure 5.** The Scatter plot and Excel table shown the fold enrichment values and -log10(q-values) for all repeat element at the ESC+ (n=565) specific TAD boundaries to the stable TAD boundaries (n=2,271). The p-values were calculated by two-sided proportion test and then were corrected by BH method.



**Figure 6.** Aggregated RNA-Seq expression profile (RPKM normalized) in 0 day, 1.5 days, and 2.5 days around ESC+ specific TAD boundaries overlapping with the MERVL repeat elements.



**Figure 7**. The motif of MERVL internal sequences is significant CTCF motif.



**Figure 8.** Fold enrichment of TFs or histone modification signals at TAD associated MERVL. The red line (value = 1) shows there is no fold enrichment.

# Tables Legends

**Table 1.** The data statistics of Hi-C data.

|  |  |  |  |
| --- | --- | --- | --- |
| Title | Day 0 | Day 1.5 | Day 2.5 |
| Accession No. | GSM2099879 | GSM2099880 | GSM2099881 |
| #Sequenced read pairs | 1,222,812,288 reads | 1,023,511,964 reads | 992,421,144 reads |
| #Alignable read pairs=normal paired+chimeric paired | 1,108,254,009 (90.63%) | 922,766,861 (90.16%) | 853,509,483 (86.00%) |
| #Unique Reads | 1,049,533,117 (85.83%) | 808,172,486 (78.96%) | 790,866,804 (79.69%) |
| #PCR Duplicates | 58,720,892 (4.80%) | 114,594,375 (11.20%) | 62,642,679 (6.31%) |
| #Hi-C Contacts | 676,477,268 (MAPQ>30) | 489,333,509 (MAPQ>30) | 497,068,272 (MAPQ>30) |
| #TADs (resolution: 10 kb) | 5,439 | 5,065 | 5,156 |

# References

1. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.

2. Eckersley-Maslin, M.A., C. Alda-Catalinas, and W. Reik, *Dynamics of the epigenetic landscape during the maternal-to-zygotic transition.* Nature Reviews Molecular Cell Biology, 2018. **19**(7): p. 436-450.

3. Xu, Q. and W. Xie, *Epigenome in early mammalian development: inheritance, reprogramming and establishment.* Trends in cell biology, 2018. **28**(3): p. 237-253.

4. Zhang, Y., et al., *Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells.* Nature genetics, 2019. **51**(9): p. 1380-1388.

5. Bonev, B., et al., *Multiscale 3D genome rewiring during mouse neural development.* Cell, 2017. **171**(3): p. 557-572. e24.

6. Perincheri, S., et al., *Characterization of the ETnII-alpha endogenous retroviral element in the BALB/cJ Zhx2 ( Afr1 ) allele.* Mamm Genome, 2008. **19**(1): p. 26-31.

7. Ge, S.X., *Exploratory bioinformatics investigation reveals importance of "junk" DNA in early embryo development.* BMC Genomics, 2017. **18**(1): p. 200.

8. Llères, D., et al., *CTCF modulates allele-specific sub-TAD organization and imprinted gene activity at the mouse Dlk1-Dio3 and Igf2-H19 domains.* Genome biology, 2019. **20**(1): p. 1-17.

9. Matthews, B.J. and D.J. Waxman, *Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver.* Elife, 2018. **7**: p. e34077.

10. Barutcu, A.R., et al., *A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus.* Nature communications, 2018. **9**(1): p. 1-11.

11. Ishiuchi, T., et al., *Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly.* Nature structural & molecular biology, 2015. **22**(9): p. 662.

12. Kruse, K., et al., *Transposable elements drive reorganisation of 3D chromatin during early embryogenesis.* bioRxiv, 2019: p. 523712.

13. Li, F., Z. An, and Z. Zhang, *The dynamic 3D genome in gametogenesis and early embryonic development.* Cells, 2019. **8**(8): p. 788.

14. Sundaram, V. and J. Wysocka, *Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes.* Philosophical Transactions of the Royal Society B, 2020. **375**(1795): p. 20190347.

15. Tan, J.Y., et al., *Splicing of enhancer-associated lincRNAs contributes to enhancer activity.* Life science alliance, 2020. **3**(4).

16. Pope, B.D., et al., *Topologically associating domains are stable units of replication-timing regulation.* Nature, 2014. **515**(7527): p. 402-405.

17. Ladstätter, S. and K. Tachibana, *Genomic insights into chromatin reprogramming to totipotency in embryos.* Journal of Cell Biology, 2019. **218**(1): p. 70-82.

18. Sima, J., et al., *Identifying cis elements for spatiotemporal control of mammalian DNA replication.* Cell, 2019. **176**(4): p. 816-830. e18.

19. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome biology, 2008. **9**(9): p. 1-9.

20. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nature protocols, 2012. **7**(3): p. 562-578.

21. Durand, N.C., et al., *Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments.* Cell systems, 2016. **3**(1): p. 95-98.

22. Yan, J., et al., *Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers.* Cell research, 2018. **28**(2): p. 204-220.