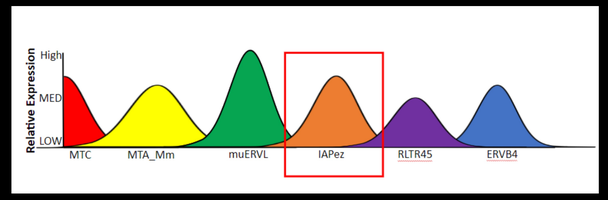
Genomic Analysis of the Effect of the IAPEz and MT2_Mm Retrotransposons on Genetic Regulation

Anmol Dash(C '26)
Department of Biomedical Sciences, University of Pennsylvania School of Veterinary Medicine
Mentor: Dr. Andrew Modzelewski
Grant: Grant for Mentoring Undergraduate Research

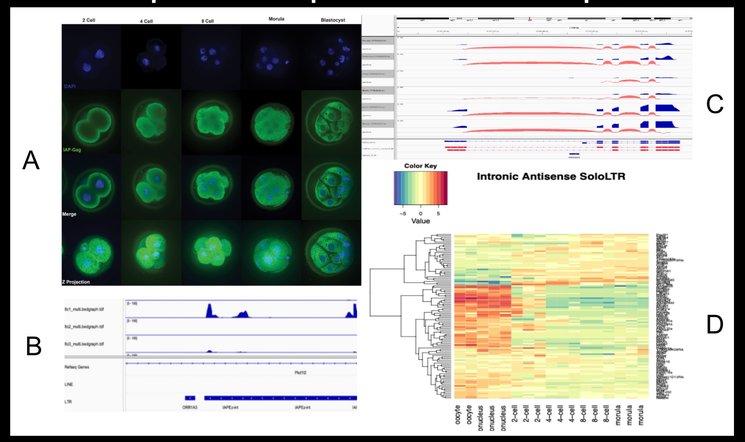
Abstract:

Long considered "junk DNA", retrotransposons make up approximately 46% of all human DNA and have recently been discovered to contribute to embryonic development and the progression of several diseases, including aging, neurodegeneration and cancers. Retrotransposons have been shown to be reliable markers of cell fate and tumorigenesis. While most investigation has been conducted into the deleterious effects of retrotransposon expression, with special focus on the LINE-1 family of retrotransposons causing deleterious insertions, relatively little investigation has been conducted into the potential beneficial effects of retrotransposons occurring through millions of years of co-evolution. Moreover, while some successful investigation has been conducted into the possibility of retrotransposon-targeted therapeutics on disease progression, those studies have largely focused on the *indirect* impact and potential of drugs targeted towards epigenetic regulation, and not *directly* on the impact of the retrotransposons themselves. Utilizing genomic indices taken from RepeatMasker, a database of millions of retrotransposon found in humans and mouse, we obtained all the sequences assigned to the IAPez and MT2_Mm transposons families that show massive expression in the early preimplantation embryo. Utilizing a variety of tools to analyze their expression in the 1-cell, 2-cell, 4-cell, 8-cell and morula stages of development, we were able to design effective probes and reagents against the MT2_Mm family. This work is done in tandem and validated by the characterization of a mouse model in which specific retrotransposon elements have been removed using CRISPR-Cas9. Globally, we were able to conduct a cluster and motif analysis of both families to reveal the adjacent genes these specific insertions possibly evolved to directly regulate during development and open the door for further exploration into the role of these sequences in the genome.



Schematic to show the expression pattern of various retrotransposons families across preimplantation development. Families are composed of 100s to 10,000s of nearly identical insertions throughout. Our two areas of focus are the IAPEz and MT2_Mm(muERVL) families.

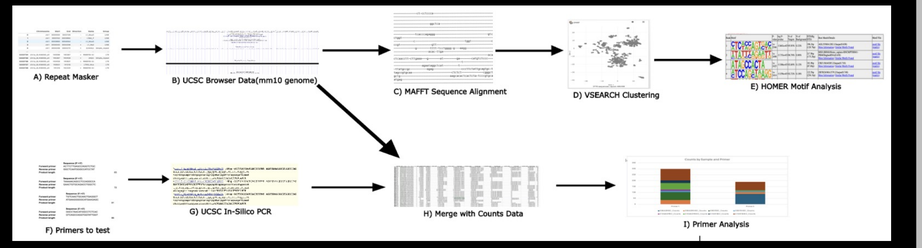
Pre-Implantation Expression of Retrotransposons



(A) Indirect Immunofluorescent localization of IAP elements throughout mouse preimplantation embryo development. Embryos were fixed and stained using immunofluorescence to monitor the expression and localization of the IAP elements. (B) Published software and custom designed software were developed to identify specific genomic coordinates of transposon expression. An example of an IAPez element in an intron of the gene Pkd12 is shown as an example. (C) Using the same software as (B), the expression of sample MT2_Mm and Oct4 retrotransposon elements are shown as an example. Note the consistent patterns of expression across samples. (D) The statistical correlation of the expression of genes near MT2_Mm elements with MT2_Mm elements is shown. Note the high correlation shown near the oocyte stage, suggesting some level of MT2_Mm gene regulation.

- Transposable elements take up nearly half of the human genome and are hypothesized to have originated from viruses that integrated into organismal host DNA.
- Out of 4 million discovered, only a small fraction are predicted to produce a protein.
- We discovered that the MT2_Mm (muERVL) and IAPez subfamilies follow a precisely regulated expression pattern at the 2C and 4C/8C stage, respectively.
- The sequence clustering and expression of both the IAP and MT2_Mm has been relatively unexplored but are both dense source of transcription factor binding sites.
- Very little guidance on designing primers targeting highly repetitive elements (IAP and MT2_Mm subfamilies) can be found in the literature, as they exist as hundreds to thousands of nearly identical copies.
- Even less is known about the heterogeneity of individual insertion under a single subfamily.

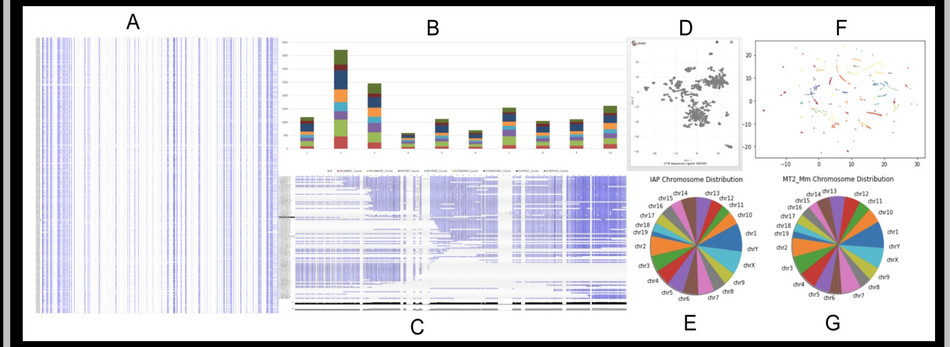
Clustering and Primer Evaluation Pipeline



(A) is a sample representation of the IAP coordinates taken from Repeat Masker. (B) is the UCSC visualization of a random MT2_Mm element. (C) is a FASTA result of aligning all IAPez sequences using MAFFT. (D) is a representation of the available MT2_Mm sequences using UMAP. (E) is a portion of the HOMER results using the IAPez subfamily. (F) is a sample set of primers used to test the primer evaluation pipeline. (G) is a sample in-silico result using the first of the primers in (F). (H) is a section of the merged results data with the counts information. (I) is a chart comparing the expression levels of primer targets, with the color indicating steps in development.

To obtain the target candidate genes for experimentation, we examined existing annotations on the mm10 genome from Repeat Masker, through which a large number of sequences were isolated. The utilization of the sequences was two-fold, first to determine a consensus sequence and analyze for potential motifs, and second to test potential primer sets for overlap and efficacy along the entire genome. The first pathway required a significant amount of preprocessing, removing outliers in terms of size and updated annotations and various irrelevant errors. We then attached the repeat masker counts to existing counts data and aligned them using MAFFT. Once we were able to align them, we used HOMER to identify motifs across the sequences and compare them to existing databases of mammalian motifs. Separately, we also clustered the sequences using VSEARCH and UMAP and re-aligned them within the retrotransposon families. The second pathway involved primers taken from arbitrarily selected sequences in IGV Browser. From there, we evaluated the success of these primers against the sequences taken from Repeat Masker, with attention to the counts covered by the primer relative to the total produced.

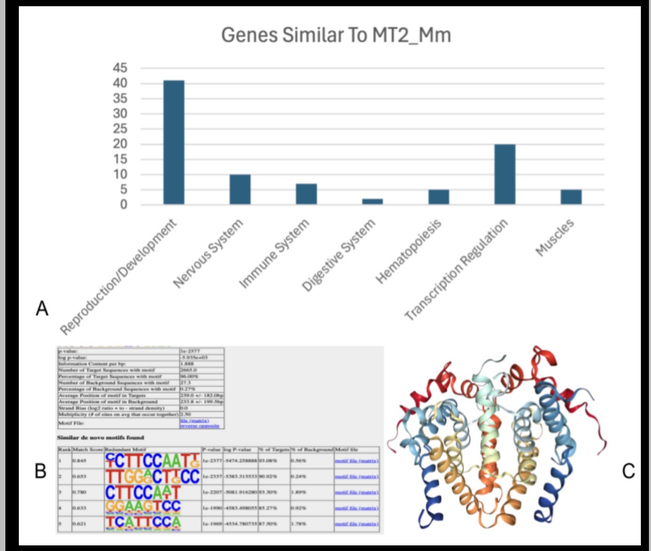
Quantifying the Impact



(A) A visualization of the alignment of the MT2_Mm family. We can see that after clustering and re-alignment (C) that the number and size of gaps is significantly reduced. (B) A sample result of the primer evaluation pipeline, with primer 3 showing the highest number of counts produced within the MT2_Mm family. (C) A visualization of a sample cluster identified in (D) from the MT2_Mm family. (D) A cluster map of the MT2_Mm family created using UMAP, revealing six distinct clustered groups for MT2_Mm. (E) Chromosomal distribution of the IAP family. Both this and the distribution of MT2_Mm genes(G) show no overrepresentation of a particular chromosome. (F) is a cluster map of the IAP family, revealing a variety of individual clustered groups. (G) The chromosomal distribution of the genes in the MT2_Mm family.

Acknowledgments
Many thanks to Dr. Andrew Modzelewski, Vamshidhar Nallamalla, and Claire An for providing the mentorship, information, and resources needed to complete this project. This work was funded in part from the Grant for Mentoring Undergraduate Research.

Results of Clustering and Motif Analysis



Conducting a motif analysis of the MT2_Mm family via the HOMER package, we found several motifs of interest (sample image of HOMER interface shown in (B)). The program also identified 100 mammalian genes of interest, whose transcription factor binding sites match MT2_Mm, 41 of which have functions relating to reproduction. The distribution of roles of the mammalian genes is shown in (A). One of those genes is the COUP-TFII gene, which has roles in organogenesis and cell fate differentiation, whose protein product is shown in (C).

Summary

- Repeat Masker provides a path through which we can analyze the role of retrotransposon elements in the mammalian genomes and evaluate the efficacy of primers against retrotransposons.
- The utilization of MAFFT, elimination of outliers, and selection of the highest expressed sequences is a viable way to align gene sequences and analyze similarities.
- VSEARCH is a viable way to obtain sequence clusters from retrotransposon subfamilies
- Motifs present in MT2_Mm elements are also present in mammalian reproductive regulatory genes, a connection which merits further exploration.