Identifying Repressive RNAs with *Xist*-Like Functions in the Mouse Transcriptome

By
Zhiyue Zhang

Senior Honors Thesis
Department of Biology
University of North Carolina at Chapel Hill

November 25, 2022

Approved:

Dr. Mauro Calabrese, Thesis Advisor

Dr. Michael Emanuele, Reader

Dr. Terry Furey, Reader

Abstract

Long non-coding RNAs (lncRNAs) play major roles in gene regulation and human health. *Xist*, the lncRNA responsible for X chromosome inactivation, has been studied for 30 years, but only a few repressive lncRNAs with similar functions have been discovered, including *Airn* and *Kcnq1ot1*. One notable obstacle in lncRNA discovery is a poor understanding of the lncRNA sequence-function relationship. To study this problem, we hypothesized that there exists a class of repressive RNAs that share 2 sequence features: k-mer content (the abundance of short motifs), and protein-binding profile (binding affinity with proteins important for *Xist* function). In my analysis of mouse trophoblast stem cells, I compared the k-mer content and protein-binding profile of expressed, chromatin-enriched RNAs to *Xist*, *Airn*, and *Kcnq1ot1*. I used the algorithm SEEKR to evaluate k-mer similarities and a computational pipeline I created to calculate protein-binding similarities. I found that *Airn* and *Kcnq1ot1* share more similarity in protein-binding profile and k-mer content than *Xist*. Furthermore, k-mer content correlates with protein-binding positively for similarities to the unspliced *Airn* and *Kcnq1ot1*, but negatively to the spliced *Xist*. We infer that such contradiction to our hypothesis is due to the confounding effect of splicing, suggesting the need to further tailor the list of RNAs in search of gene repressors. Future directions include removal of RNAs with short half-life in the chromatin region, analysis of regional k-mer content, and biochemical testing of true functions of candidate RNAs.

Introduction

Genomes store the instructions to produce and maintain organisms. To understand the execution and regulation of such information, DNA transcription and translation have been extensively studied, with much attention given to proteins and coding RNAs. However, only 2% of human genes are protein-coding; 98% of the human genome encodes for non-coding RNAs that are translated into little to no proteins (Dunham et al., 2012). A major class of non-coding RNAs are long non-coding RNAs (lncRNA) longer than 200 base pairs (Diamantopoulos et al., 2018). Research on lncRNAs in the past decade have shown their key roles in gene regulation: through interaction with chromatins, RNAs, and proteins, they modulate multiple levels of gene expression, such as transcription, splicing, translation, RNA stability, as well as the localization and function of proteins. The malfunction of lncRNAs and the consequent misregulation of genes are associated with a wide range of human diseases: multiple types of cancer, as compiled in databases including Lnc2Cancer and the Cancer LncRNA Census; developmental disorders, such as the Rett syndrome and autism; brain disorders, such as Huntington's disease and Alzheimer's disease; disorders in the heart, immune system, and so on (Petazzi et al., 2013; Statello et al., 2021; Tang et al., 2017). The most potent lncRNA discovered so far is *Xist*. *Xist* is responsible for X-chromosome inactivation - the process of inactivating one of the two X chromosomes in female mammal embryonic cells. By wrapping around the X chromatin, *Xist* recruits chromatin-modifying proteins to repress transcription, thus achieving dosage compensation - equal amount of gene

expression between males and females (Brown et al., 1992). *Airn* and *Kcnq1ot1* are 2 repressive lncRNAs with *Xist*-Like functions: *Airn* transcription induces silencing of an insulin-like growth factor receptor important for healthy metabolism; *Kcnq1ot1* is necessary to silence multiple genes that are normally paternally repressed and involved in the pathogenesis of Beckwith-Wiedemann Syndrome and colorectal cancer (Latos et al., 2012; Mancini-DiNardo et al., 2006).

Considering such essential roles lncRNAs play in gene regulation and human health, accurate annotations of lncRNAs would advance biomedical research, such as studies of lncRNAs as disease etiologies, biomarkers, and novel therapeutic tools targeting gene expression. However, numerous lncRNAs lack complete identification and functional characterization; for the lncRNAs already mapped to the genome, the proportion of experimentally characterized or disease-associated lncRNAs is less than 1%. The significant lag of lncRNA annotations behind those of protein-coding genes results from a variety of reasons: the relatively low expression level of lncRNAs, the weak conservation of lncRNA sequences during evolution, and our poor understanding of the lncRNA sequence-function relationship (Uszczynska-Ratajczak et al., 2018). Prominently, despite abundant evidence of *Xist*'s powerful and important biological function, only a couple of *Xist*-like lncRNAs have been found after more than 30 years of its discovery, and all of them were discovered by chance.

To contribute to understanding the lncRNA sequence-function relationship, we aimed to develop methods that predict lncRNA function based on its sequence alone.

To search for novel *Xist*-like RNAs, we hypothesized that there exists a class of repressive, *Xist*-like lncRNAs that share 2 sequence features: k-mer content and protein-binding profile. K-mers are short motifs of length k within a genetic sequence. K-mer content refers to the frequency of every k-mer for a given RNA sequence and a given length of k. The other feature, protein-binding profile, refers to the binding signals or affinities between an RNA and RNA-binding proteins; in this project, we studied the interaction between RNAs and 27 proteins essential for *Xist* gene-silencing, as outlined in the Methods section. Specifically, we hypothesized that *Xist*, *Airn*, and *Kcnq1ot1* have highly similar k-mer content and protein-binding profile, and that k-mer content positively correlates with protein-binding profile such that RNAs with k-mer content more similar to *Xist*, *Airn*, and *Kcnq1ot1* are more likely to display protein-binding signals similar to these 3 model repressive lncRNAs.

To this end, I performed computational analysis on the sequencing data from mouse trophoblast stem cells (TSCs). To include the possibility that some RNAs not currently annotated as lncRNA might also act as gene repressors, I selected all RNAs that are expressed and enriched within the chromatin region in TSCs. I computed the k-mer similarities between these selected RNAs and 3 model lncRNAs (*Xist*, *Airn*, and *Kcnq1ot1*) using an algorithm previously developed by the Calabrese Lab called SEEKR (Kirk et al., 2018). I then developed a computational pipeline that calculates the protein-binding profiles of RNAs based on a large panel of RNA immunoprecipitation data with proteins important for *Xist* functions. As a result, I found that *Airn* and *Kcnq1ot1* share more similarity in protein-binding profile and

k-mer content than *Xist*. Furthermore, k-mer content and protein-binding profile were positively correlated for similarities to the unspliced *Airn* and *Kcnq1ot1*, but negatively correlated for similarities to the spliced *Xist*. As of *Xist*, we infer that this contradiction of our hypothesis may be due to the confounding effect of protein-mediated RNA splicing. In conclusion, we recognize the need to further tailor our selection of candidate RNAs and calculation of sequence features. Future directions include incorporation of half-life data to filter for RNAs that function in the chromatin region for a relatively long time, computation of regional k-mer content, and eventually, biochemical experimentation of the true functions of our final list of candidate RNAs that are sequentially similar to *Xist*.

Methods

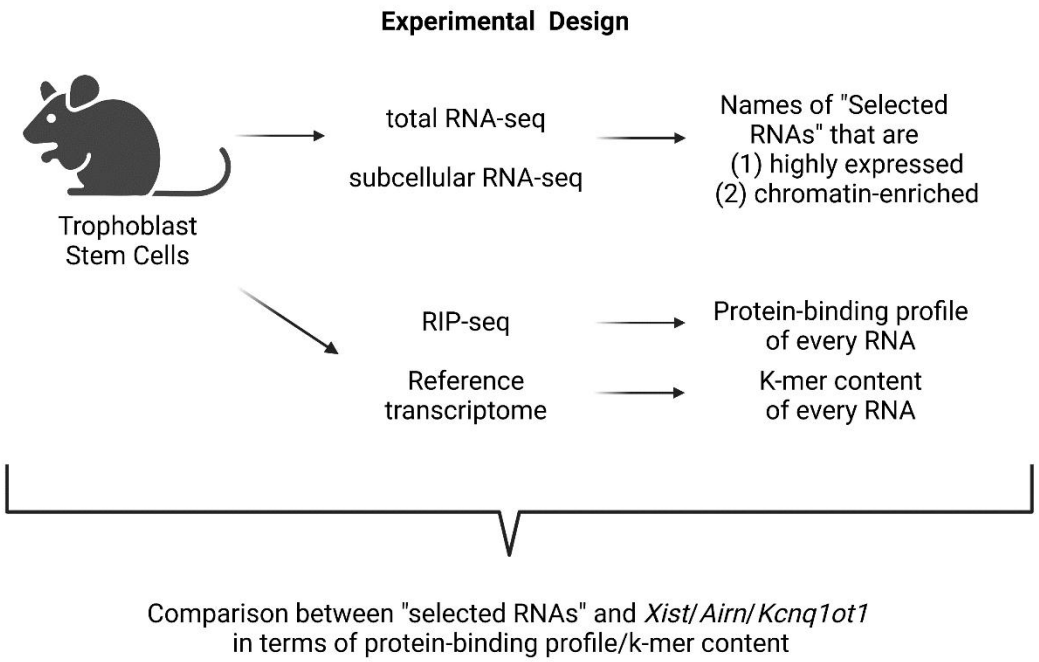


Fig 1. Overview of experimental design.

Cell model: mouse trophoblast stem cells (TSCs)

All data used in this study were generated from mouse trophoblast stem cells (TSCs), which are cells derived from the trophectodermal stem cell population that mediates implantation and gives rise to the placenta. In TSCs, *Xist* was known to be extremely active performing gene silencing, thus offering a good standard of interaction levels between RNAs and RNA-binding proteins (Calabrese et al., 2015).

Reference genome and transcriptome

The reference genome used in this study was the Genome Reference Consortium Mouse Build 38, mm10 (Church et al., 2011). The reference transcriptome used was the GENCODE mouse build vM5 (Frankish et al., 2019).

Selecting expressed, chromatin-enriched RNAs: total and subRNA-seq data

To select only RNAs that are highly expressed and chromatin-enriched in TSCs, I used total RNA-seq and subcellular RNA-seq (subRNA-seq) data to determine thresholds of expression and chromatin enrichment.

Total RNA-seq detects both coding and non-coding RNA; since non-coding RNAs are important to this study, total RNA-seq provides an appropriate threshold of transcript abundance. The total RNA-seq data used in this study were previously published by the lab (Schertzer et al., 2019).

SubRNA-seq detects sequencing of RNA after cellular fractionation of tissue samples; since this study aimed to find *Xist*-like RNAs involved in gene regulation

(interacting with chromatin), we selected for RNAs enriched in the chromatin region relative to the cytoplasm. The subRNA-seq data of chromatin and cytoplasm were generated by Mickey Murvin from the Calabrese Lab.

To determine thresholds of expression and chromatin enrichment, I aligned both total and subRNA-seq data to the mouse transcriptome and computed the total, cytoplasmic, and chromatinic abundance of each RNA in units of the normalized TPM (transcripts per million) using Kallisto, a program for quantifying abundances of transcripts from RNA-seq data (Bray et al., 2016). This analysis was done on UNC's Longleaf, a Linux-based computing system. I then plotted the total RNA abundance in histogram. Since many transcripts were not expressed (TPM = 0), I used log transformation to correct the skewed data. I plotted the expression levels in units of $\log_2(\text{TPM} + 0.001)$; 0.001 was added only to prevent taking the undefined logarithm of zero and had minimal effect on the overall distribution. I will refer to this unit as $\log_2(\text{TPM})$ in the Results section. Based on the histogram, I determined the threshold of high expression to be a value approximately around the inflection point, removing lowly expressed transcripts that were likely weakly functional and/or highly unstable (Fig. 3a).

Chromatin enrichment of each RNA was defined as:

$$Chromatin\ Enrichment\ = \frac{Chromatinic\ TPM}{Cytoplasmic\ TPM\ +\ Chromatinic\ TPM}$$

I plotted the chromatin enrichment in histogram and determined the threshold of chromatin enrichment to be a value approximately around the inflection point, which excluded transcripts that likely function in the cytoplasm (Fig. 3b). Plotting both

histograms were performed in Python.

In summary, RNAs that satisfied these 2 conditions were selected for further analysis: (1) total abundance > threshold of high expression (2) chromatin enrichment > threshold of chromatin enrichment.

Detecting RNA-protein interactions: RNA immunoprecipitation sequencing

Protein-binding profile refers to the binding signals or affinities between an RNA and RNA-binding protein (RBP), namely the strength of RNA-protein interactions. RNA-protein interactions were measured through RNA immunoprecipitation sequencing (RIP-seq). RIP-seq differs from RNA-seq in one essential step of antibody pulldown: after the chromatin was sheared, antibodies of RBP were used to immunoprecipitate the RBP of interest together with the bound RNAs; the bound RNAs would then be purified and sequenced.

In this study, Mickey Murvin from the Calabrese Lab performed RNA immunoprecipitation on TSCs, using a different RBP antibody each time. RIP-seq data were generated for 27 different RBPs: ALY/REF, G9A, HNRNPC, HNRNPK, HNRNPM, HNRNPU, JARID2, LBR, MATR3, NUDT21, PABPN1, PTBP1, RBM15, RING1B, RYBP, SAFB, SPEN, SRSF1, SUPT16H, SUZ12, U2AF35, XRN2, TIA1, CIZ1, U2AF65, NXF1, SFPQ. Past research has identified all these RBPs as *Xist*-binding proteins playing key roles in *Xist*-mediated gene silencing (Chu et al., 2015). A control dataset with Immunoglobulin G (IgG) was also generated. IgG is a non-specific antibody that estimates the background noise levels and was used for

normalization against non-specific binding.

Computing protein-binding profile

To quantify RNA-protein interactions, I created a computational pipeline that reads as input the RIP-seq data from a certain RNA-binding protein (RBP) antibody and the control (IgG) and produces as output the binding signals between the given RBP and RNAs of the reference transcriptome. I ran the pipeline on Longleaf for all 27 groups of RIP-seq data, each targeting a different RBP. The final output was a matrix of binding signals between the reference transcriptome and the 27 RBPs; each matrix entry identified the strength of interaction between an RNA-protein pair.

The workflow of this pipeline consists of 5 steps (Fig 2). Details of the programs used follow. RIP-seq data are short reads in the format of FASTQ files. First, to know where in the genome each read came from, I aligned the RBP and IgG RIP-seq data separately to the mouse genome using STAR, an aligner for RNA-seq data (Dobin et al., 2013).

To find areas of enriched RNA-protein binding, I performed strand-specific peak-calling for RBP alignment using MACS, a peak-calling program (Zhang et al., 2008). Since MACS is typically used for ChIP-Seq (DNA-protein binding), necessary modifications were made. For DNA-protein binding, a peak contains reads from both strands; for RNA-protein binding, a peak should be specific to one strand, because reads on different strands indicate binding activities with different RNAs, such as *Xist* and *Tsix*, the RNA antisense to *Xist*. Therefore, to adapt RIP-seq alignment to MACS,

I divided the alignment from STAR into 2 groups by strand, randomized the strand information within groups, and called strand-specific peaks with MACS. This step removed some of the false positive binding signals and produced the exact genomic coordinates within which enriched binding was detected.

I then analyzed which MACS peaks were enriched over IgG. This step was necessary because there existed varying levels of non-specific binding in different regions of the genome. Since IgG is a non-specific antibody, IgG RIP-seq data were used as a control against background noise. For both RBP and IgG alignment, I counted the reads assigned to each MACS peak using featureCounts, a read summarization program for "features" of genomic regions (Liao et al., 2014). To compare RBP reads against IgG reads, I also normalized for sequencing depth, which is a technical bias of sequencing—when more sequencing depth produces more read count for gene expressed at same level. I converted the raw counts assigned to each peak to RPM (reads per million mapped reads), and found the peaks enriched over IgG based on the RPM values of RBP and IgG, defined as following:

$$Peaks\ enriched\ over\ IgG, or\ "true"\ peaks:$$

$$RBP\ RPM\ at\ this\ peak > 2 \times IgG\ RPM\ at\ this\ peak$$

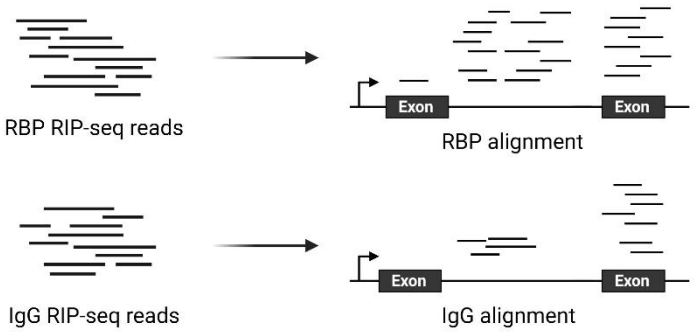
Next, I extracted the RIP-seq reads of RBP and IgG that overlapped with these "true" peaks using Samtools, a suite of programs for interacting with high-throughput sequencing data (Danecek et al., 2021). This step converted the genomic coordinates of the "true" peaks back to short reads ready for transcriptome alignment.

Lastly, to know which transcripts were bound to this RBP and quantify the

strength of such interactions, I aligned the short reads at the "true" peaks to the mouse transcriptome using Kallisto, a program for quantifying abundances of transcripts from RNA-seq data (Bray et al., 2016). Kallisto computed the estimated counts for each RNA, which I converted to RPM to normalize for sequencing depth. Finally, to account for the noise of non-specific binding, I subtracted the RPM of IgG from the RPM of RBP for each RNA. The final values of RPM defined the binding signal between RNAs and the RBP.

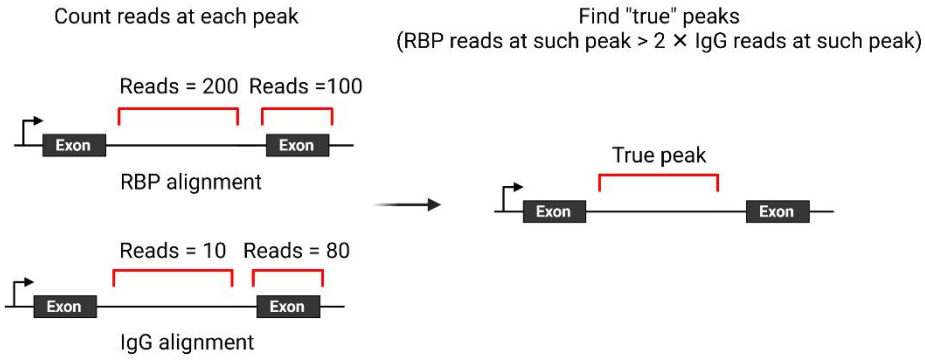
$$Binding\ signal\ between\ an\ RNA\ and\ a\ RBP = RPM\ of\ RBP\ for\ this\ RNA - RPM\ of\ IgG\ for\ this\ RNA$$

(1) Align RIP-seq reads of RBP and IgG to the genome

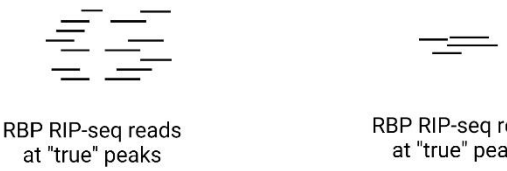


(2) Call peaks of protein-binding based on RBP alignment

(3) Find "true" peaks enriched over IgG



(4) Extract RIP-seq reads at "true" peaks



(5) Align RIP-seq reads at "true" peaks to the transcriptome

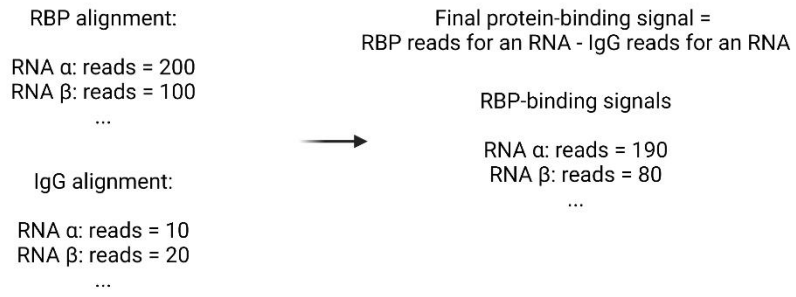


Fig 2. Schematic of computational pipeline generating protein-binding profile to quantify RNA-protein interactions. RBP = RNA-binding protein of interest.

Calculating and comparing k-mer content: SEEKR

K-mers are short motifs of length k within a genetic sequence. K-mer content refers to the frequency of every k-mer for a given RNA sequence and a given length of k. To compare the k-mer content between selected RNAs (highly expressed and chromatin-enriched) and the 3 model repressive RNAs (*Xist*, *Airn*, *Kcnq1ot1*), I used the algorithm SEEKR, a sequence comparison method previously developed by the Calabrese Lab (Kirk et al., 2018). A motif length of k=6 was used, which, based on the Kirk paper, had the best predictive power of *Xist*-like repression compared with other k-mer length.

SEEKR performed the analysis in several steps. First, it counted the occurrence of all 6-mers in one-nucleotide increments across each RNA. Second, it normalized the 6-mer counts for each RNA by RNA length and standardized across the group (z-score conversion). Lastly, it determined the relative similarity between each selected RNA and the 3 model repressive RNAs via Pearson's correlation.

Correlating protein-binding profile with k-mer content

K-mer similarities of selected (expressed, chromatin-enriched) RNAs to *Xist*, *Airn*, and *Kcnq1ot1* were calculated using SEEKR. To calculate protein-binding

similarities, I standardized across the binding signals of each RBP (z-score conversion) and determined the similarity between each selected RNA and *Xist*, *Airn*, and *Kcnq1ot1* via Pearson's correlation. As a result, each RNA had 6 values associated with it: k-mer similarity to *Xist*, protein-binding similarity to *Xist*, k-mer similarity to *Airn*, protein-binding similarity to *Airn*, k-mer similarity to *Kcnq1ot1*, and protein-binding similarity to *Kcnq1ot1*.

To visualize how similar *Xist*, *Airn*, and *Kcnq1ot1* were in terms of k-mer and protein-binding, I plotted histograms of each of the 6 values and highlighted similarities between these 3 model lncRNAs.

To evaluate the correlation between k-mer content and protein-binding profile, for each of *Xist*, *Airn*, and *Kcnq1ot1*, I created a scatterplot of k-mer similarities versus protein-binding similarities and calculated the Pearson's correlation coefficient between the 2 groups. This part was performed in R.

Analyzing Splicing in Top-Ranked RNAs

To analyze the sequence feature of splicing in RNAs most similar to *Xist*, I extracted top-ranked RNAs in the 99.5th percentile of k-mer similarities to *Xist*, and performed binomial tests on the number of unspliced RNAs. This part was performed in R.

Results

Selection of highly expressed, chromatin-enriched transcripts in search of Xist-like gene repressors

Xist is a repressive lncRNA essential for human health, but only a few RNAs with similar functions have been found, notably *Airn* and *Kcnq1ot1*. We have hypothesized that there exists a larger class of such *Xist*-like RNAs with repressive functions. To search for such *Xist*-like RNAs, I investigated the transcriptome of mouse trophoblast stem cells (TSCs), where *Xist* has high level of gene-silencing activity (Calabrese et al., 2015). An active gene regulator must be expressed and would likely localize to the chromatin. Therefore, I used total RNA-seq data to determine the threshold for expression (Fig. 3a) and subcellular RNA-seq data for chromatin enrichment (Fig. 3b) and then filtered for transcripts that were highly expressed and chromatin-enriched. The mouse reference transcriptome contains more than 136,000 RNAs; after filtering, I selected 10% of these transcripts for further analysis. To account for the incompleteness of RNA annotation and the possibility that some RNAs without the label of lncRNA might also function as gene repressors, I did not limit the search to only lncRNAs. In sum, I selected a list of highly expressed, chromatin-enriched transcripts that more likely function as gene repressors than the rest of the mouse transcriptome.

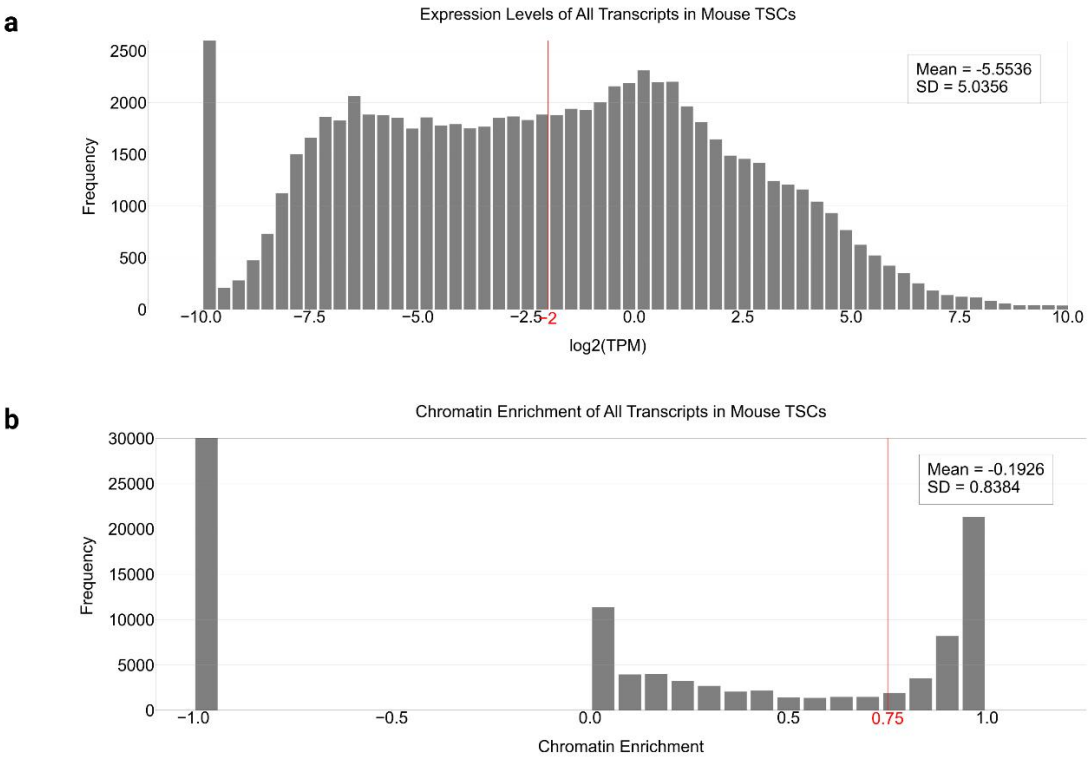


Fig 3. Selection of highly expressed, chromatin-enriched transcripts.

a, Expression levels of all transcripts in mouse TSCs (trophoblast stem cells), in units of $\log_2$(TPM). TPM = transcript per million. The red vertical line indicates the threshold determined for high expression. **b**, Chromatin enrichment levels of all transcripts in mouse TSCs. The red vertical line indicates the threshold determined for chromatin enrichment.

Protein-binding profile quantifies RNA-protein interactions

The discovery of novel lncRNAs is hindered by a poor understanding of the lncRNA sequence-function relationship (Uszczynska-Ratajczak et al., 2018); in other words, it is not known what sequence features unite lncRNAs of a similar function. We have hypothesized that *Xist*-like repressive RNAs share 2 sequence features: k-mer content and protein-binding profile. K-mer content refers to the frequency of every k-mer (short motif with length k) for a given RNA sequence and a given length of k. For each selected RNA (highly expressed and chromatin enriched), I calculated its k-mer content using the algorithm SEEKR (Kirk et al., 2018).

The other feature, protein-binding profile, refers to the binding signals or affinities between an RNA and RNA-binding proteins (RBPs). To search for *Xist*-like transcripts, I analyzed the RNA immunoprecipitation sequencing (RIP-seq) data of 27 *Xist*-binding proteins and Immunoglobulin G (IgG), the control against background noise (non-specific binding). I developed a computational pipeline that computed the protein-binding profile for the whole mouse transcriptome based on RIP-seq data of the RBP of interest and IgG (Fig. 2). This pipeline consists of the following steps that corrected for several technical biases of immunoprecipitation and sequencing. RIP-seq reads were aligned to the mouse genome. To reduce false positive binding signals, areas of concentrated alignment were found as peaks on each strand. To find areas of specific binding over background noise, RIP-seq reads within each RBP peak were counted for RBP and IgG, and only peaks with RBP reads enriched over IgG reads were kept as “true” peaks. RIP-seq reads were then extracted at these “true”

peaks and aligned to the mouse transcriptome. Finally, for each transcript, the count of IgG reads was subtracted from that of RBP reads to reduce background noise. Sequencing depth was accounted for through unit conversion to reads per million (RPM). Length normalization was avoided, considering that length is a meaningful factor that potentially allows lncRNAs to interact more with proteins. As a result, this computational pipeline generated protein-binding profiles that quantified the interactions between *Xist* RBPs and the whole transcriptome, which was further filtered to include only selected RNAs.

Airn and Kcnq1ot1 share more similarity in protein-binding profile and k-mer content than Xist

K-mer contents and protein-binding profiles were calculated for all selected RNAs. We hypothesized that we would observe similar k-mer contents and protein-binding profiles for *Xist*, *Airn* and *Kcnq1ot1*. I computed the similarity of k-mer content and protein-binding profile between selected RNAs and *Xist*, *Airn*, and *Kcnq1ot1* using Pearson's correlation (Fig. 4). *Airn* and *Kcnq1ot1* showed high similarity in protein-binding profile and k-mer content: for both features, *Kcnq1ot1* was more similar to *Airn* than 95% of the selected RNAs; *Airn* was more similar to *Kcnq1ot1* than 90% of the selected RNAs. In comparison, *Airn* and *Kcnq1ot1* were less similar to *Xist*: for both features, *Kcnq1ot1* was more similar to *Xist* than 75% of selected RNAs; *Airn* was more similar to *Xist* than 78% of selected RNAs for protein-binding profile, but only 52% for k-mer content.

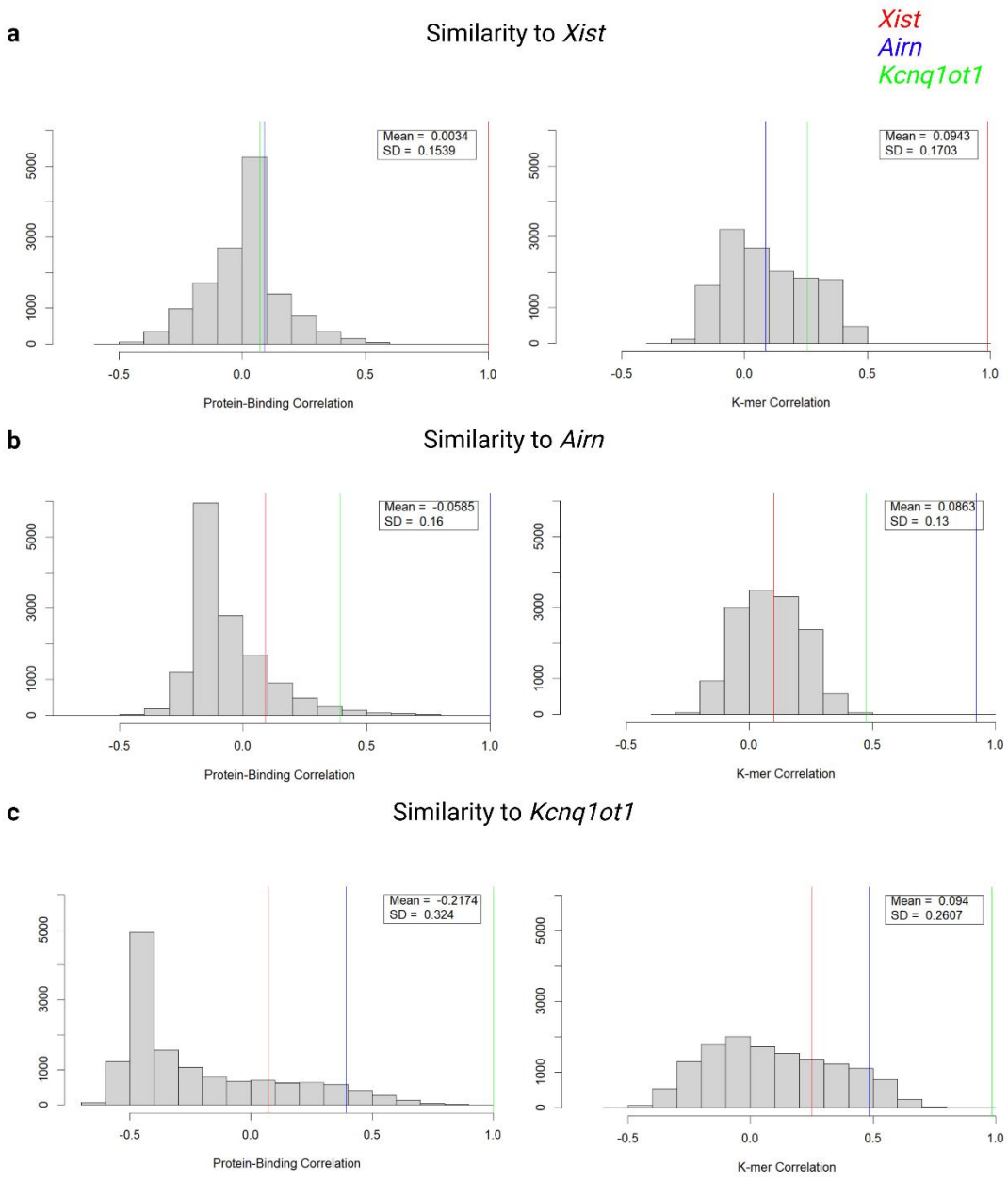


Fig 4. Similarity between selected RNAs and model repressive lncRNAs in protein-binding profile and k-mer content. a, Similarity between *Xist* and selected RNAs in protein-binding profile (Left) and k-mer content (Right). The three vertical lines are correlation values for *Xist* (red), *Airn* (blue) and *Kcnq1ot1* (green). **b**, Same as (a) but for *Airn* and selected RNAs. **c**, Same as (a) but for *Kcnq1ot1* and selected RNAs.

Protein-binding correlates with k-mer positively for similarity to Airn and Kcnq1ot1, but negatively for similarity to Xist

Similarities of k-mer content and protein-binding profile to *Xist*, *Airn*, and *Kcnq1ot1* were evaluated for all selected RNAs. We hypothesized to observe a positive correlation between k-mer and protein-binding similarities; in other words, RNAs with more *Xist*-like k-mer content would have more *Xist*-like protein-binding profile, and vice versa. I calculated the Pearson's correlation coefficients between k-mer and protein-binding similarities to *Xist*, *Airn*, and *Kcnq1ot1* (Fig 5). For similarity to *Xist*, k-mer negatively correlated with protein-binding. For similarity to *Airn* and *Kcnq1ot1*, k-mer positively correlated with protein-binding.

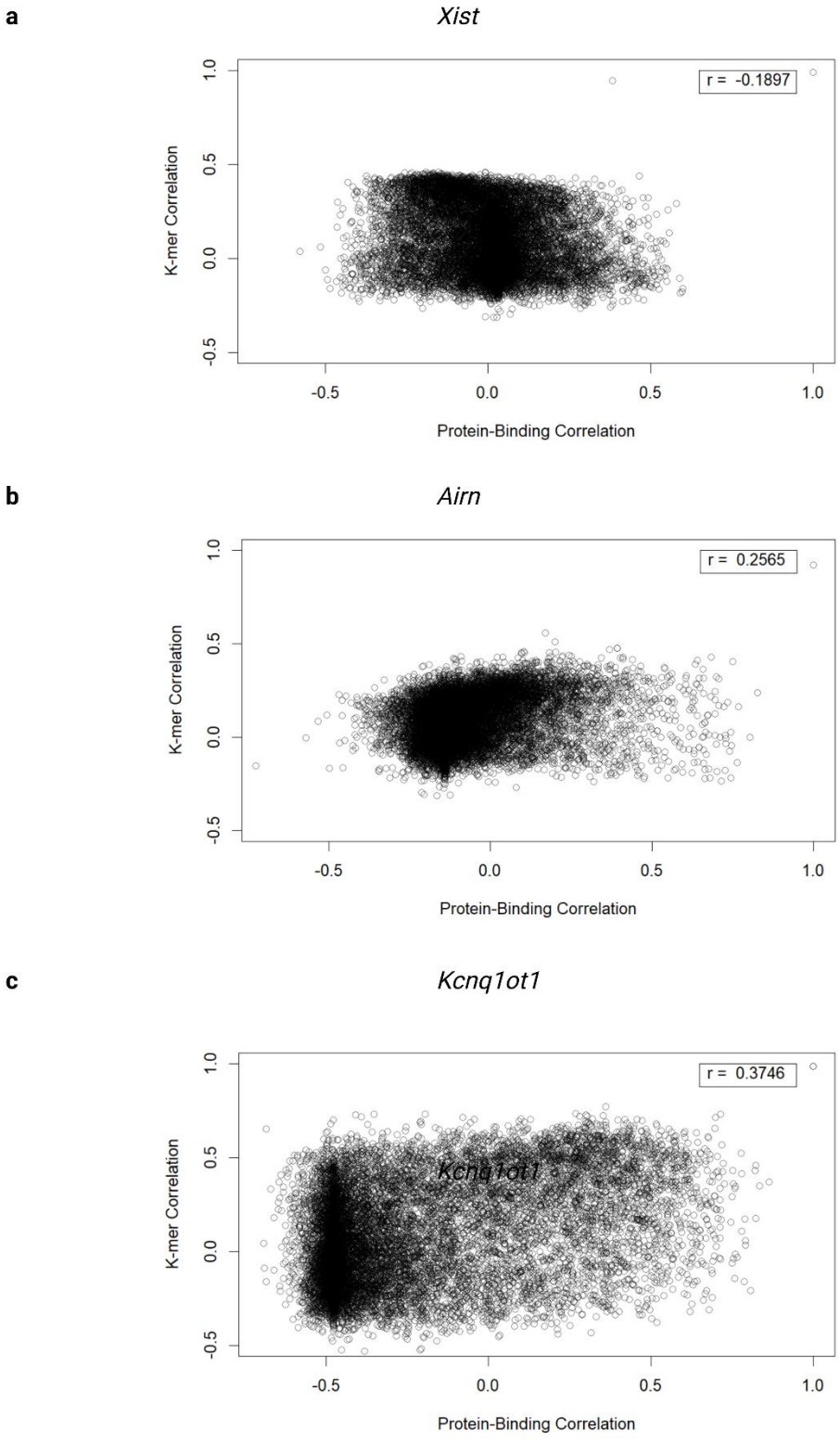


Fig 5. Correlation between similarities to model repressive lncRNAs in protein-binding profile and k-mer content. a, Similarities to *Xist*. **b**,

Similarities to *Airn*. **c**, Similarities to *Kcnq1ot1*. Pearson's correlation coefficients are shown here as "r".

Discussion

Collectively, my data support that within the expressed, chromatin-enriched transcripts of mouse trophoblast stem cells, k-mer content correlate with protein-binding profile positively for similarities to *Airn* and *Kcnq1ot1*, but negatively for similarities to *Xist*. For these 3 lncRNAs known to have similar repressive functions, *Airn* and *Kcnq1ot1* share more similarity in k-mer and protein-binding with each other than with *Xist*.

Previously, we have hypothesized to observe positive correlations between k-mer and protein-binding for all 3 lncRNAs, and high similarities within these 3 lncRNAs in both sequence features. The deviation of *Xist* from our hypothesis is likely due to splicing, as only *Xist* is spliced among the 3 model lncRNAs. The *Xist*-binding proteins chosen to compute the protein-binding profile mediate not only gene silencing by *Xist*, but *Xist* splicing (Chu et al., 2015). For example, PTBP1 is essential for *Xist* function and is well known to regulate splicing (Kafasla et al., 2012). Therefore, spliced RNAs might possess an advantage over unspliced RNAs in having an *Xist*-like protein-binding profile. However, when comparing k-mer content, the top-ranked *Xist*-like RNAs were significantly more likely to be unspliced (p-value = 1.708e-05 in the 99.5th percentile).

To address the confounding effect of splicing, we will further tailor our selection of RNAs besides thresholds of expression and chromatin enrichment. We are currently measuring chromatin-specific RNA half-life (turnover), which is the time an

RNA is present in the chromatin region before either translocation to the cytoplasm or RNA degradation. We expect that filtering for RNAs with relatively long half-life would exclude many RNAs that are spliced but do not function in the chromatin region.

Another future direction is to improve the calculation of k-mer content from whole sequence to functional regions, as only parts of the model lncRNA sequences were shown to have a significant function in gene silencing (Kirk et al., 2018) . Eventually, after we narrow down the list of selected RNAs and refine the method for quantifying k-mer content, we will test the functions of candidate RNAs through biochemical experiments.

Acknowledgments

I thank my PI and direct mentor Dr. Mauro Calabrese for his mentorship and guidance throughout this project. I am grateful for the professional and personal growth I had as part of the Calabrese Lab for the past 2 years. I thank Mickey Murvin for performing the cellular fractionation and RNA immunoprecipitation assays that generated the sequencing data crucial to this study, and other lab members for their constant help and support. I thank Dr. Amy Maddox, Hazel Havens, and my BIOL692H cohort for their generous feedback. I thank the UNC community for inspiring me to enter computational biology, especially the OUR office and COMP110 taught by Professor Kris Jordan. I thank my family and friends for always standing by me while I struggle through research.

All the figures in this thesis were created or formatted in BioRender.com.

References

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5). https://doi.org/10.1038/nbt.3519

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafrenière, R. G., Xing, Y., Lawrence, J., & Willard, H. F. (1992). The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, *71*(3). https://doi.org/10.1016/0092-8674(92)90520-M

Calabrese, J. M., Starmer, J., Schertzer, M. D., Yee, D., & Magnuson, T. (2015). A survey of imprinted gene expression in mouse trophoblast stem cells. *G3: Genes, Genomes, Genetics*, *5*(5). https://doi.org/10.1534/g3.114.016238

Chu, C., Zhang, Q. C., da Rocha, S. T., Flynn, R. A., Bharadwaj, M., Calabrese, J. M., Magnuson, T., Heard, E., & Chang, H. Y. (2015). Systematic discovery of Xist RNA binding proteins. *Cell*, *161*(2). https://doi.org/10.1016/j.cell.2015.03.025

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H. C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*(7). https://doi.org/10.1371/journal.pbio.1001091

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Diamantopoulos, M. A., Tsiakanikas, P., & Scorilas, A. (2018). Non-coding RNAs: the riddle of the transcriptome and their perspectives in cancer. *Annals of Translational Medicine*, *6*(12). https://doi.org/10.21037/atm.2018.06.10

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1). https://doi.org/10.1093/bioinformatics/bts635

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414). https://doi.org/10.1038/nature11247

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1). https://doi.org/10.1093/nar/gky955

Kafasla, P., Mickleburgh, I., Llorian, M., Coelho, M., Gooding, C., Cherny, D., Joshi, A., Kotik-Kogan, O., Curry, S., Eperon, I. C., Jackson, R. J., & Smith, C. W. J. (2012). Defining the roles and interactions of PTB. *Biochemical Society Transactions*, *40*(4). https://doi.org/10.1042/BST20120044

Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J., & Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, *50*(10). https://doi.org/10.1038/s41588-018-0207-8

Latos, P. A., Pauler, F. M., Koerner, M. v., Şenergin, H. B., Hudson, Q. J., Stocsits, R. R., Allhoff, W., Stricker, S. H., Klement, R. M., Warczok, K. E., Aumayr, K., Pasierbek, P., & Barlow, D. P. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science*, *338*(6113). https://doi.org/10.1126/science.1228110

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7). https://doi.org/10.1093/bioinformatics/btt656

Mancini-DiNardo, D., Steele, S. J. S., Levorse, J. M., Ingram, R. S., & Tilghman, S. M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes and Development*, *20*(10). https://doi.org/10.1101/gad.1416906

Petazzi, P., Sandoval, J., Szczesna, K., Jorge, O. C., Roa, L., Sayols, S., Gomez, A., Huertas, D., & Esteller, M. (2013). Dysregulation of the long non-coding RNA transcriptome in a Rett syndrome mouse model. *RNA Biology*, *10*(7). https://doi.org/10.4161/rna.24286

Schertzer, M. D., Braceros, K. C. A., Starmer, J., Cherney, R. E., Lee, D. M., Salazar, G., Justice, M., Bischoff, S. R., Cowley, D. O., Ariel, P., Zylka, M. J., Dowen, J. M., Magnuson, T., & Calabrese, J. M. (2019). lncRNA-Induced Spread of Polycomb Controlled by Genome Architecture, RNA Abundance, and CpG Island DNA. *Molecular Cell*, *75*(3). https://doi.org/10.1016/j.molcel.2019.05.028

Statello, L., Guo, C. J., Chen, L. L., & Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. In *Nature Reviews Molecular Cell Biology* (Vol. 22, Issue 2). https://doi.org/10.1038/s41580-020-00315-9

Tang, J., Yu, Y., & Yang, W. (2017). Long noncoding RNA and its contribution to autism spectrum disorders. In *CNS Neuroscience and Therapeutics* (Vol. 23, Issue 8). https://doi.org/10.1111/cns.12710

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., & Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. In *Nature Reviews Genetics* (Vol. 19, Issue 9). https://doi.org/10.1038/s41576-018-0017-y

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., & Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9). https://doi.org/10.1186/gb-2008-9-9-r137

Code Availability

All the code used for this study is available at

https://github.com/zhiyuezhang7/tsc_lncRNA_project/tree/main/22Fall.