1 Germline RNA signatures underlying transposable element activity in mice

Partners: Lab. Panagiotis Alexiou, CEITEC MU and Lab. Joan Barau, IMB Mainz

1.1 SUMMARY

Transposable elements (TEs) reemerged as important controllers of genome regulation, however computationally sorting among active, inactive and regulatory copies remains challenging. Mouse germ cells effectively use a small RNA-based system to detect, classify and react to active TEs. We will attempt to identify the loci where the antisense RNA that fuels this system is produced and deployed. Machine learning applied to this data will be used to identify a genomic signature that predicts TE activity, gain insight into the origins of antisense small RNAs and how they are converted into epigenetic information. We expect our effort to be the seed of a framework allowing the consistent prediction of TE activity based on genomics data.

1.2 SCIENTIFIC BACKGROUND

Transposable elements (TEs) make up roughly half of the mouse and human genomes, far exceeding the amount of protein-coding genes[1]. Despite a historical view of TEs as parasitic genetic elements that are detrimental to genome integrity, a recent number of studies have shown that TEs also contribute to gene regulation in a tissue- and cell-specific fashion[2,3]. A framework to categorize TEs into active, inactive and regulatory may thus enhance our capacity to understand cell-specific genome regulation and its evolution.

One of the biggest challenges associated with the functional categorization of TEs lies in their abundance, diversity and repetitiveness[4]. Because sequence mappability tends to be limited, it is difficult to ascertain the activity of a TE based solely on genome-wide chromatin or RNA profiles. The germ cells of animals natively possess an outstanding system of sorting TEs into active and inactive. This pathway needs to be very accurate and efficient: mistargeting can lead to the silencing of genes, and failure to silence TEs leads to infertility due to TE reactivation during meiosis. Strikingly, this pathway relies on information encoded into small RNAs that are 21 to 35 nucleotides in length, far below the read length of current sequencing technologies. These are called piwi-interacting RNAs (piRNAs) after their association with PIWI proteins which are at the center of the piRNA pathway[5]. In mouse embryonic germ cells this pathway converts TE-derived RNA into piRNAs that are in sense and antisense orientation to a canonical TE mRNA[6]. piRNA-mediated transcriptional silencing of TEs is thought to occur when a nuclear PIWI loaded with an antisense piRNA finds a nascent TE mRNA. The stabilized match between the nascent transcript and the piRNA leads to the assembly of a repressive complex that culminates with the DNA methylation of the TE promoter[7]. PIWI loaded with antisense piRNAs and their stabilization to TE mRNA are thus essential for effective piRNA pathway specificity.

Intriguingly, the origin of TE-derived antisense RNA remains obscure. It has been proposed that antisense TE sequences are produced from transcription of so-called ‘piRNA clusters’[8].

However, these are no different from standalone TE insertions, and unlike true piRNA clusters found in meiotic cells[9], their transcriptional output is poorly characterized. Two possible reasons may account for the under-representation of these transcripts in previous data. Antisense transcripts may undergo fast cleavage and processing into piRNAs, quickly disappearing from the long RNA pool. Alternatively, these transcripts may not be polyadenylated, disappearing from RNA sequencing commonly performed on polyA-enriched fractions. In line with these hypotheses, piRNA clusters in *Drosophila* germ cells are biochemically defined by chromatin modifications and its readers, and sense and antisense RNA is produced via a non-canonical transcription and RNA export mechanism[10,11].

1.3 GOAL AND RATIONALE

Our goal is to identify molecular signatures that allow the accurate prediction of TE activity in germ cells.

Because the piRNA pathway is a biological system that very efficiently achieves this goal, we will leverage its key molecular signatures to build a dataset of *bona fide* active and inactive TEs. This will allow us to use machine learning to attempt to predict TE activity based on DNA sequences.

Aims:

I. Identify the genomic locations where antisense TE-derived RNA is produced and deployed

II. Define closely related cohorts of TEs that are targets (active) and non-targets (inactive)

III. Define the molecular signature that best predict TE activity

1.4 WORK PROGRAMME

<u>*Aim I*</u>: To identify the loci where antisense TE-derived RNA is produced, we will sequence the RNA that accumulates when PIWI-mediated processing is impaired. To expose accumulated antisense RNA we will use germ cells from double knockout (dKO) 16.5 day-old embryo mutants for the two embryonic piwi argonautes, MILI and MIWI2[12]. dKO and wild-type control mice in the reporter background *Oct4-EGFP* (B6;129S4-*Pou5f1*^{tm2Jae}/J) will be used to isolate germ cells by FACS and extract total RNA. Stranded RNA-seq libraries will be sequenced after ribosomal RNA depletion. Genomic locations where antisense piRNAs are deployed will be identified by iCLIP (individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation)[13]. For this, we will target SAFBL, a paralog of an RNA binding protein that has been recently identified in a proteomics screen for factors involved in the piRNA-guided silencing of TEs in the lab of *Partner 1 (Barau)*. Experimental work and sequencing of these two datasets will be performed by *Partner 1 (Barau)*.

<u>*Aim II*</u>: To define closely related cohorts of TEs that are targets (active) and non-targets (inactive) of the piRNA pathway we will combine the loci identified on Aim I with existing datasets on DNA methylation from MILI and MIWI2 mutant germ cells[12] and DNA and histone methylation from DNA methyltransferase (DNMT3C) mutants available from *Barau et al., 2016*[14] and unpublished data from *Partner 1 (Barau)*. Overlaps will be identified and sorted based on the type of TE sequences they contain. Cohorts of loci classified into active and inactive will be defined for every major mouse TE subfamily using their RepeatMasker and

Dfam models and manually inspected for defining features (e.g. promoter and 3' UTRs). These analyses will be performed by *Partner 2 (Alexiou)* and *Partner 1 (Barau)*.

Aim III: To define the molecular signature that best predicts TE activity we will feed the information about the sequences of the curated cohorts of defined active and inactive TEs into a machine learning pipeline. We will explore various architectures based on Convolutional and Recurrent Neural Networks[15]. We will optimize and train a model for the classification of TEs between the active and inactive classes. We will build an interpretation segment that will retrieve the most important characteristics that can be used to separate these two classes. Once a computational framework has been defined and trained, we will test its ability to identify active TEs in meiotic germ cells of piRNA pathway mutants and in mouse embryonic stem cell knockout lines of TE repressors. Data from TE reactivation in these two biological systems is available from *Partner 1 (Barau)*, and it will provide a benchmark to the trained machine learning algorithm. This computational framework and benchmarking will be performed by *Partner 2 (Alexiou)*.

1.4.1 Expected Outcomes, Synergies and Benefit to RNA Research

We expect the work outlined here to provide a model for the prediction of active and inactive TE copies in the mouse genome. This model, alongside the data generated will also contribute to advance the understanding of this small RNA pathway whose functioning is essential for fertility. Prediction of functional status is a significant problem in the field of genomics applied to the study of TEs (*Barau lab*), and the usage of machine learning algorithms (*Alexiou lab*) may enhance our ability to direct future studies on their roles as genome regulators. The growing body of work on TE-related phenomena at the interface of RNA and chromatin biology is a novel, exciting frontier on the study of genome regulation. We believe this collaboration will help set the stage to apply for joint funding focused on exploring TE-based genome regulation in contextes outside of germ cells such as embryonic and neuronal development. We will initially target the submission of a collaborative project using the Weave initiative (https://weave-research.net) via our respective national funding agencies (DFG in Germany and GACR in Czech Republic). Additional sources of funding will be also explored.

1.5 REFERENCES

1. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu. Rev. Genet.* 54, 539–561 (2020).
2. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543, 550–554 (2017).
3. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405 (2008).
4. Teissandier, A., Servant, N., Barillot, E. & Bourc'his, D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mob. DNA* 10, 52 (2019).
5. Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nat. Rev. Genet.* 20, 89–108 (2018).
6. Czech, B. & Hannon, G. J. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends Biochem. Sci.* 41, 324–337 (2016).
7. Li, Z., Tang, X. & Shen, E.-Z. How mammalian piRNAs instruct de novo DNA methylation of transposons. *Signal transduction and targeted therapy* vol. 5 190 (2020).
8. Aravin, A. A., Sachidanandam, R., Girard, A.,...Hannon, G. J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316, 744–747 (2007).
9. Li, X. Z. *et al.* An ancient transcription factor initiates the burst of piRNA production during

early meiosis in mouse testes. *Mol. Cell* 50, 67–81 (2013).
10. Andersen, P. R., Tirian, L., Vunjak, M. & Brennecke, J. A heterochromatin-dependent transcription machinery drives piRNA expression. *Nature* 549, 54–59 (2017).
11. ElMaghraby, M. F. *et al.* A Heterochromatin-Specific RNA Export Pathway Facilitates piRNA Production. *Cell* 178, 964–979.e20 (2019).
12. Vasiliauskaitė, L. *et al.* A MILI-independent piRNA biogenesis pathway empowers partial germline reprogramming. *Nat. Struct. Mol. Biol.* 24, 604–606 (2017).
13. Alexiou, P., Maragkakis, M., Mourelatos, Z. & Vourekas, A. cCLIP-Seq: Retrieval of Chimeric Reads from HITS-CLIP (CLIP-Seq) Libraries. *Methods Mol. Biol.* 1680, 87–100 (2018).
14. Barau, J. *et al.* The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* 354, 909–912 (2016).
15. Georgakilas, G. K. *et al.* Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci. *Sci. Rep.* 10, 9486 (2020).

1.6 CHRONOGRAM AND BUDGET

For the execution of this seed project we request a total of 3000,00 € (three thousand euros). This will cover costs related to mouse breeding costs, FACS core facility usage, and Genomics Core facility usage with sequencing, and data analysis. Personnel and consumable running costs will be provided by the labs. We also request a secondment fund of 500,00 € to allow travel to CEITEC or IMB for a joint meeting and experimental/data analysis setup.

11-month chronogram

Activity	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Mouse breeding *(Barau)*	X	X	X	X	X	X					
TE annotation *(Alexiou & Barau)*				X	X	X					
Cell isolation & sequencing *(Barau)*					X	X					
Data analysis *(Alexiou & Barau)*						X	X	X			
Model construction *(Alexiou)*							X	X	X	X	
Testing *(Alexiou & Barau)*									X	X	X
Follow-up funding *(Alexiou & Barau)*											X