Piwi-interacting RNA (piRNA) are a class of small non-coding RNA (sncRNA) molecules thought to mediate retrotransposon silencing. Since transposable elements (TEs) can serve as mutagenic factors that contribute to genomic instability, piRNA function is extremely important for genome defense. While the traditional view of piRNA pathways restricted them to germline tissues, increasing evidence suggests that piRNA and Piwi proteins are present in somatic cells, including those of the central nervous system. Retrotransposition in the human brain is thought to be important for increasing neuronal diversity and specialization, by introducing genomic diversity, also known as somatic mosaicism. However the role of piRNA in these processes remains to be discovered.

Various computational tools have been developed for the detection of piRNA in small-RNA sequencing data. The first generation of such tools relied on the observation of piRNA gene clusters within the genome. Practically, those tools statically determined the existence of piRNA gene clusters by calculating pairwise genomic distances between distinct alignments. The second generation of piRNA detection tools utilized various supervised machine learning algorithms for sequence-based piRNA detection. Nevertheless, all those algorithms have one critical limitation which is the result of relying on highly parallelized aligners (e.g. STAR and Bowtie). These widely-used aligners cannot fully-align reads mapped to hundreds or even thousands of loci. Yet piRNA genes may have tens of thousands of well-defined genomic loci in repetitive elements such as TEs or heterochromatin. Consequently, commonly used aligners often miss many bona fide piRNA alignments.

We developed a novel piRNA detection method and apply it to discover piRNAs in human prefrontal cortex. Our approach is based on unsupervised learning of sncRNA populations by multidimensional genomic features, including not only their sequence composition but also the number of distinct alignments, their pairwise distances, and locations relative to other genomic elements. First, we obtain non-redundant sequences by alignment score based clustering of the reads using CD-HIT-EST. Consensus sequences are then aligned to the human genome with Seqmap, an aligner designed for finding all possible loci where a read could originate from. Seqmap alignments are next used for genomic feature extraction and HDBSCAN is applied to identify feature-based clusters of reads.
Read clusters significantly enriched with known piRNA sequences are considered novel piRNA.

This work represents the first application of multidimensional clustering of all possible small RNA alignments for the characterization of genomic features informative of piRNA and the discovery of novel piRNA genes. The use of unsupervised learning provided unprecedented results, which are challenging to obtain using supervised learning approaches due to the risk of overfitting the model to currently known piRNA features. Future work should experimentally validate our piRNA predictions.