

cCLIP-Seq: Retrieval of Chimeric Reads from HITS-CLIP (CLIP-Seq) Libraries

Panagiotis Alexiou, Manolis Maragkakis, Zissimos Mourelatos, and Anastassios Vourekas

Abstract

HITS-CLIP (High-Throughput Sequencing after in vivo Crosslinking and Immunoprecipitation, CLIP-Seq) libraries contain fragments of the RNA sequences bound in vivo by an RNA binding protein (RBP). Such fragments, especially if they represent RNA duplexes bound in vivo by the RBP, can occasionally be ligated together to form *chimeric* CLIP tags. Chimeric CLIP tags from Argonaute CLIP libraries can provide the exact base pairing profiles of small RNAs with their target RNA sequences, thus solving a critical problem in the field of post-transcriptional regulation. We recently reported an analysis of chimeric reads from the *Drosophila* Piwi protein Aubergine, which revealed a novel mechanism for mRNA entrapment within germ RNP granules. We term this novel approach chimeric CLIP (cCLIP) and present here the main steps that a researcher can take after the acquisition of the deep sequencing data, for the identification of candidate chimeric reads in Piwi CLIP libraries. Extending the scope beyond small-RNA binding proteins, we believe that cCLIP can be utilized to elucidate the in vivo functions of RNA-binding proteins in general, and especially those that modulate RNA secondary structures. We, therefore, also describe aspects of the generalized chimeric read identification problem, which can find use in the analysis of the CLIP libraries of any RNA-binding protein.

Key words HITS-CLIP, CLIP-Seq, RNA-IP, Argonaute, Piwi, Next generation sequencing, Illumina, cDNA, Immunoprecipitation, Chimeric reads, RNA-binding protein, Ribonucleoprotein complexes, In vivo, Transcriptomic analysis, Base-paired RNA, miRNA target sequences, piRNA target sequences, Posttranscriptional RNA processing, Gene silencing

1 Introduction

High-Throughput Sequencing after Crosslinking and Immunoprecipitation (HITS-CLIP, CLIP-Seq) has become the leading method to probe the target specificity of RNA-binding proteins (RBPs) in general and Argonaute family proteins in particular. UV crosslinking of freshly harvested live tissue or live cells induces

Panagiotis Alexiou and Anastassios Vourekas contributed equally to this work.

the formation of covalent bonds between RNA and proteins that are in complex, and this allows the immunopurification of the cross-linked complexes under very stringent conditions [1–3]. CLIP libraries contain fragments of the RNA sequence that were directly bound by the studied protein in vivo, called CLIP tags. Recently, our group and others have discovered CLIP tags that contain fragments of RNA sequences ligated together [4–9]. These reads were called chimeric (also referred to as hybrid), and it was determined that they represent base-paired RNA sequences, bound in vivo by the RNA-binding protein that is being probed by the CLIP approach, and ligated together by virtue of their proximity while bound by and crosslinked to the RBP (Fig. 1). Chimeric reads between small RNAs and their complementary RNA targets have been discovered in Argonaute (Ago) and Piwi CLIP libraries [4–6, 8, 9], and have been analyzed to provide the small RNA (miRNA or piRNA) target sequence in vivo. This discovery reduces the need for prediction approaches for target sequence identification in an Argonaute CLIP experiment, approaches that are blind to the availability of the target sequence in vivo, and have known limitations concerning the specificity and sensitivity of target sequence detection. Similar to the chimeric reads are the hybrid reads identified by the hiCLIP method, which revealed secondary structures bound by Staufen 1 [7].

Chimeric reads can be formed by endogenous activities, in addition to being an objective of a specific step of the CLIP experimental protocol [6, 9]. Our group has adapted and optimized the HITS-CLIP protocol [2] for Piwi proteins, and has previously published a detailed description of this protocol [10]. Although we did not include a specific step to promote the generation of chimeric reads, numerous chimeras were identified in mouse (unpublished data) and *Drosophila* Piwi protein CLIP libraries [5], in agreement with previous observations [6, 9]. It was suggested that an endogenous ligase activity present in the lysates of crosslinked cells can ligate RNAs cleaved by endonuclease RNase T1 (provided exogenously) to produce 5' -OH and 2'-3' cyclic phosphate ends

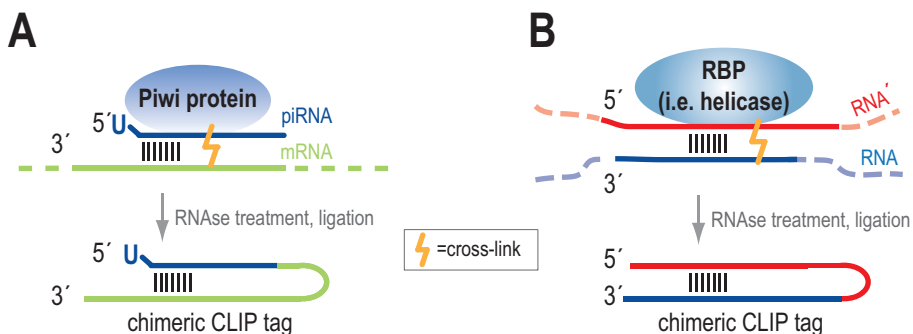


Fig. 1 Formation of chimeric reads in CLIP libraries. (a) Chimeric reads in Piwi CLIP libraries. (b) Hypothesis on how chimeric reads can be formed in CLIP libraries of any RNA binding protein that binds RNA duplexes

(these are slowly resolved into 3'-P) [6]. These ends can be substrates for the endogenous RtcB tRNA ligase activity, which joins the spliced tRNA halves after the removal of introns [11], therefore RtcB can play the role of the suggested activity that generates chimeric reads in the absence of exogenously provided ligase. We have observed the telltale sequence bias of RNase A cleavage on CLIP tags from mouse testis [12] and *Drosophila* early embryo (0–2 h) libraries. RNase A, like RNase T1, generates 5' -OH ends and 2'-3' cyclic phosphate ends, therefore both activities required for the generation of chimeric reads can be endogenously present in the cell lysate. We have verified that the chimeric reads generated by the endogenous activities accurately capture authentic, in vivo piRNA base pairing with mRNAs in the context of a Piwi-piRNA-mRNA tripartite complex [5]. It is also possible that a small degree of other nucleolytic and ligase activities can promote the generation of some chimeras. The subsequent steps of our protocol should prevent any other RNA ligase byproducts; therefore, our data are compatible with fortuitous chimeric read generation (possibly during the immunoprecipitation step) by endogenous nucleolytic and ligase activities [5]. We have noticed that such activities are apparently not always present in the cell and tissue lysates at sufficient levels to promote the formation of chimeras; therefore, the researcher should employ one of the relevant published protocols, if chimeras are an experimental objective.

We have observed the presence of chimeric reads in libraries of various RNA-binding proteins (unpublished data). These concern both *in cis* and *in trans* RNA duplex elements. We are certain that the analysis of these chimeras can reveal a trove of information on the type of structures that different RBPs recognize, and how the activities of certain RBPs can potentially modulate these structures.

This method paper aims to provide details on how to bioinformatically retrieve and analyze chimeric reads from CLIP libraries (Fig. 2), assuming that the researcher has employed one of the relevant methods that have been described adequately elsewhere. We use as example CLIP libraries of the Piwi protein Aubergine, and we try to set the framework for generalizing the chimeric read identification approach for any RNA-binding protein that binds structured RNAs.

Candidate chimeric reads are found in the pool of the unmapped CLIP tags (*see Note 1*). We filter reads in which an intact piRNA sequence can be found either at the 5' or 3' end of an unmapped read. We have observed the occurrence of chimeras containing piRNAs at both the 5' and 3' ends at comparable rates, and at least in Aubergine CLIP libraries the properties of such chimeras are similar. Therefore, we include both types of chimeras in downstream analyses. We then map the remaining fragment of the chimera against the genome, and for the purposes of the analysis of the entrapment of germ plasm localized mRNAs, we keep only chimeras that contain both a piRNA and a sequence fragment that can be uniquely aligned to an mRNA.

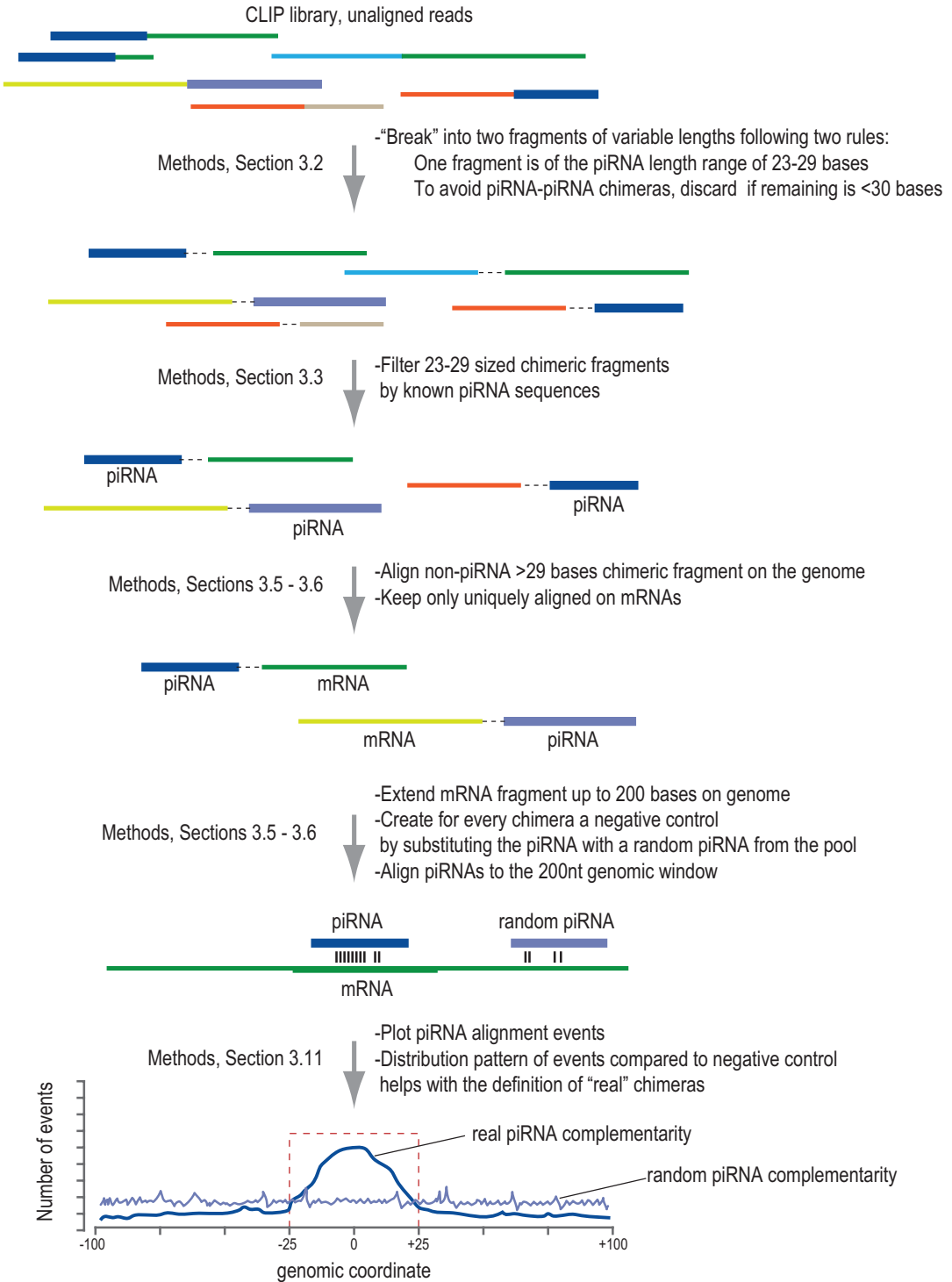


Fig. 2 Graphic outline of the retrieval of chimeric reads from a Piwi CLIP library. Throughout the text, we provide additional guidelines for applying this methodology to address the general problem of retrieving chimeric reads from a CLIP library of any RNA binding protein

Critical for the successful retrieval and subsequent analysis of chimeric reads from CLIP libraries is the approach followed for uncovering the base pairing patterns between the two RNA fragments found within the chimeric reads. We note that the use of RNA folding algorithms for determining candidate chimeras has several weaknesses, and in our opinion should be used with caution. First, the chimeric read is expected to fold back into a hairpin, but since this structure may have never existed *in vivo*, considering the thermodynamic score of this structure as a proxy for the legitimacy of the chimera can be misleading. Protein binding can modify the thermodynamic stability of an RNA duplex, and in fact many RNA-binding proteins such as RNA helicases in the presence of ATP can force the formation of RNA structures or duplexes that are not favored thermodynamically [13]. Further, due to the randomness of the nucleolytic trimming of the crosslinked RNAs, part of the base-paired sequences might be missing, or parts of the mRNA target sequence might be able to fold back on itself.

Having the above in mind, we followed a straightforward approach that reduces the assumptions about the base pairing profiles of the piRNA and mRNA fragments of chimeric reads. We employed a local alignment strategy that maps the piRNA sequence to a genomic window, extended $-/+100$ bases from the midpoint of the genomic alignment of the mRNA fragment. This allows the uncovering of the full base-paired sequence in case of missing nucleotides, and also provides a reliable estimate of the signal-to-noise ratio (here as noise we consider the random complementarity that a piRNA can show to genomic sequences neighboring the captured complementary site). Our pipeline delivers the chimeras and the complementarity profiles of the pairs of RNA fragments found in them (in this example piRNAs and mRNA fragments). We utilize the plot of the complementarity events relative to the mRNA fragment midpoint as a way to evaluate the authenticity of the complementarity events and therefore the chimeras, by comparing the distribution of these events with the negative control (random piRNA complementarity), and the events occurring by chance in sequences neighboring the mRNA fragment (Figs. 2 and 3). The researcher can use filters present in our pipeline to select any subset of the chimeras based on the complementarity score and on the position of the complementarity event relative to the mRNA fragment midpoint.

2 Required Programs and Data

2.1 Programs

1. Perl (<https://www.perl.org/>).
 - (a) GenOO [14] (<https://github.com/genoo/GenOO/>).
 - (b) Modern::Perl (<http://search.cpan.org/~chromatic/Modern-Perl-1.20161005/lib/Modern/Perl.pm>).

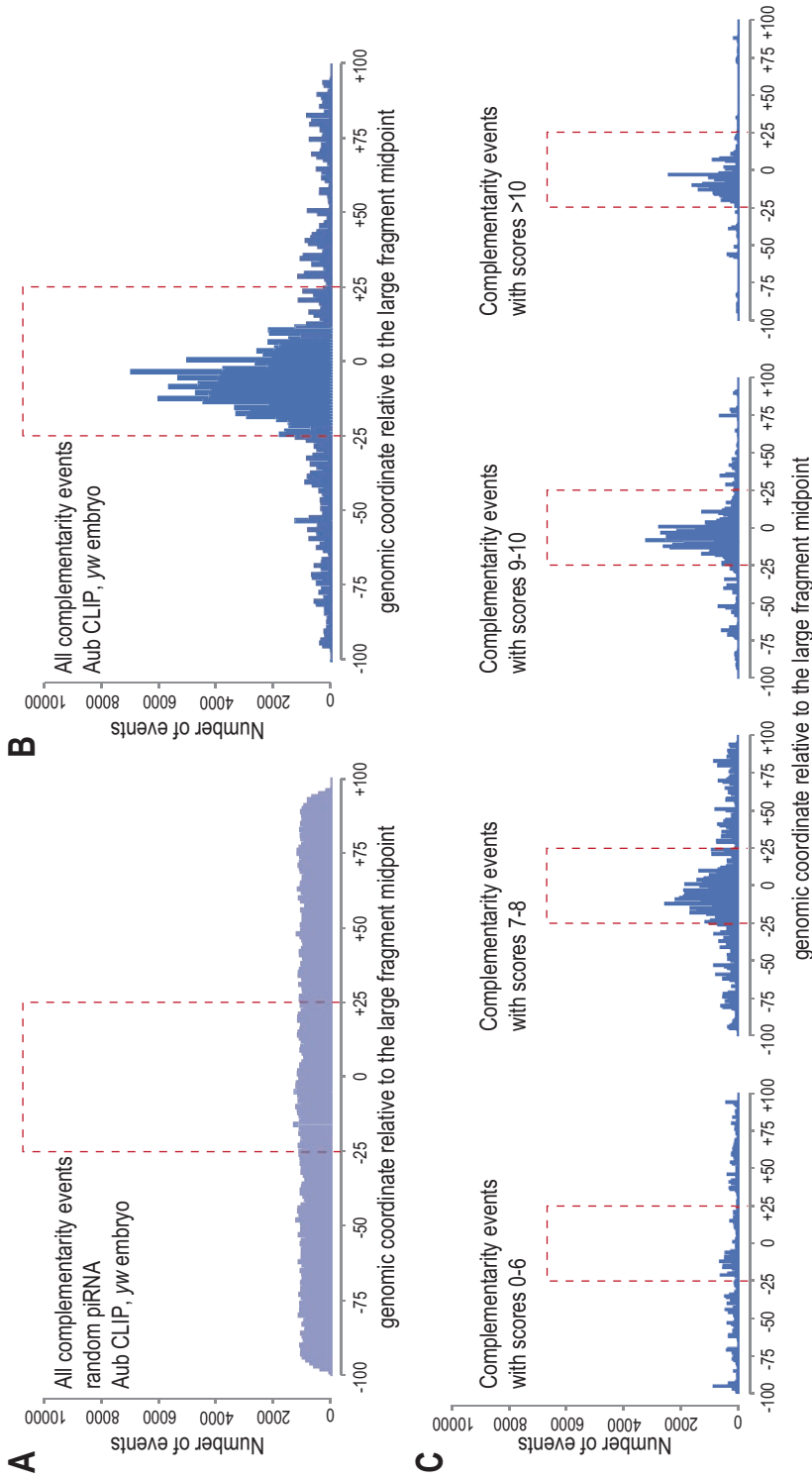


Fig. 3 piRNA-mRNA complementarity events (counts) within $-/+100$ bases from the midpoint of the mRNA fragment of chimeras found in CLIP libraries of *Drosophila* Piwi protein Aubergine. **(a)** Negative control (random piRNA) complementarity events for one Aubergine CLIP library. **(b)** All complementarity events for the same library as in **(a)**. **(c)** Complementarity events for the same library as in **(a)** and **(b)**, plotted separately for four alignment score groups. The *dashed red* line demarcates the subset of the events that occur within $-/+25$ bases from the midpoint of the mRNA fragment. Visual observation reveals that these events are substantially more abundant compared to negative control and to events occurring further away from the midpoint of the mRNA fragment, for scores >6 . Our pipeline provides the option to extract any subset of chimeric reads by the complementarity score or the relative position from the mRNA fragment

- (c) Getopt::Long (<http://search.cpan.org/~jv/Getopt-Long-2.49.1/lib/Getopt/Long.pm>).
 - (d) File::Path (<http://search.cpan.org/~riche/File-Path-2.12/lib/File/Path.pm>).
2. R (for plots only) (<https://cran.r-project.org/>).
 - (a) Optparse (<https://cran.r-project.org/web/packages/optparse/>).
 3. STAR aligner (<https://github.com/alexdobin/STAR>).
 4. Samtools (<http://www.htslib.org/>).
 5. Bedtools (<http://bedtools.readthedocs.io/>).

2.2 Data

1. Chromosome sequences for reference genome (FASTA format).
2. Gene locations (.gtf format file).

3 Methods

This is a step-by-step description of the bioinformatic pipeline used to retrieve the chimeric reads from a Piwi CLIP library. The segments of the functional bash script that can run this pipeline can be found in every section. A functional, complete bash script that can be used to run this analysis, along with all the required analysis scripts, are also deposited online (<https://github.com/palexio/CLIP-chimaeric/>).

3.1 Preprocessing and Alignment

The input CLIP reads have been prepared by trimming adaptor sequences followed by alignment to the genome using the STAR aligner (*see Note 2*).

3.2 Split Unmapped Reads

Script: split_unmapped_reads.pl

Unmapped reads are split into “small” and “large” fragments and converted into .fasta format files. The flags *--min-small 23* and *--max-small 29* limit the length of small sequence output. The flag *--min-large 30* only allows large fragments 30 nucleotides or longer to be produced. Chimeras are split at every point that allows for pairs of fragments within the given length constraints to be produced (*see Notes 3–5*). For example, a 55 nt long read would be broken in pairs of (23,32), (24,31), (25,30), (30,25), (31,24) and (32,22) nucleotides, within the constraints used here.

```
perl dev/split_unmapped_reads.pl \
  --ifile $UnmappedReadsFile.fastq \
  --min-small 23 \
  --max-small 29 \
```

```

--min-large 30 \
--out-small results/$samplename/frag-
ments.small.fa \
--out-large results/$samplename/frag-
ments.large.fa

```

3.3 Filter Small Fragments with Library of Known Small Reads (piRNAs)

Script: filter_fa_by_sam_sequence.pl

We want to retain only possible chimeric pairs that contain a real piRNA sequence. To this end, the small fragments are filtered based on the sequences found in the aligned reads of a piRNA CLIP library. Other sequence filtering steps can be performed here.

```

cat $Aligned.SmallReadsFile.bam \
| samtools-1.2 view \
-h \
- \
| perl dev/filter_fa_by_sam_sequence.pl \
--ifile results/$samplename/frag-
ments.small.fa \
--ffile - \
> results/$samplename/fragments.small.
filtered.fa

```

3.4 Filter Large Fragments with Passed Small Fragments

Script: filter_by_name.pl

After the “small fragment file” has been filtered, we remove from the “large fragment file” all large fragments that were paired with small fragments that did not pass the filter (*see Note 6*).

```

perl dev/filter_by_name.pl \
--ifile results/$samplename/fragments.
large.fa \
--ref-file results/$samplename/frag-
ments.small.filtered.fa \
> results/$samplename/fragments.large.
filtered.fa

```

3.5 Align Fragments to the Genome

Tool: STAR 2.5.2b aligner

We align the filtered large and small fragment files separately to the reference genome using the STAR aligner. The attached bash script contains the STAR settings used in this step. Neither alignment allows introns based on known genes. Again, small fragments are allowed to map in up to 50 genomic locations with equally good score.

```

### ALIGN LARGE FRAGMENTS
STAR-2.5.2b \

```



```

--genomeDir data/genome/dm3/ \
--readFilesIn results/$samplename/
fragments.large.filtered.fa \
--readFilesCommand cat \
--runThreadN 12 \
--outSAMtype SAM \
--outSAMattributes All \
--outFilterMultimapScoreRange 0 \
--alignIntronMax 1 \
--outFilterMatchNmin 8 \
--outFilterMatchNminOverLread 0.9 \
--outFileNamePrefix
results/$samplename/fragments.large.filtered.

### ALIGN SMALL FRAGMENTS
STAR-2.5.2b \
--genomeDir data/genome/dm3/ \
--readFilesIn results/$samplename/
fragments.small.filtered.fa \
--readFilesCommand cat \
--runThreadN 12 \
--outSAMtype SAM \
--outSAMattributes All \
--outFilterMultimapScoreRange 0 \
--alignIntronMax 1 \
--outFilterMatchNmin 8 \
--outFilterMatchNminOverLread 0.9 \
--outFileNamePrefix
results/$samplename/fragments.small.filtered. \
--outFilterMultimapNmax 50

```

3.6 Post-Alignment Filters

Tool: bedtools intersect

We use *bedtools intersect* and a General Transfer/Feature Format (.gtf/.gff) file of gene location to filter large fragments that map on genes. The output of this process should be two .sam files—one containing the filtered and aligned small fragments, and one containing the filtered and aligned large fragments (*see Notes 7 and 8*).

```

cat results/$samplename/fragments.large.
filtered.Aligned.out.sam \
| grep 'NH:i:1' \
| samtools-1.2 view \
-b \
-T $GenomeFile.fa \
- \
> results/$samplename/fragments.large.
filtered.Aligned.bam

```

```

bedtools intersect \
  -a results/$samplename/fragments.
large.filtered.Aligned.bam \
  -b $GenesFile.gtf \
  -split \
  -u \
  | samtools-1.2 view \
  - \
  > results/$samplename/fragments.large.
filtered.Aligned.OnGenes.sam

```

3.7 Keep Unique Pair *Script: filter_best_pairs.pl*

At this step, we still possibly have multiple pairs of fragments per original chimeric read. Considering the genomic alignment of both the short and long fragments, we choose the pair of fragments with the highest sum of matching nucleotides. If there are multiple pairs with the same total alignment score, a random pair is selected.

```

perl dev/filter_best_pairs.pl \
  --infile1 results/$samplename/fragments.
large.filtered.Aligned.OnGenes.sam \
  --infile2 results/$samplename/fragments.
small.filtered.Aligned.out.sam \
  --ofile1 results/$samplename/fragments.
large.filtered.Aligned.OnGenes.paired.sam \
  --ofile2 results/$samplename/fragments.
small.filtered.Aligned.paired.sam

```

3.8 Get Fragment Sequences

Script: sam-to-fasta.pl

We extend the large reads to a 200-nucleotide (nt) window using the `--out-length 200` flag and convert to .fasta file format. We also convert small reads to .fasta file format without extending their length (*see Note 9*).

```

perl dev/sam-to-fasta.pl \
  --sam results/$samplename/fragments.
large.filtered.Aligned.OnGenes.paired.sam \
  --chr_dir data/$species/chromosomes/ \
  --out-length 200 \
  > results/$samplename/fragments.large.
filtered.Aligned.OnGenes.paired.200nt.fa

perl dev/sam-to-fasta.pl \
  --sam results/$samplename/fragments.
small.filtered.Aligned.paired.sam \
  --chr_dir data/$species/chromosomes/ \
  --exact \
  > results/$samplename/fragments.small.
filtered.Aligned.paired.fa

```

3.9 Make Pairs Table

Script: make_pairs_table.pl

This step combines the two .fasta sequence files produced in the previous step to create a tab delimited pairs table. We run the same command a second time with the *--shuffle* option to make the same number of random pairs (*see Note 10*).

```
perl dev/make_pairs_table.pl \
  --ifile1 results/$samplename/fragments.
small.filtered.Aligned.paired.fa \
  --ifile2 results/$samplename/fragments.
large.filtered.Aligned.OnGenes.paired.200nt.fa \
  > results/$samplename/pairs.tab

perl dev/make_pairs_table.pl \
  --ifile1 results/$samplename/fragments.
small.filtered.Aligned.paired.fa \
  --ifile2 results/$samplename/fragments.
large.filtered.Aligned.OnGenes.paired.200nt.fa \
  --shuffle \
  > results/$samplename/pairs.shuffled.tab
```

3.10 Align

Script: align.pl

We align the small fragments to the extended 200-nt fragments (for both real and random pairs). We use a slightly modified Smith–Waterman [15] alignment method (weights: match = +1, mismatch = -1, gap = -2). Differences of our alignment versus Smith–Waterman: (a) No penalties are given to non-matching nucleotides on the edges of the alignment. (b) If there are multiple optimal alignment scores, one is picked randomly. (c) Alignments in which part of one sequence is outside the boundaries of the other sequence are not considered. The fields of the output alignment file are as follows: “chimeric read name”, “sequence small fragment”, “sequence large fragment (200 nt)”, “length small fragment”, “length large fragment”, “alignment score”, “midpoint location on large fragment”, “binding vector large fragment”, “binding vector small fragment reverse complement”, and “reverse complement sequence small fragment” (*see Notes 11 and 12*).

```
perl dev/align.pl \
  --ifile results/$samplename/pairs.tab \
  > results/$samplename/pairs.aligned.tab

perl dev/align.pl \
  --ifile results/$samplename/pairs.shuf-
fled.tab \
  > results/$samplename/pairs.shuffled.
aligned.tab
```

3.11 Plots

Script: plots.R

Using the alignment files as input (real and random alignments) this R script produces several plots that help understand the alignment and therefore provide valuable insight into the attributes of the complementarity between the piRNAs and their target sequences. Plot-1 is a scatter plot of real versus random alignment scores. Plot-2 is a histogram of real and random alignment scores. Plot-3 is a histogram of the distribution of real and random alignment midpoint locations on the 200-nt extended large fragment.

```
Rscript dev/plots.R \
  --ifile results/$samplename/pairs.
aligned.tab \
  --sfile results/$samplename/pairs.shuf-
fled.aligned.tab \
  --ofile results/$samplename/pairs.
aligned.pdf
```

Script: filter_alignment.pl

Allows user defined filtering of the alignment file. Flags --min-score, --max-score, and --min-midpoint, --max-midpoint, give the constraints for output alignments. These settings allow the user to filter a subset of the chimeras based on their alignment characteristics for further analysis downstream.

4 Notes

1. Gapped read alignment can interfere with the retrieval of the candidate chimeras, especially for CLIP libraries in which local secondary structures are expected to be captured by chimeric reads. In this case, the researcher should opt for non-gapped alignment of the CLIP libraries, and also develop a rationale for separating chimeric reads from standard CLIP tags that may have small deletions in their sequence.
2. Using the flag `-outReadsUnmapped` Fastx will output unmapped reads into a separate file. Using the flag `-outFilterMultimapNmax 50` allows up to 50 equal score mappings per read instead of the default ten. This is recommended when working with repetitive sequences such as piRNAs. Any aligner that produces a standard SAM/BAM file can be used at this step.
3. In order to avoid piRNA:piRNA chimeras, the remaining sequence after generating the piRNA-sized small fragment is required to be >29-nt long.
4. Pairs with the small fragment both on the 5' end and the 3' end of the chimeric read can be output.
5. Extending the method to any size is a matter of tweaking the `-min-small`, `--max-small`, `--min-large`, `--max-large` flags to the

desired lengths. If these flags are left undefined then pairs of fragments of all potential sizes will be produced. The researcher should be aware that many sequence fragments with a size of ≤ 20 bases cannot be unambiguously aligned to the genome. This ambiguity increases, as the size gets smaller. We chose to process only chimeric reads that contain sequence fragments that are unambiguously identified within a piRNA library or within mRNA exons. This requirement reduces uncertainty in downstream analyses such as complementarity between the two chimeric read fragments.

6. This is an optional step, but as the next step will be a resource consuming alignment to the genome, it is more efficient to filter unwanted reads before aligning rather than after.
7. Using the `-split` option we take into consideration only the exonic regions of genes.
8. At this point, any additional filters can be applied to the aligned files.
9. The `-chr_dir` flag denotes a folder containing a file per chromosome named `$ChrName.fa`.
10. The random pairs consist of a random small sequence (piRNA) of the same length as the original small fragment of chimeric read, paired with the same long fragment sequence. In this way, we avoid biases of sequence length distribution in the alignment scores.
11. The binding vector shows “1” for binding nucleotides within the binding area, “0” for non-binding nucleotides within the binding area, “g” for nucleotides that have a gap across from them in the alignment (insertions), “s” for soft clipped non-matching nucleotides (outside the binding area and not matching in the extension), “m” for soft clipped matching nucleotides (outside the binding area and matching in the extension). The binding area is the area in which binding nucleotides count toward the final alignment score—the area of local alignment.
12. Areas of the small fragment outside the core alignment area that spans from the beginning of the traceback cell to the first encountered cell of score zero are expanded in an ungapped alignment and do not count toward the alignment score or midpoint. In other words, soft-clipped nucleotides are not penalized and do not affect the alignment score but are reported.

Acknowledgments

We thank members of the Mourelatos laboratory for discussions. This research was supported by NIH grant GM072777 and a grant from ALS Therapy Alliance to Z.M.

References

1. Licatalosi DD et al (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456:464–469
2. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460:479–486
3. Ule J et al (2003) CLIP identifies novel regulated RNA networks in the brain. *Science* 302(5648):1212–1215
4. Kameswaran V et al (2013) Epigenetic regulation of the DLK1-MEG3 MicroRNA cluster in human type 2 diabetic islets. *Cell Metab* 19:135–145
5. Vourekas A, Alexiou P, Vrettos N, Maragkakis M, Mourelatos Z (2016) Sequence-dependent but not sequence-specific piRNA adhesion traps mRNAs to the germ plasm. *Nature* 531:390–394
6. Grosswendt S et al (2014) Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol Cell* 54:1042–1054
7. Sugimoto Y et al (2015) {hiCLIP} reveals the in vivo atlas of {mRNA} secondary structures recognized by Staufen 1. *Nature* 519:491–494
8. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153:654–665
9. Moore MJ et al (2015) miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nat Commun* 6:8864
10. Vourekas A, Mourelatos Z (2014) HITS-CLIP (CLIP-Seq) for mouse piwi proteins. *Methods Mol Biol* 1093:73–95
11. Popow J et al (2011) HSPC117 is the essential subunit of a human tRNA splicing ligase complex. *Science* 331:760–764
12. Vourekas A et al (2015) The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. *Genes Dev* 29:617–629
13. Yang Q, Fairman ME, Jankowsky E (2007) DEAD-box-protein-assisted RNA structure conversion towards and against thermodynamic equilibrium values. *J Mol Biol* 368:1087–1100
14. Maragkakis M, Alexiou P, Mourelatos Z (2015) GenOO: a modern perl framework for high throughput sequencing analysis. *bioRxiv*. doi:[10.1101/019265](https://doi.org/10.1101/019265)
15. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197