Micro DNA

Audrey Weaver

May 8, 2025

1 Introduction

MicroDNA refers to small extrachromosomal circular DNA elements. These circular chunks of DNA typically range from 100-1000 base pairs in length. They have been detected in in eukariotic cells and are thought to be the result of DNA replication, transcriptional variations, or other repair processes.

In this project, we used an algorithm to detect the circular fragments of microDNA. Soft clips were identified and isolated from the BAM file. These are sequences which do not fully align with the sequence of the chromosome, indicating that they may be circular junctions. Reads were then filtered by length and grouped by position and sequence to detect potential junctions. Reliant on the identification of repeated sequences and soft clips from a BAM file, the algorithm then used smith-waterman to align the potentially overlapping reads on either side of the circle and detect the microhomology between them.

A confidence score was then calculated to weigh all factors contributing to the likelihood that the sequence is in fact a circle. Similar to the score given by the SequenceMatcher package, we used Smith-Waterman which had already beem developed for an assignment to match all of the microhomology sequences and give them a score. We were inspired by the scores given by SequenceMatcher and wanted to use our own algorithm while still being able to have a comparable score for each read. Mainly focused on sequence similarity and their junction proximity within the genome, our goal was to prioritize biologically realistic microDNA circles and validate them.

2 Results

A total of 23,640 soft-clipped reads were extracted from the aligned BAM file using a minimum clip length of 12 base pairs. These reads were split into start- and end-clipped groups and compared to identify potential microDNA circles. Junction pairs were required to be on the same chromosome, within 1,000 base pairs of each other, and have matching or highly similar clipped sequences as determined by local sequence alignment. While all data in the BAM file studied was originating from the same chromosome, this is something that must be considered for future use of this algorithm so we implemented considerations for the chromosome the sequences were found on.

After scoring each candidate using a confidence metric that weights microhomology against circle length, 948 distinct candidate circles were identified. We applied additional validation criteria requiring at least one supporting read on both the start and end of the circle, at least two total reads, and a minimum confidence score of 0.3. This reduced the candidate set to a smaller group of high-confidence circles, summarized in Table 1.

Most validated circles were short, with distances under 200 base pairs and perfect or near-perfect sequence alignment between junction reads. One outlier (chr1:17,912,842–17,912,938) had 63 total supporting reads and a repeated clipped sequence (TTTTTAAACATC), suggesting possible over-amplification or origin from a low-complexity region.

Despite the unusually high number of supporting reads and the repetitive nature of its clipped sequence, this candidate passed all validation criteria. Specifically, it had soft-clipped reads on both ends of the predicted circle junction, each with the same clipped sequence (TTTTTAAACATC). It also had a total

Table 1: Top 5 Validated microDNA Circles (Escaped for LaTeX)

Chromosome	Start	End	Start Seq	End Seq	Start Reads	End Reads	Total
NC_000001.10	17912842	17912938	TTTTTAAACATC	TTTTTAAACATC	23	40	63
NC_000001.10	121485155	121485216	GAATATCCACTT	TGAATATCCACT	3	1	4
NC_000001.10	121485177	121485188	TGGAATTTGCAA	TGGAATTTGCAA	2	1	3
NC_000001.10	121485185	121485358	GTTTGTAAAGTC	TTTGTAAAGTCT	2	1	3
NC_000001.10	121485356	121485378	TTTTGTGGAATT	TTTTTGTGGAAN	2	1	3

support count of 63 reads with 23 at the start, 40 at the end. The clipped sequences showed perfect microhomology, and the base pair distance between junctions was only 96 bp, resulting in a high confidence score. Strong read support, matching sequences, and short circle length were all features indicating that the signal is consistent and not an result of poor alignment or random clipping.

3 Methods

To identify potential microDNA circles, we developed a Python pipeline that detects and scores soft-clipped reads from aligned sequencing data. The input is a coordinate-sorted BAM file with its index. Soft-clipped reads were extracted using pysam by checking for soft clipping at either the 5' or 3' end of the read (CIGAR notation S). Reads were filtered based on a minimum soft clip length, which we set at 12 bp for most of the analysis to avoid short, possibly noisy clips.

Each clipped read was recorded along with its position, CIGAR string, and clipped sequence. Reads were separated into "start-clipped" and "end-clipped" groups based on the side of the alignment where the clip occurred. To identify candidate circles, we paired clipped reads on the chromosome that were within 1,000 bp of each other and had similar clipped sequences. This was chosen due to the assumption that most microDNA circles to not excede 1,000 base pairs in length.

To measure sequence similarity between clipped reads, we implemented a Smith-Waterman local alignment algorithm. The alignment score was normalized and used to calculate microhomology between each start-end pair. We then applied a confidence score that weighted sequence similarity against the genomic distance between positions. Candidates with higher similarity and shorter distances received higher confidence scores.

$$\text{confidence} = \text{microhomology} \times \left(1 - \frac{\text{distance}}{\text{MAX\_DISTANCE}}\right) \tag{1}$$

The final output includes all paired junctions with their positions, clipped sequences, distance, microhomology score, and calculated confidence. For validation, we counted the number of supporting clipped reads for each junction and considered circles with at least one read on both ends and total support of two or more to be valid. Confidence scores and clipped sequence quality were used to further analyze results.

All scripts were written to include potential variability in inputs and outputs according to what was changed throughout development of this algorithm. Minimum length of clip and possibility for validating multiple start positions at once allowed for trial with the BAM file and the algorithm and were kept for potential future users.

3.1 Reproducibility

To replicate these experiments, clone the repository and then run the following commands from the root directory of the repository.

```
$ git clone https://github.com/audreyw04/microDNA.git
$ python3 detecting.py --bam ./data/SRR413984.sorted.NC_000001.10.bam \
                       --min_clip 12 \
```

```
                        --out clips_12bp.tsv

$ python3 validation.py --circles microdna_circles.tsv \
                        --reads clips_12bp.tsv \
                        --start_positions 121485177 121485185 121373488 17912842 121485155
                        121485356 121485363 121485210 121484883 121484600 121485000
                        121484512
```

3.2 References

References

[1] OpenAI. (2024). *ChatGPT* (April 4 version) [Large language model]. https://chat.openai.com/

ChatGPT was used to assist with refining ideas, debugging code and editing for this assignment.