Discovering and Characterizing MicroDNA in a Cancer Cell Line

Vincent Bowen

University of Colorado Boulder

CSCI 4444: Computational Genomics

Dr. Ryan Layer

May 7, 2025

1 Introduction

Circular DNA, particularly microDNA, have received significant academic attention for its potential role in genomic instability, cancer progression, and gene regulation. MicroDNA are short, circular DNA molecules that originate from the nuclear genome and have been detected in a variety of cell types, including cancerous tissues. Despite their prevalence, their formation mechanisms and biological functions remain poorly understood.

This analysis presents a computational pipeline designed to detect and characterize microDNA using high-throughput sequencing data. The approach involves identifying soft-clipped reads from a BAM file, aligning these clipped sequences to one another and to a reference genome, and computing a scoring metric to evaluate the likelihood of circularization. Additionally, the performance of various pairwise sequence alignment algorithms was empirically compared to identify the most suitable tool for this task. MicroDNA events were manually verified, to ensure the accuracy of the report. This work aims to streamline the discovery of microDNA and provide a reproducible framework for further experimental testing with alternative genomes.

2 Results

A list of potential circles (microDNA) are written to a user-specified output file. If the output file is designated with a `.csv` or `.tsv` extension, the report includes the Start Position, End Position, Soft-Clip Alignment Score, Starting Soft-Clip Reference Alignment Score, Ending Soft-Clip Reference Alignment Score, and a computed Evidence Score for each candidate. Alternatively, if the user specifies a `.txt` output, the report is expanded to include the full alignment strings: the alignment between the start and end soft-clips, the alignment between the start soft-clip and the reference genome, and the alignment between the end soft-clip and the reference genome. The reports are sorted by the Evidence Score, with the highest being first, and the lowest being last.

979 candidate circles were generated, with 139 having strong supporting evidence (a positive Evidence Score) using the provided BAM and FASTA files (See the public data directory in Google Drive). The other potential circles were included to allow the user to validate circles manually using IGV if desired. A sample of both file classes can be seen in Figures 1 and 2

```
COMPARING SOFT CLIPS
Start: 2383556 End: 2383739
Score: 57.499999999999986
target            0 ----------GCCTGGCAGGTAGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42
                  0 ----------|||||||||||||||||||||||||||||||||||---------- 52
query             0 CCCCGGCCCTGCCTGGCAGGTAGCAGCCCCCACTGCATTGCT---------- 42

COMPARING START WITH REFERENCE
Score: 49.0
target            0 TAGGGGCAGCGGGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42
                  0 ....|||||...|||||||||||||||||||||||||||||| 42
query             0 GCCTGGCAGGTAGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42

COMPARING END WITH REFERENCE
Score: 34.0
target            0 CCCCGGCCCTGCCTGGCAGGTAGCAGCCCCGTGGAAGTATTT 42
                  0 |||||||||||||||||||||||||||||||.......|...| 42
query             0 CCCCGGCCCTGCCTGGCAGGTAGCAGCCCCCACTGCATTGCT 42

Evidence Score: 621.5095931892079
--------------------------------------------------------------------------------

COMPARING SOFT CLIPS
Start: 2383556 End: 2383738
Score: 54.899999999999984
target            0 ----------GCCTGGCAGGTAGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42
                  0 ----------|||||||||||||||||||||||||||||||||||---------- 53
query             0 CCCCCGGCCCTGCCTGGCAGGTAGCAGCCCCCACTGCATTGC---------- 42

COMPARING START WITH REFERENCE
Score: 49.0
target            0 TAGGGGCAGCGGGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42
                  0 ....|||||...|||||||||||||||||||||||||||||| 42
query             0 GCCTGGCAGGTAGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42

COMPARING END WITH REFERENCE
Score: 34.0
target            0 CCCCCGGCCCTGCCTGGCAGGTAGCAGCCCCGTGGAAGTATT 42
                  0 |||||||||||||||||||||||||||||||.......|... 42
query             0 CCCCCGGCCCTGCCTGGCAGGTAGCAGCCCCCACTGCATTGC 42

Evidence Score: 600.3372681281617
--------------------------------------------------------------------------------

COMPARING SOFT CLIPS
Start: 2383556 End: 2383737
Score: 50.09999999999999
target            0 ----------GCCTGGCAGGTAGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42
                  0 ----------|||||||||||||||||||||||||||||||--|-||-------- 54
query             0 TCCCCCGGCCCTGCCTGGCAGGTAGCAGCCCCCACTGCA--G-TG--------- 42

COMPARING START WITH REFERENCE
Score: 49.0
target            0 TAGGGGCAGCGGGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42
                  0 ....|||||...|||||||||||||||||||||||||||||| 42
query             0 GCCTGGCAGGTAGCAGCCCCCACTGCATTGCTGAGCCTGGAA 42
```

Figure 1: Sample of Text File Report

	A	B	C	D	E	F	G
1	start_pos	end_pos	read_read_sc	start_ref_sco	end_ref_scor	evidence_score	
2	2383556	2383739	57.49999999	49	34	621.5095931892079	
3	2383556	2383738	54.89999999	49	34	600.3372681281617	
4	2383556	2383737	50.09999999	49	34	561.7303832908004	
5	19583699	19583705	54.9	74	74	378.793343	
6	179606025	179606031	54.9	74	54	358.1843474676696	
7	179606025	179606032	52.3	74	49	344.03034343486644	
8	22182896	22183193	64.40000000	-91	-46	195.26806857050357	
9	35928488	35928731	63.90000000	-126	-46	159.81265980071228	
10	35928488	35928729	58.69999999	-126	-46	120.53453611957478	
11	35928488	35928728	56.09999999	-126	-46	100.76923076923067	
12	35928488	35928727	53.49999999	-126	-46	80.92290648105767	
13	35928488	35928725	52.29999999	-126	-46	71.93792453286264	
14	35928488	35928726	50.89999999	-126	-46	60.997973168612944	
15	121485044	121485382	68.60000000	-126	-126	30.350247273220017	
16	121485044	121485384	66.5	-126	-126	18.79194630872483	
17	121484866	121485202	66.30000000	-126	-126	18.052516411378615	
18	121484090	121484424	66.00000000	-126	-126	16.564290151149304	

Figure 2: Sample of Comma Separated Value Report

Using the `.txt` file containing the actual alignments, true circularization events could be manually verified. These potential circles were additionally hand validated in IGV by loading the FASTA and BAM file, then analyzing the sequence at the start position, and comparing that to the sequence at the end position.

3 Methods

3.1 Performance of Pairwise Sequence Aligners

After quickly observing substantial performance bottlenecks with the developed Smith-Waterman algorithm, the performance of various sequence aligners was tested. Four Pairwise Sequence Aligners were compared, with their performance recorded in Figure 3. The query size was set to 42 base pairs long, to simulate alignment on BAM reads with increasing target lengths. The empirical time to run each algorithm was recorded.

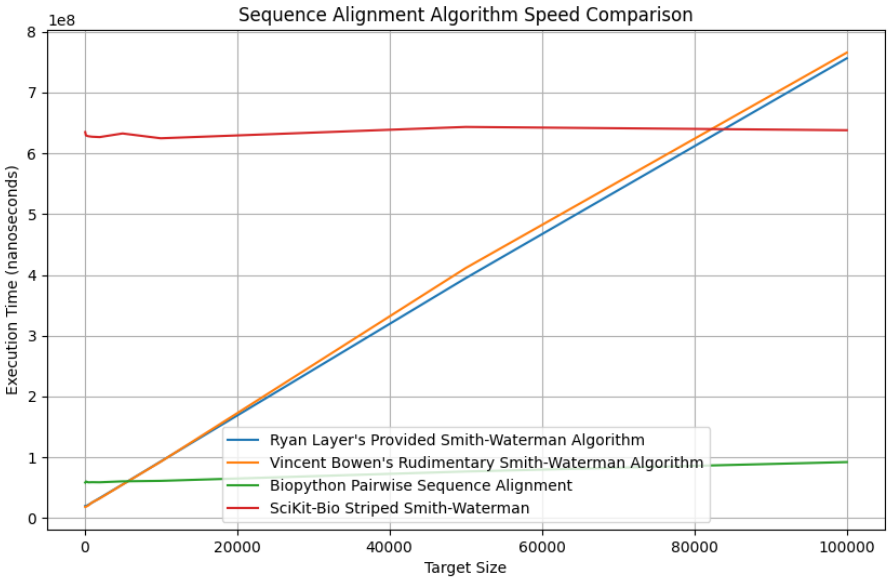

Figure 3: Performance of Various Pairwise Aligner Algorithms

Four algorithms were tested:

1. Ryan Layer's Smith-Waterman Algorithm: Basic Python implementation using a fill matrix and traceback.
2. Vincent Bowen's Smith-Waterman Algorithm: Basic Python implementation, almost identical to the algorithm developed by Ryan.
3. Sci-Kit Bio's Striped Smith-Waterman Algorithm: Python library performing a striped (banded) Smith Waterman Alignment.
4. Biopython's Pairwise Aligner: Highly configurable Python library, automatically selecting between: Needleman-Wunsch, Smith-Waterman, Gotoh (three-state), and Waterman-Smith-Beyer global and local pairwise alignment algorithms, and the Fast Optimal Global Alignment Algorithm (FOGSAA) based on user-inputted scoring schema. For this experiment, the Gotoh algorithm was used.

The algorithms developed by Vincent and Ryan exhibited nearly identical performance, which can be attributed to the fact that Vincent's implementation was derived from Ryan's methodology and reference materials. Both approaches demonstrated suboptimal performance when applied to larger targets due to a steep linear increase with respect to target length. The SciKit-Bio alignment algorithm showed unexpectedly poor performance on shorter target sequences—a notable limitation given that the majority of experimental targets were approximately 42 base pairs in length. In contrast, Biopython's Pairwise Aligner exhibited marginally lower performance on

extremely short targets but rapidly outperformed the other algorithms as target size increased. Notably, the linear growth rate of Biopython's computational cost was significantly smaller compared to the Vincent and Ryan's algorithm.

Due to the superior performance of Biopython's implementation and its extensive configurability, it was selected as the alignment tool. A key advantage lies in the fine-grained control over alignment parameters, including direct specification of the alignment algorithm itself and the scoring schema. Parameters such as: mode, match score, mismatch score, gap opening score, and gap extension score can be individually adjusted. Furthermore, for soft-clipped sequences, additional control is available through the left and right query gap opening scores, as well as the left and right query gap extension scores. This level of specificity was preferred to the more limited configuration options—restricted to basic: gap, mismatch, and match penalties in the algorithms developed by Vincent and Ryan.

3.2 Detection of MicroDNA

To detect microDNA, an algorithm was developed to: identify circular DNA sequences, aggregate candidate findings, and compute a quantitative metric reflecting the strength of evidence for true circularization.

The steps of the algorithm are as follows:

1. Load and process the BAM file to extract aligned sequencing reads and the FASTA file to serve as the reference genome for downstream analysis.

2. Identify soft-clipped regions at the start and end of reads by analyzing the CIGAR strings for the presence of matching segments adjacent to soft-clips.

3. Aggregate start and end soft-clips based on their genomic positions, recording the number of reads exhibiting these characteristics at each site.

4. Apply an empirically defined threshold to determine high-support sites. In this study, a minimum of 10 supporting reads at a given position was required for further analysis. All other sites were filtered out.

5. Perform an brute-force pairwise alignment between all identified starting and ending soft-clips to evaluate potential circular junctions.

   Here the alignment was performed using the Gotoh algorithm with the following parameters: a match score of 2, a mismatch penalty of -2, a gap opening penalty of -1, and a gap extension penalty of -0.5, configured in global alignment mode. Additionally, the aligner was configured with minimal penalties for gap opening and extension at the ends of the query: the left and right query gap opening scores, and the left and right gap extension scores were all set to -0.1. This schema prefers shifting the position of the sequences versus a mismatch, a desirable property when identifying regions of central homology, and matching junction tags.

6. Retain candidate pairs with an alignment score of at least 50 and a minimum separation of 5 base pairs between the clipped regions to ensure sequence homogeneity between the junction tags. These candidate sequences are then aligned against the reference genome.

   Here the alignment was performed using the Gotoh algorithm with the following parameters: a match score of 2, a mismatch penalty of -3, a gap opening penalty of -25, and a gap extension penalty of -6, configured in global alignment mode. This schema strongly penalized both gap opening and extension, encouraging the aligner to favor mismatches instead. This behavior is desirable when aligning soft-clipped sequences, which are expected to differ from the reference genome at the soft-clipped regions, but should not be shifted through gap creation. This preserves the positional integrity of the sequences.

7. Compute an evidence-based metric to quantify the likelihood that a given candidate represents true mi-

croDNA (circle) formation. This score was defined as follows.

$$Score = \frac{(SCAS * 8 + SSCRAS * 2 + ESCRAS * 2)}{Penalty}$$

Where,

$$Penalty = ((span - 200)/200) * *2 + 1$$

Additionally, $SCAS$ is the Soft-Clip Alignment Score, the score between the start and end soft-clips, $SSCRAS$ is the Starting Soft-Clip Reference Alignment Score, the score between the start soft-clip and the reference genome, and $ESCRAS$ is the Ending Soft-Clip Reference Alignment Score, the score between the ending soft-clip and the reference genome. Arbitrary but empirically motivated weights were applied to these scores to emphasize the alignment between the soft-clips themselves ($SCAS$), as this was observed to be a strong indicator of circularization.

A *Penalty* term was also introduced, defined as a quadratic function that rewards a homologous span close to 200 base pairs. This design penalizes extremely short or excessively long regions of homology, which are unlikely to represent true circular DNA junctions based on empirical observations.

8. Sort the candidates based on the defined metric, and write to a user-defined .txt, .csv, or .tsv file.

3.3 Reproducibility

To replicate these experiments, clone the repository and then run the experiments and analysis as follows:

The data files must be downloaded from the public data directory because they are too large for GitHub. Then move them to the `data` folder.

```
$ git@github.com:vincedbowen/micro-dna.git
$ cd micro-dna
```

1. Pairwise Aligner Performance Test

```
$ cd pairwise-aligner-comparer
$ python compare_pwa.py
```

2. MicroDNA Analysis

```
$ cd micro-dna-finder/cli
$ python3 main.py \
  -bf ../data/SRR413984.sorted.NC_000001.10.bam \
  -ff ../data/GCF_000001405.13_GRCh37_genomic.NC_000001.10.fna \
  -of ../results/results.<file-extension>
```

Where file-extension is `txt`, `tsv`, or `csv`.