MicroDNA

Drew Aparicio

May 7, 2025

1 Introduction

MicroDNA are small, circular, non-coding DNA molecules whose role in cellular life is not fully understood. Due to the use of microDNA as biomarkers for cancer, the detection of these circular microDNA molecules is crucial in understanding how they relate to the biology of cancer and how they may be used in cancer treatment. This experiment searches a BAM alignment file (binary compressed SAM file) containing many 42-bp reads from the reference genome and finds possible circular microDNAs throughout this genome, outputing a "score" of how likely the circle exists. By utilizing CIGAR strings that signify junction tags, the Smith-Waterman alignment algorithm, and the IGV software, 26 possible microDNA molecules were discovered. Using an program such as this may prove beneficial in the quick detection of microDNAs by researchers in the future.

2 Results

In total, there were 26 likely circles (seen below) found in the DNA sequence that met the set criteria, 13 of which had a score of 80 or above, representing a very high probability. All 26 circles had a length between 150 and 400 bp, which is consistent with the typical range from the literature of 100-400 bp. There was a significant number of scores in the 60s, representing relatively low confidence that a circular microDNA does exist there. However, these scores are still high enough to report as at least likely. The most probable reason for these low scores is due to errors in reads that caused the junction tags to not align perfectly with the start and end of the circle. Additionally, some of these likely circles did not show evidence of a microhomology, which significantly reduced their overall score, as the existence of a microhomology is thought to be the reason microDNAs appear. Taking this into consideration, the circles that lacked microhomologies still showed enough evidence, through the alignment of their junction tags with the ends of the sequence, of a microDNA molecule.

POSITION	LENGTH	SCORE
956749-956934	185 bp	87.91
2066285-2066645	360 bp	84.32
2403270-2403652	382 bp	68.17
2403298-2403643	345 bp	85.32
11458983-11459179	196 bp	62.99
17998305-17998648	343 bp	95.25
18062885-18063265	380 bp	66.88
21698974-21699333	359 bp	93.31
26053913-26054114	201 bp	75.80
27478693-27479037	344 bp	82.95
31775822-31776185	363 bp	70.98
35928489-35928760	271 bp	86.55
42105303-42105483	180 bp	88.56
53792372-53792739	367 bp	80.94
116247456-116247806	350 bp	86.98
121485089-121485434	345 bp	68.34
144058864-144059048	184 bp	78.50

```
161940384-161940765       381 bp       74.36
167461877-167462071       194 bp       86.83
181943749-181944106       357 bp       76.56
201093850-201094038       188 bp       89.35
206521552-206521736       184 bp       78.20
230653575-230653895       320 bp       65.57
230653581-230653895       314 bp       73.78
232062795-232063012       217 bp       62.91
236595941-236596163       222 bp       84.10

Number of circles: 26
```

3 Methods

3.1 Finding Circles

To find the circles, the algorithm first searches the BAM file for reads whose CIGAR string contains a soft-clip region then a matching region (start junction tag) or a matching region then a soft-clip region (end junction tag). These reads are then compiled into verified junctions according to three criteria. Specifically, there must be at least 50 reads that contain the soft-clipped region but no more than 500, and the soft-clipped region must be at least 4 bp in length. The length criteria are in place to ensure later alignment does not incorrectly assume short alignments are perfect evidence of a circle. In other words, if more than four bases from the junction tag agree with the opposite end of the sequence, then it is likely evidence of a circle rather than the result of simple chance. From here, the junctions are labeled as either a start junction or end junction. The algorithm parses the junction list for each start junction ("+") and searches the nearby (within 1000 bp) end junctions ("-"). Next, each pair is passed to the Smith-Waterman alignment algorithm to find whether there is a microhomology and to give a score of how well the tags align to the sequence ends. These alignment scores and the number of tags for the pair of junctions, along with the possible microDNA's start and end positions, are added to a preliminary list of circles.

3.2 Scoring

This list of circles is then re-scored using the following weights: 80% for alignment score and 20% for the number of junction tags. The junction tag score takes each circle's average number of junction tags and normalizes it to all other junctions using a max/min normalization. The final weighted score for each possible circle is calculated, and circles that score above a 60 are added to the final list of circles that have the highest probability of existing within the genome.
The minimum score of 60 was chosen through trial and error, increasing each time to eliminate circles that had little likelihood of existing, based on visual inspection. Visual inspection was carried out using the IGV software (`https://igv.org/`). The alignment of the few lowest scoring circles' junction tags to their sequence ends was determined, and if the circle showed little to no evidence of alignment (with or without a microhomology), the minimum criteria was raised to eliminate these circles from the final list.

3.3 Reproducibility

To replicate the experiment, first make sure that the `pysam` package is installed (if not already):

```
$ pip install pysam
```

Also, ensure that both the BAM and FASTA files from the shared Google Drive are uploaded into the `/data` folder.

Then, clone the repository (`https://github.com/drewapar/MicroDNA`) and then run the following commands from the root directory of the repository to find and output all circles with their scores.

```
$ git clone https://github.com/drewapar/MicroDNA.git
$ cd MicroDNA
$ python src/microDNA.py
```