MicroDNA Project

Tavin Turner

May 8, 2025

1 Introduction

MicroDNA are short, circular, non-coding DNA molecules which form a loop upon themselves. The detection of MicroDNA is deeply important to beginning to understand their role in believed influences, particularly cancer biology. We demonstrate an effective method to produce viable MicroDNA candidates from sequence alignment data, which has been designed for flexibility to parameterize scoring thresholds.

2 Method

From short-read sequencing data, we can accumulate potential circles based on their fundamental patterns: a circle junction and a homologic body. The circle junction is often split in reads, but always unmatched with the reference genome, so we refer two parts: the start junction tag and the end junction tag.

We first iteratively identify alignments with start junctions, then identify end junctions among other alignments in the remainder of the start alignment's run, and consider the pair a candidate circle. Candidate circles are then scored by the size of their largest overlap o – that is, how much of the circle junction is coded in both the start and end junction – and the circle's overall depth d. Specifically, $s = o\frac{d}{50}$.

To identify these reads in alignment data, we depend on CIGAR characteristics, which summarize the alignment behavior of a read, including M (match) and S (soft clipping). The junctions should be soft clipped.

A read with a beginning junction has a soft-clip, then match. A read with an end junction has a match, then a soft-clip. If a read with an end junction is found in the range of the match sequence of a beginning junction, they are considered as a circle candidate.

Circle candidates can be then pruned by a score threshold, default to 1.0, and reported.

3 Results

The results appear promising and relatively performant. For every run, 10 circles were manually validated for plausibility by random selection, from the output (Figure 1). Each time, they did not disobey any rule of microDNA and appear successful, although more study would be needed to select a suitable threshold.

```
SRR413984.19567624: 6S36M / GAGAATCTTCTTCGACTGCCAAGAGCGGTCCAAGGCCAAGCT @ 15503
 score = 4.20 :: 37M5S / CTTGGACCGCTCTTGGCAGTCGAAGAAGATTCTCCTGGAGAA / ...
SRR413984.7436320: 11S31M / TCAAACAAAAAAAAAAACAAAGTGCTAGACTTGGCCAGGCACG @ 553691
 score = 1.80 :: 37M5S / TTTTGTTTGAGACGGAGTCTCACTCTGTCGCCCAGGCCAGTC / 553712 / ...
SRR413984.12298944: 4S38M / GCATGAGTAGGTGGCCTGCAGTAATGTTAGCGGTTAGGAGGA @ 569551
 score = 2.52 :: 39M3S / GGGGTGGCGCTTCCAATTAGGTGCATGAGTAGGTGGCCTGCA / 569569 /  ...
SRR413984.17867790: 5S37M / TCCTCGGCTTCTCCACCTGTACAGGCCAAGGGGAAGCAGGCC @ 659477
 score = 1.96 :: 33M9S / CATTGGGAGGGTCTCCGGTTCCCTGAGCCTGTCCTCGGCTTC / 659503 / ...
```

Figure 1: Truncated sample output displaying potential circles and their score, for threshold of 1.0.

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/itsTurner/microdna.git \
            microdna
$ cd microdna
$ pip install -r requirements.txt
$ python3 src/explore.py \
    -r data/SRR413984.sorted.NC_000001.10.bam \
    -t 1.0 \
    -o docs/results/circles.txt
```