

PolyAStatist - tool for collecting statistics about poly-A tails of non-LTR retrotransposones.

Poly-A tail is important characteristic of Non-LTR retrotransposons. It could implicitly show age of transposition event and help with finding repeats de novo. PolyAStatist works with output files of Repeat masker and outputs percentage of repeats with poly-A tails to total number of repeats.

Requirements:

g++ 4.8 (or higher) or Microsoft Visual C++ compiler with c++11 support.
Windows or Unix OS

Installation:

Just run make from command line

Usage:

To use PolyAStatist you should pass it list of files to process. You could pass as many files as you want, but files should go in pairs, where the first one is RepeatMasker output file (.out extension) and second is corresponding fasta file (.fa or .fasta extension).

Examples of correct usage:

```
./PolyAStatist chr1.fa.out chr1.fa chr2.fa.out chr2.fa chr3.fa.out chr3.fa chr4.fa.out chr4.fa
./PolyAStatist chrX.fa.out chrX.fa
```

Examples of incorrect usage:

```
./PolyAStatist chr1.fa chr1.fa.out chr2.fa chr2.fa.out chr3.fa chr3.fa.out chr4.fa chr4.fa.out
./PolyAStatist chr1.fa.out chr2.fa.out chr3.fa.out chr4.fa.out chr1.fa chr2.fa chr3.fa chr4.fa
```

You could find full information about repeat masker usage at <http://www.repeatmasker.org/>

But it is enough to run repeat masker like this:

RepeatMasker myFile.fasta

Dont forget about species flag - it improves sensitivity a lot.

ATTENTION

PolyAStatistic works only with files, generated from single sequence.

Workflow

To find polyA tails polyAStatistic make several steps:

- 1) Filter LINE Retsrotransposones from another repeatitive elements.
- 2) Merging some cases, when RepeatMasker incorrectly determine repeat with 5' inversion as pair of repeats.
- 3) Cut off to short repeats (<1000 bp) as nonrepresentative.
- 4) Check several positions on 3' end of repeat for poly-A tail
- 5) Output the results

To change some features (as length of repeat of filter another family of repeats) you could change the code as you wish - it has some comments for clarification.

Output

PolyAStatist has following output:

The first line has one of two variants:

- a) In ... found - if PolyAStatist worked with single pair of files
- b) In ...and x others found - if PolyAStatist worked with x + 1 pairs of files

The second line is header od table:

Family Of Repeat;Number of repeats;Number of repeats with poly-A tails;Percentage of repeat with poly-A tails

Other lines contains rows of table in following format:

LIP3;1440;672;0.4666

LIPB3;888;672;0.7567

L1HS;1560;1344;0.8615

...

Feedback

For complaints and suggestions please write to 1dayac@gmail.com to Meleshko Dmitrii.

you could find source code at <https://github.com/1dayac/AUSpringRepeats>