4ureliek / Parsing-RepeatMasker-Outputs

Branch: master ▾ Parsing-RepeatMasker-Outputs / README.md

Find file Copy path

4ureliek parseRM_merge_interrupted.pl 6c23f4c on Oct 19, 2017

1 contributor

162 lines (116 sloc) 7.73 KB

Parsing-RepeatMasker-Outputs => parseRM

Last update: 2017 Oct 18

Of interest if you are using the software RepeatMasker, for transposable elements (TEs) annotation [see http://www.repeatmasker.org/].

This is a collection of perl scripts I wrote for me and my lab to facilitate TE annotation (see list below) and the extraction of information from Repeat Masker output file ".out" or ".align". These scripts may help refining a de novo TE library, but won't help you to actually make one.

They have been used in several publications:

Mitra, Li et al (2012) Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. PNAS 110 (1) http://www.pnas.org/content/110/1/234.long - Figure 1

Hamilton, Kapusta et al (2016) Structure of the germline genome of Tetrahymena thermophila and relationship to the massively rearranged somatic genome. eLife 2016;10.7554/eLife.19090 https://elifesciences.org/content/5/e19090 - Figure 6

Kapusta and Suh (2016) Evolution of bird genomes — a transposon's-eye view. Ann. N.Y. Acad. Sci http://onlinelibrary.wiley.com/doi/10.1111/nyas.13295/full

Kapusta, Suh and Fechotte (2017) Dynamics of genome size evolution in birds and mammals. PNAS Plus http://www.pnas.org/content/114/8/E1460.full (http://www.biorxiv.org/content/early/2016/10/16/081307)

Note that for all of the scripts, a complete usage will be obtained by simply launch them without any argument or option -h

===

parseRM.pl

```
WHAT IT DOES:
Parse RepeatMasker outputs .out but also .align which is much better for landscape graphs
(and this scripts is replacing parseRM_GetLandscape.pl)

There are 3 non exclusive parsings types (they can be set together):
 -p To get a summary of the masking, as well as amount or DNA,
      counts of fragments, + several other details for each repeat name
      (all-repeats file), family, class and total amount (summary file)
      To deal well with the .align positions instead of segments are considered,
      so it is slow (several hours on a bird genome)
 -a To determine the amounts of DNA in a genome that is masked by repeats
      of different lineages / %divergence categories
 -l To split the amount of DNA by bins of %div or My, allowing to generate
      landscape graphs for each repeat name, family or class (one output for each)

Note: if all 3 options -a, -t and -l are set, there will be an additional output
      with bins by %div or My, but by age categories (specified in -a)

Type perl parseRM.pl without options for more details & examples,
or perl parseRM.pl --help for a full usage and more examples.
```

===

parseRM_GetNesting.pl

```
WHAT IT DOES:
    This script reads a Repeat Masker output (.out) and find nested groups
    see below for more details on method and outputs
    Note that the TE that suffered the insertion is the nesting TE. The TE that inserted is the nested TE.
    It provides 3 outputs:
        - only nested/nesting blocks
        - original lines of RM output, with annotations of nesting/nested
        - all lines, but with corrections of coordinates. TEs fragmented by nesting events are merged in one line
          (noted by additional column with number of frags in it)
          note that coordinates of the nesting TE will be true, but WRONG TO CALCULATE LENGTH MASKED BY IT.
```

```
            For that, use additional column with real lenght
           [NOTE: SEEMS BUGGED - use parseRM_merge_interrupted.pl instead if corrected coordinates is what you need]

  METHOD (from Qi Wang)
      For each TE, the previous and next TE in the file are compared. If the previous and next TEs have:
          - the same Repeat Name
          - the same strand
          - genomic End of the previous TE is within 50 bp of the genomic Start of the nested TE
          - genomic Start of the next TE is within 50 bp of the genomic End of the nested TE
          - repeat End (in consensus) of the previous TE is within +/- 20 bp of the repeat Start (in consensus) of
  the next TE
          Then the TE is determined to be nested within the previous/next TE.

  SUMMARY OF THE NESTED/NESTING STRUCTURES FOUND BY THE SCRIPT (frg = fragment):
      Frg A in B:                           [BBBBBB][AAAAAA][BBBBBB]
      Frg A in B in C:              [CCCCCC][BBBBBB][AAAAAA][BBBBBB][CCCCCC]
      Frg A in B in C in D: [DDDDDD][CCCCCC][BBBBBB][AAAAAA][BBBBBB][CCCCCC][DDDDDD]

      Two indep. nested frg in C:                      [CCCCCC][XXXXXX][CCCCCC][XXXXXX][CCCCCC]
      Two indep. nested frg in C, nested in D: [DDDDDD][CCCCCC][XXXXXX][CCCCCC][XXXXXX][CCCCCC][DDDDDD]

      Two frg in C:                      [CCCCCC][BBBBBB][AAAAAA][CCCCCC]
      Two frg in C, nested in D: [DDDDDD][CCCCCC][BBBBBB][AAAAAA][CCCCCC][DDDDDD]

      Three frg in C:                      [CCCCCC][AAAAAA][EEEEEE][BBBBBB][CCCCCC]
      Three frg in C, nested in D: [DDDDDD][CCCCCC][AAAAAA][EEEEEE][BBBBBB][CCCCCC][DDDDDD]
```

==

parseRM_GetNesting.pl

```
  Usage [v$VERSION]:
     perl $SCRIPTNAME -i <genome.out> [-t <TEinfo.tab>] [-m] [-n] [-v] [-l] [-h]

  This script reads a Repeat Masker output (.out) and corrects coordinates based on
  the nesting blocks (last column of RM.out). Once this is done, if -m is set
  the script will then look for more interrupted repeats.
```

==

parseRM_Coverage.pl

```
  WHAT IT DOES:
      This script will output the coverage of a repeat plotted on the consensus length
      (if the plotting fails, -Rfile will output command lines for R to get the plots)
      It is identical to the script in the Utils of this repository: https://github.com/4ureliek/TEanalysis
      Here, for a simple parsing of repeat masker output, use -type RMout

      Check the usage for the many options (such as filtering etc)!
      perl parseRM_Coverage.pl -h

  NOTE:
      Requires the Statistics::R perl module
```

==

parseRM_ExtractSeqs_P.pl

```
  WHAT IT DOES:
      This script reads a Repeat Masker output (.out) and extract sequences.
      It will reconstructed interrupted repeats, and allows extraction of a random subset + its complementary

      Check the usage for the many options (such as filtering etc)!
      perl parseRM_ExtractSeqs_P.pl -h

  NOTE:
      Requires the forks perl module
```

==

parseRM_simple.pl

```
  DEPRECATED (now that parseRM.pl -p is faster)
  Parse RepeatMasker outputs (.out only) to get summary info for each repeat as well as masked amounts by class and
  family.

  For the repeats, it will provide:
   - fragment number (frg nb):
        all
        frg nb from start to end of consensus
        frg nb corrected for interupted repeats (using the "ID". e.g. from the Repeat Masker .out)
   - %div, ins, del: pondered average, median
   - length masked + %genome by this repeat
   - amount of DNA that is masked several times (usually 2) by this element

  Summary will provide amounts and % masked by various class, families
  Plus the script provides a set of files with overlap info to help giving real amounts/%


==================================================================
```

parseRM_GetLandscape.pl

```
  DEPRECATED - use parseRM.pl with the option --land instead, and on .align files for smoother graphs.
```