4ureliek / ReannTE

Branch: master ▾ ReannTE / README.md Find file Copy path

4ureliek Edit README of ReannTE f59bd40 on Dec 15, 2016

1 contributor

112 lines (89 sloc) 5.92 KB

# ReannTE

Last Update: 2016 12 15

Scripts to facilitate transposable elements consensus sequences curation

========== ReannTE_Filter-mRNA.pl

```
   WHAT IT DOES:
   Blastx the (consensus) sequences against a database that can be defined, unless remote blast is used.
   (if -remote is chosen, the default database is refseq_mrna)
   Then the sequences are filtered out from the input file if they correspond to unclassified TEs
   (no class/family defined, or the class or family are "unclass" or "unknown")

   perl <scriptname.pl> -i <fa> [-b <blast-path>] [-e <XX>] [-forceB <X>] [-remote]
   OR
   perl <scriptname.pl> -i <fa> [-b <blast-path>] [-e <XX>] [-forceB <X>] [-db <fa>] [-dbt <XX>] [-bt <XX>]

   MANDATORY ARGUMENTS:
   -i <fa>    => fasta file

   [OPTIONAL ARGUMENTS]:
   -blast <path> => path = localisation of ncbi blast software
                    if no path provided, path = /home/software/ncbi-blast-
  2.2.25+
   -e <XX>       => XX = threshold, evalue (default = 10-10). It sets the minimum evalue to eliminate a sequence.
   -forceB       => set x to chose how to behave if previous <fa>.blast.out exists
                    x = 0 (default), chose this to avoid redoing the blast if <fa>.blast.out file already exists
                    x = 1, chose this to save existing <fa>.blast.out (renamed), but still rerun blast
                    x = 2, chose this to delete the pre-existing <fa>.blast.out file (therefore blast will be
  redone)
   -remote       => use the -remote option of blast if you don't have the -db. This takes a while.
   -db <fa>      => database to blast against [not relevant if -remote]
   -dbt <XX>     => dbtype option of makeblastdb [default = nucl] [not relevant if -remote]
   -bt <XX>      => blast type [default = tblastx] [not relevant if -remote]

   REQUIREMENTS:
   - Blast software
   - Bioperl
```

========== ReannTE_FilterLow.pl

```
   WHAT IT DOES:
   This script uses Repeat Masker to mask low complexity / simple repeats of the input fasta file
   (for example, RepeatScout output)

   It eliminates the ones that are more than XX% masked (-p option)
   2 fasta outputs: retained sequences and rejected sequences

   perl <scriptname.pl> -i <fa> [-r <RMpath>] [-p <XX>

   MANDATORY ARGUMENTS:
   -i <fa>    => fasta file

   [OPTIONAL ARGUMENTS]:
   -r <path>  => path = localisation of repeat masker software
                 if no path provided, path = /home/software/RepeatMasker
   -p <XX>    => XX = threshold, in % (default = 80%). It sets the minimum low complexity masked % required to
  eliminate the sequence

   REQUIREMENTS:
   - Repeat Masker software, crossmatch engine
   - Bioperl (Bio::DB::Fasta, Bio::SeqIO)
```

========== ReannTE_MergeFasta.pl

```
   WHAT IT DOES:
```

```
 This script facilitates merging two consensus libraries
 - mask a with b (and b with a just to have access to it in case if needed)
 - parses the masking outputs to evaluate overlaps
 - make choices and flag sequences to keep or not. Note that all info are printed in an output,
   to facilitate manual verification (advised)

 perl <scriptname.pl> -a <seqs_1.fa> -b <seqs_2.fa> [-p <x>] [-s <x>] [-forceRM <x>] [-gc <XX>] [-RM <path>]
[-project <name>] [-CheckLow <XX>]

 MANDATORY ARGUMENTS:
 -a <seqs_1.fa> => first fasta file
 -b <seqs_2.fa> => second fasta file

 [OPTIONAL ARGUMENTS]:
 -p <x>           => priority setting to favor or not one of the files when choice of sequence to keep
                     x = a or b, give priority to file a or b when choice is not clear
                             x = no (default), both sequences will be kept
 -s <XX>          => \"span\" corresponds to the minimum percentage of the sequence that is masked by another one
to consider eliminating it
                     The value [default = 80] will be used as a threshold to make choices on sequences to keep.
                     For ex, if >XX% of sequenceA is masked by <XX% of sequenceB, sequenceB is kept.
                             However, if <XX% of sequenceA is masked by <XX% of sequenceB, both are kept.
 -forceRM        => set this to chose how to behave if previous .out exist
                     x = 0 (default), chose this to avoid remasking if .out files already exist for files set as
-a and -b
                     x = 1, chose this to let RM check for existing .out (RM will move them if they do)
                     x = 2, chose this to delete the pre-existing .out files (therefore masking will be redone)
 -gc <XX>        => GC content (%) of the genome of the species considered, for use of good matrix in repeat
masker
 -RM <path>      => path = localisation of repeat masker software
                             if no path provided, path = /home/software/RepeatMasker_405
 -project <name> => name = will be in the name of the output files, including the merged fasta
                             if nothing provided, default = \"MergeFasta\"
 -CheckLow <XX>  => chose this option to remove low complexity sequences before doing anything to merge libraries.
                             XX = threshold, in % (80% is advised). Set the minimum low complexity masked %
required to eliminate the sequence.
 -v              => verbose mode, make the script talks to you
 -v              => print version if only option
 -chlog          => print change log (updates)
 -h|help         => Print this help

 REQUIREMENTS:
 - Repeat Masker software
 - that ALL sequences have a unique name (e.g. name before the #)
   if several different consensus have the same names between the 2 libraries this will create errors
   you can use sed (see below) to add a number in front of all sequences of one of the files to avoid that issue
   in the case of merging 2 repclass outputs for ex: sed 's/>/>1_/' seqs_1.fa > seqs_1.ok.fa
```