The manual of MITE-Hunter

09/08/2010

Yujun Han

- **What can MITE-Hunter do?**

MITE-Hunter was designed to identify miniature inverted-repeat transposable elements (MITEs) as well as other small (< 2Kb, default) class 2 nonautonomous transposable elements (TEs) from genomic DNA datasets. Class 1 TEs and long TEs can't be found by MITE-Hunter. TEs with too many mutations and mismatches in the terminal inverted repeats (TIRs) may not be detected.

- **The input file:**

Only one input file is required by MITE-Hunter and it is the genomic DNA sequences in fasta format. Such input file needs to be indexed, which can be done automatically by MITE-Hunter. To make sure your date is safe; please save a copy because MITE-Hunter may change the file format.

- **Other software required by MITE-Hunter:**

NCBI BLAST - http://www.ncbi.nlm.nih.gov

Muscle - http://www.drive5.com/muscle/

mDust - http://compbio.dfci.harvard.edu/tgi/software/

(If the links provided above don't work, just google it)

- **Installing MITE-Hunter:**

Please run MITE_Hunter_Installer.pl, which needs 5 parameters:

1) -d the full path to the folder of MITE-Hunter.
2) -f formatdb program
3) -b blastall program
4) -m mdust program
5) -M muscle program

Full path name is required if necessary and no space was allowed in the folder names. For example:

perl MITE_Hunter_Installer.pl -d /iob_home/srwlab/vehell/scratch/Work_Room/Temp_Room/Yujun/ -f formatdb -b blastall -m /iob_home/srwlab/vehell/scratch/Tools/mdust/mdust -M /iob_home/srwlab/vehell/scratch/Tools/Muscle/muscle3.6/muscle

- **Running MITE-Hunter:**

Although MITE-Hunter is composed with many perl scripts, it can be used by typing one command, eg:

perl MITE_Hunter_manager.pl -i /iob_home/srwlab/vehell/scratch/Data_Room/Rice/rice_b5 -g Rice -n 3 -S 12345678 &

In the above example, 3 cpu were applied and "-S 12345678" means it starts from the very beginning step to the last step. If for some reason the program stopped in the middle, you can run it again starting from where it stopped. **But if you want to repeat a step, please delete the old output files for that step, otherwise the program may not work.**

For another example:

perl MITE_Hunter_manager.pl -i /iob_home/srwlab/vehell/scratch/Data_Room/Maize/maize -g Maize -n 5 -S 12345678 –P 0.25 &

-P 0.25 means that MITE-Hunter only sample 25% of the whole dataset to search for TEs.

- **Three Important parameters for MITE-Hunter manager:**

-i (Genomic Sequence file)

-P 1 or decimal fraction (default 1)

This is the parameter that you may change. 1 means MITE_Huner will find TEs from the whole given fasta file. However, for some huge genomes such as human or maize, it will take too much time. In this case, you may input, for example 0.2, which means MITE_Hunter only use 20% of the given sequence and do the TE discovery. I recommend using 1 for genomes that is smaller than 700 Mb.

-g A word (default "genome")

This is the name for your task. All output file names start with the word you input here. Names of all the discovered TEs will also start with this word.

- **Other parameters:**

-F 0 or 1 (default 0)

The genomic sequence file needs to be indexed before being used. By default, MITE_Hunter automatically does it.

-w integer (default 2000)

The maximal length of a TE that can be found by MITE_Hunter.

-s integer (default 1500)

-n integer (default 5)

This is the maxcimal group number to be used.

-c integer (default 5)

This is the maxcimal CPU number to be used.

-d decimal fraction (default 0.2)

For a candiate TE, if its sequence has more than this percentage (0.2 = 20%) of low complexcity sequences, such as "AAAAA...", "TATATATA..." or "GGGGG..", it will be filtered.

-f integer (default 60)

This is the length of the flanking sequences MITE_Hunter will use to help judging whether a TE is real or not. In the multiple alignment file, by default, you will see 60 bp flanking sequence on each side of the TE copies.

-t integer (default 10)

This is the length of terminal inverted repeat (TIR) that is used to find candidate TEs.

-M integer (default 10)

This is the length of the longest target site duplication (TSD) for detecting candidate TEs.

-l 1 (default)

This is the maximal unmatched bp in the TIR region for detecting candidate TEs.

-p decimal fraction (default 0.2)

For a candidate TE, if its sequence has more than this percentage (0.2 = 20%) of low complexity sequences, such as "AAAAA...", "TATATATA..." or "GGGGG...", it will be filtered.

-L Integer (default 90)

This is the minimal shared length between TEs that will be grouped together. This is a parameter for generating exemplars.

-I Integer (default 80)

This is the minimal identity between TEs that will be grouped together. This is a parameter for generating exemplars.

-m Integer (default 3)

This the minimal copy number for a qualified TE.

-T DNA sequences separated by “_” (default TA_)

This is the given TSD sequence. The default is "TA_", and it means if a candidate has a 2 bp TSD, it should be "TA". If you want to change it (I don't recommend this), input as "TC_TGC_".

-C 0 or 1 (default 0)

If the input here is "0", MITE_Hunter uses the whole sequences as query to find its copies and check whether they are real TEs.

If the input here is "1", MITE_Hunter uses 200 bp from each end of the query to find its copies and check whether they are real TEs.

-A Integer (default 90)

If a TE candidate has more than this number of continues low complexity bp, it will be filtered.

-S (default 1)

There are 8 steps. Input the steps that you want MITE-Hunter to do, such as "12" or "345678" (if you finished step 1 and 2).

- **The output:**

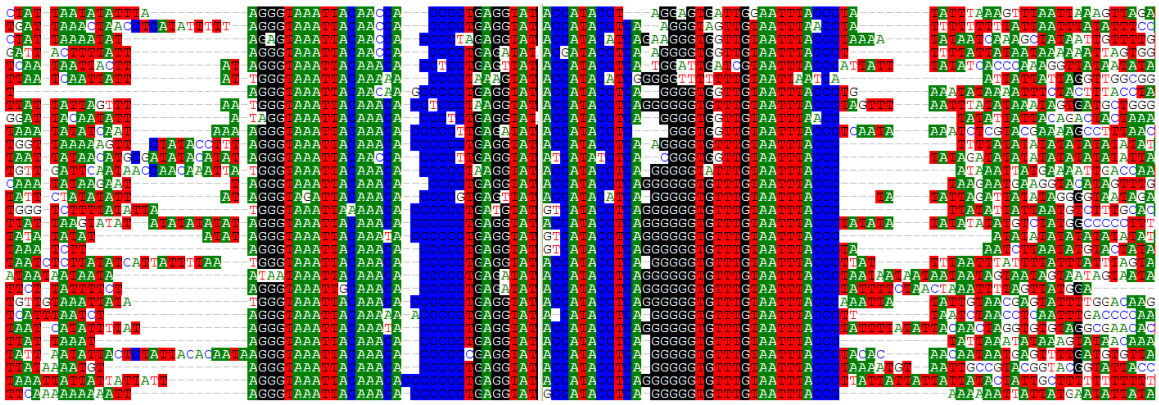
The output files of MITE-Hunter include consensus TE sequences grouped into families and their multiple alignment files. Those with ".aln.elite" are mutiple sequence alignment (MSA) files. TE consensus sequences are in files that have "Step_8" in their names. Each "Step_8" file contains a TE family, except for two. One is "Step_8.singlet" that contains all TEs that don't have similar homologs in the output, and the other is "Step_8.paired" that contains putative compound TEs. A compound TE is composed with two different or same TEs and should be excluded. To manually check or classify a TE sequence, a user can open the MSA file of the TE to check the TIR and TSD structures. I recommend using Bioedit, which can label letter in different color, set consensus

percentage and split the window so that both of the head and tail of the alignment can be viewed at the same time.

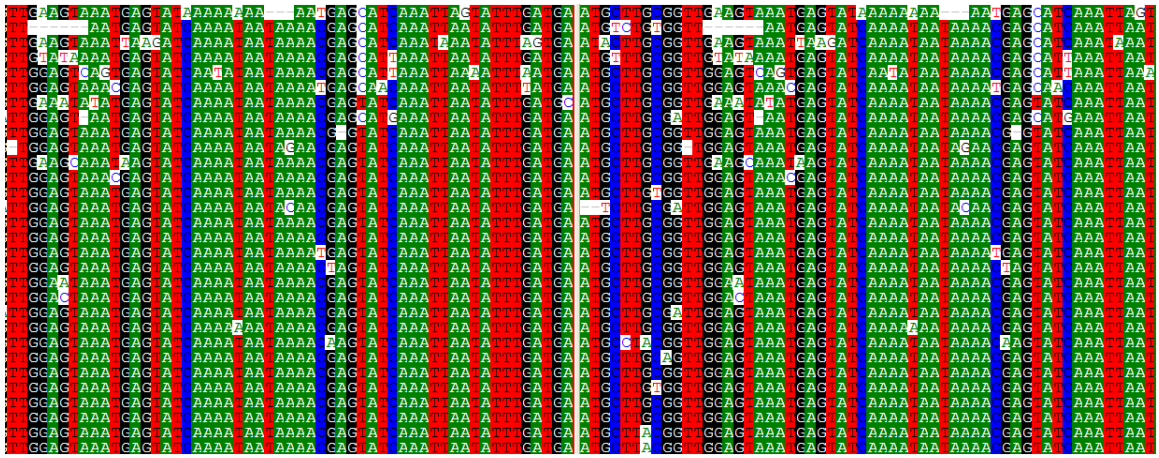
- **How to find out whether a TE is real or false from the MSA file?**

  Because MITE-Hunter retrieves homolog TE copies with 60 bp flanking sequences (default), in the MSA file, sequences within 60 bp on each side shouldn't have similarity among copies.

A good example: TIR and TSD can be determined.



A bad example: TIR and TSD can't be determined.

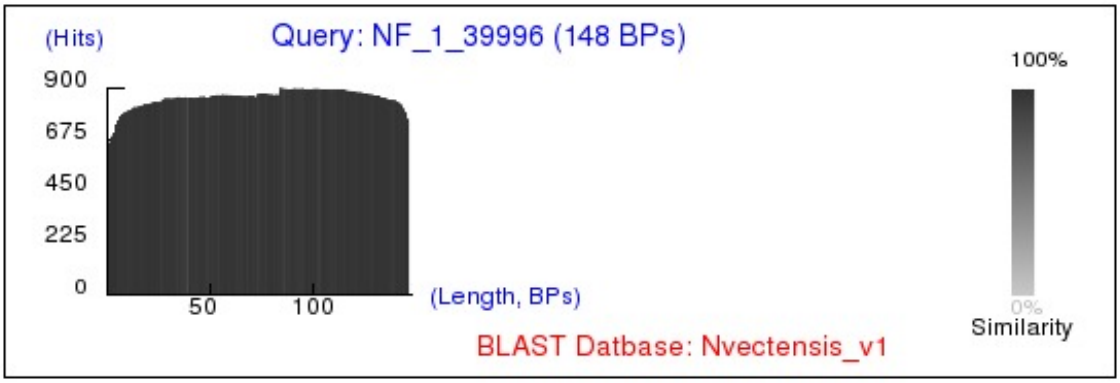


- **How to find out whether a TE is a compound?**

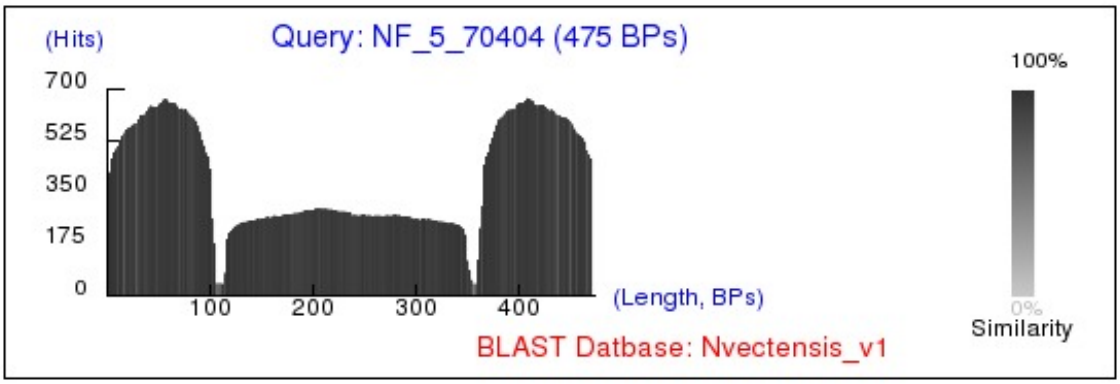
There are several ways to do this. I recommend using TARGeT to check the copy number coverage along the element. TARGeT can be visited at http://target.iplantcollaborative.org/

You need clik the DNA query option, copy and paste the TE sequence and choose the database, then click BLAST. A figure will show up like those below.

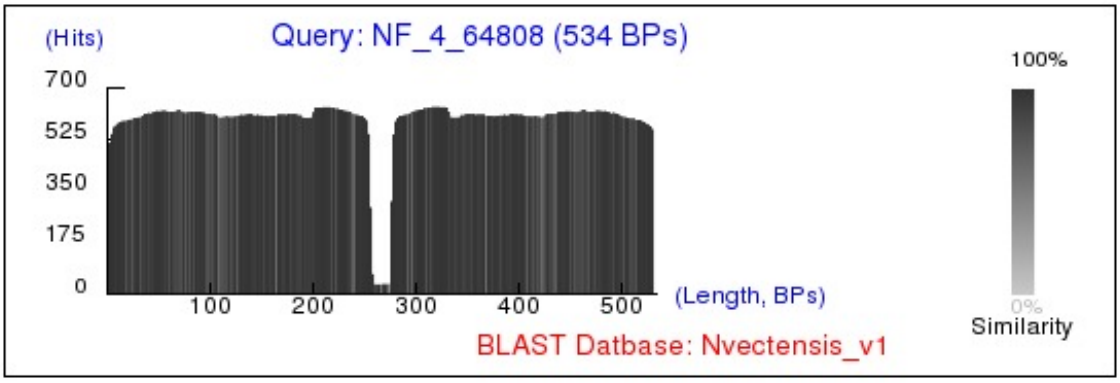
A good example: the coverage is even along the element.



A bad example: Coverage is not even because this element is composed with two different TEs that have different copy number in the genome.



Another bad example: Coverage is almost even except the gap in the middle, because this element is composed with two identical TEs that have same copy number in the genome.


(Hits)
Query: NF_4_64808 (534 BPs)
700
525
350
175
0
100
200
300
400
500
(Length, BPs)
100%
0%
Similarity
BLAST Datbase: Nvectensis_v1