User Manual v1.2

LoRTE

Detecting transposon-induced genomic variants using low coverage PacBio long read sequences

Eric Disdero & Jonathan Filée

February 2017

1. Introduction

LoRTE is a population genomics tool designed to detect insertions and deletions of transposable element sequences between a reference genome and the genomes of closely related populations, strains or species. It has been designed to use long read sequences such as PacBio reads.

2. Installation

Requirements

You will need Python 2.7 to run LoRTE and the ncbi-blast+ package (at least the 2.2.28 version or a more recent one). BioPython is also required. LoRTE has been designed to require low amounts of RAM, but large datasets (>5Gb of reads) would require a significant volume of memory (>12GB of RAM).

Install LoRTE

Download the tarball file and type:

```
tar -xvf LoRTEv1.2.tar.gz
chmod +x LoRTEv1.2.py
```

LoRTE reads a configuration file called «LoRTE-Parameters». You will need to edit this file using a text editor to specify the different paths leading to your data as well as to define your own search parameters.

3. Running LoRTE

Type:

```
python LoRTEv1.2.py /home/user/LoRTE-Parameters/
```

4. Input files

LoRTE need five input files:

1. A *reference genome* in fasta format, example:

```
>2R
ATATATAGGGTATAT...
>2L
TATATATATAGGGG...
etc...
```

2. An *annotated list of Transposable Elements and their coordinates* in the reference genome, for example in a tabular format: 'name of the TE' 'name of the contig or chromosome' 'start' 'stop' :

TE_1	2R	22855992	290131
TE_2	X	11465540	11465725
TE_3	2L	13560515	13565210
TE_4	X	15316007	15317462
TE_5	X	14483763	14483948

3. A *Transposable Element consensus file* in fasta. Example:

```
>MARINA#DNA/TcMar-Mariner RepbaseID: MARINA
AGAGAGAGAT...
>R1B_DVi#LINE/R1 RepbaseID: R1B_DVi
CAGTTTTTTTT...
etc...
```

Drosophila TE consensi can be downloaded via FlyBase at ftp://ftp.flybase.net/releases/current/precomputed_files/transposons/transposon_sequence_set.embl.txt.gz

4. A fasta file with *the reads.* Example:

```
>m160313_143700_42263_c10092286255003209605121667_s1_X0/15/15459_18010 RQ=0.849
cactggaagactcggagttatgtgcccttcgagaggaccaagaagagaacggaaaatgggtaggaaattattcccataaaaa...
>m160313_143700_42263_c10092286255003209605121667_s1_X0/16/0_10091 RQ=0.838
aaaaagtgagaggagataatgtctctggaaatatagagtatcaatagctggatcgataatgcagcgttactgttctgttcgcgctt...
etc....
```

5. The «LoRTE-Parameters» file with your options (see next section)

5. Editing the «LoRTE-Parameters» file

Most of the parameters used by LoRTE are flexible and need to be indicated in the LoRTE-Parameters file. Here is the complete list of the parameters with some details in italic :

```
>Complete path to the Reference Genome:
/home/user/LoRTE/my_reference_genome.fa

>Complete path to the TE list:
/home/user/LoRTE/my_TE_List

>Complete path to the Reads :
/home/user/LoRTE/my_PacBio_reads

>Complete path of the TE consensus:
```

/home/user/LoRTE/my_TE_consensus_list.fa

>Name of results folder (full path+folder, it will be created)
/home/user/LoRTE/my_results

>E-value for BLASTN & MEGABLAST
1e-40

We recommend the use of 1e-40 due to the error rate of the PacBio sequences. More stringent criteria should be used for error-free technologies (ex: 1e-60)

>Read coverage of your reads
10

LoRTE provide reliable results with low level of coverage (<10X) but higher coverage can be useful to identify low frequency events and to estimate the level of polymorphism.

>Format of the TE list: nameTE,nameContig,SStart,SStop,Sens('XXXXX'if not defined), separator_col, first line of data(line 0 can be descriptive of a blast output) ex:0,3,2,4,"XXXXX","\t",0 for a BLAST output file in which the order would be: nameTE, qstart, qend, Subject, Sstart, Sstop separated by tabulation, orientation is not precised, start with no description line)
4,0,1,2,XXXXXX,\t,0

Examples :

0,1,2,3,XXXXXX,\t,0 correspond to:

```
TE_1    2R      2285599         2290131
TE_2    X       11465540        11465725
```

0,1,2,3,4,\t,0 correspond to:

```
TE_1    2R      2290131         2285599         -
TE_2    X       11465540        11465725        +
```

3,0,1,2,XXXXXX,\t,0 correspond to:

```
ACPB03022661.1          79683   80998   TE_1
ACPB03022661.1          57199   58459   TE_2
```

>Maximum window size between two flanking sequences:
10000

A larger window should be specified if your TE list contain unusually long TEs.

>Length of the flanking sequences that will be collected
200

This parameter specifies the size of the flanking sequences that will be collected, 200nt provide the best results in our tests but a longer sequence (>1kb) should also be useful for repeat-rich genome.

>Number of CPU to be used
2

6. Output files

LoRTE creates a Summary folder that report a list of five output files:

1. «*List-Present-Absent*»

For each annotated TE in the reference genome, this tabular file indicates their presence/absence status in the reads. The first line summarizes the global statistics of the analysis. The table indicates the number of reads that support a call :
- PositiveHit-TE column indicate the presence of the corresponding TE
- NegativeHit(Tot) colomun indicate the total number of read testifying for the absence of the TE (>50nt + <50nt)
- NegativeHit<50nt column indicate the presence of a DNA fragments <50nt that does not align with the TE
- NegativeHit>50nt column indicate the presence of a DNA fragments >50nt that does not align with the TE

- The results column report the final decision (TE present / TE absent / ambiguous etc..). «Unresolvable locus» indicates a TE present in repeat-rich region in which a reliable call is impossible.

2. «*List-Present-Absent-Sequence*»

This file reports the sequences in fasta format for each deletion event detected by the program. For each event, the first sequence corresponds to the reference genome sequence and the others correspond to read sequences (TE sequence + flanking sequences). For the reads, identical sequences are reported only once. This file allows the user to visualize the calls by copying and pasting each block of sequence in an alignment viewer.

3. «*List-Present-Absent-Polymorph-sequence*»

This file reports the sequences in fasta format for each polymorphic locus detected by the program in a format similar to the previous file.

4. «*List-new-insertions-sequence*»

This file reports the sequences in fasta format for each new insertion identified in the reads. The format is similar to the «List-Presence-Absence-Sequence» except that there is a number after each TE identifier. This number indicates the total number of reads supporting the call. For example:

```
1264_Mariner_H 4
>m160322_063607_42263_c10098623291692_s1_X0/135087/8021_16095	2323	4052
GTTGAGGTTACTGGT....
>3R	12835	13255
ATATTCACTCGAACAA....
```

The number **4** indicates that four different reads support an insertion of a Mariner_H element in the chromosome 3R region of the reference genome.

5. «*List-new-insertions-sequence-Polymorph*»

This file reports the sequences in fasta format for each new polymorphic insertion identified in the reads. For each call, the first sentence indicate to which TE family belong the insertion. The first sequence correspond to the new insertion (and the number of read supporting this prediction). The second sequence correspond to the empty site in the reference genome. The remaining sequences corespond to the empty sites identified in the reads.

7. Contact

In case of questions and suggestions of improvement, do not hesitate to contact us:

eric.disdero@gmail.com
jonathan.filee@egce.cnrs-gif.fr