PiRATE: a Pipeline to Retrieve and Annotate TEs

2 December 2017

Wrote by:
Jérémy BERTHELIER

Supervised by:
Grégory CARRIER
TABLE OF CONTENT

1. Requirement ........................................................................................................................................... 3
2. How to run your PiRATE-Galaxy ........................................................................................................... 5
3. Started with your PiRATE-Galaxy .......................................................................................................... 7
4. Started with the PiRATE pipeline ........................................................................................................ 8

   STEP 0: Load and prepare your dataset ................................................................................................. 9
   STEP 1: Detection of putative TEs ....................................................................................................... 15
   STEP 2: Sort your detected sequences ................................................................................................. 19
   STEP 3: Classification ......................................................................................................................... 21
   STEP 4: Manual Check ......................................................................................................................... 22
   STEP 5: Annotation ............................................................................................................................ 24
Requirement

- The PiRATE-Galaxy is installed on a virtual machine name PiRATE-VM. The PiRATE Virtual Machine (PiRATE-VM) can be downloaded at the following URL:
  http://doi.org/10.17882/51795

- To use the PiRATE-VM a virtual machine monitor needs to be installed, for example, VirtualBox
  https://www.virtualbox.org/

- Once your virtual machine monitor is installed, you need to import the PiRATE-VM.
Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

- Made changes according to your computer setting.

Your network setting have to be correctly configured to use the PIRATE-Galaxy.

- Open the PIRATE-VM and the “Jeremy” account, the password is: jeremy07
2. How to run your PiRATE-Galaxy

Now the PiRATE-Galaxy can be launched.

1. Open the Konsole
2. Enter:
   ```
   sh /home/jeremy/galaxy/run.sh
   ```

3. Right click on the URL
4. open it.

The PiRATE-Galaxy is alive!

If not… Check that the network setting of your VM is correctly configured.
1. Started with your PIRATE-Galaxy

“Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.”

(Giardine et al., 2005) https://usegalaxy.org/
4. Started with the PiRATE pipeline
STEP 0: Load and prepare your dataset

- Connect you as administrator to the PiRATE-Galaxy

Username: administrator
Password: administrator
• Three types of data is needed to perform the complete PIRATE pipeline:

1) A genome assembly (FASTA)
2) Illumina raw data (FASTQ)
3) Illumina raw data (FASTA)

• How to load your data?

Your genome assembly should have a weight below 1 Go and can be directly download from your computer to the Galaxy environment with the “choose local file” button and launch “Start”.
However, your Illumina raw data should have a weight of more than 1 Go. Thus it will be necessary to import it with FTP:

1) You need to copy-past your files in the directory “administrator@pba.fr” of the PiRATE-VM/home/Jeremy/Documents/administrator@pba.fr
2) You can now load them into the PiRATE-Galaxy

Download from web or upload from disk

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Available files:
- Name: 1.tq  Size: 1.54 MB  Created: 11/29/2017 02:09:46 PM
- Name: 2.tq  Size: 2.47 MB  Created: 11/29/2017 02:09:46 PM

FTP files

This Galaxy allows you to upload files via FTP. To upload some files, log in to the FTP server at 134.246.55.42 using your Galaxy credentials (email address and password).
1) The genome assembly (FASTA)

To avoid any problems with the tools, your FASTA file containing the genome assembly needs to be formatted with short and simple header names (example: Chromosome1)

```
>Chromosome1
AAATTTAAAAATTTGGCCCCCCAAACCCAAACCCCAACCCCAACCCCAATTTTTAA
...
>Chromosome2
AAATTTAAAAATTTGGCCCCCCAAACCCAAACCCCAACCCCAACCCCAATTTTTAA
...
```

You can rename the headers of your FASTA file with the tool “Rename headers” in the “Text Manipulation” section.

- Your FASTA sequences must be formatted with 60 pb per line.
  You can do this task with the tool “FASTA within” in the “Text Manipulation” section.

Your genome assembly is ready!

2) The Illumina raw data (FASTA and FASTQ)

You should probably have your data in the FASTQ format, it’s ok for the tools dnaPipeTE and RepARK.

However, RepeatExplorer uses FASTA file:

a) Use a single data file for RepeatExplorer

You can convert your single FASTQ file into FASTA with the tool “FASTQ to FASTA converter” in the “Convert Formats” section.

b) Use paired data (advised) for RepeatExplorer
To use the tool RepeatExplorer with the paired data option, the headers of your FASTQ files need to finish by the indication “/1” (forward) or “/2” (reverse).

Exemple:

```
>readname/1
AAATTTTAAATTTGGGCCAAACCCAAACCACAAAACCCAACCCCAATTTTTAA

>readname/2
AAATTTTAAATTTGGGCCAAACCCAAACCACAAAACCCAACCCCAATTTTTAA
```

If not, you can add them by using the tool “Add suffix” in the “Text Manipulation” section.
Do this manipulation for each of your file (forward and reverse). This can be time consuming.

Then, you need to join your forward and reverse file in once with the tool “FASTQ interlacer” in the section “Join, Substract and Group”.

Then, you can convert your FASTQ file containing the forward and reverse into one FASTA file by using the tool “FASTQ to FASTA converter” in the “Convert Formats” section.

Your Illumina raw data in FASTQ and FASTA are ready!
STEP1: Detection of putative TEs

The detection step of PiRATE is flexible, you can use every tools one after one or only select your favorite ones.
Here is the list of the available tools with their authors and the URL:

**Approach 1: Similarity-based detection**


  This tool detects putative TE sequences from the comparison of the genome assembly and a nucleotide databank of known TEs. It is possible to use the default databank of PiRATE or yours.

- **TE-HMMER** is a made-self-tool using HMMER (Eddy and others, 1995) and BLAST (Altschul et al., 1990)

  This tool detects putative TE sequences from the comparison of the genome assembly of your studied organism and a databank composed of profile HMM of known TEs. It is possible to use the default databank of PiRATE or yours.

**Approach 2: Structural-base detection**

- **LTRharvest** (Ellinghaus et al., 2008) [http://www.zbh.uni-hamburg.de/?id=206](http://www.zbh.uni-hamburg.de/?id=206)

  This tool detects LTR from a genome assembly.


  This tool detects LINE from a genome assembly.

- **Helsearch** (Yang and Bennetzen, 2009) [http://omictools.com/helsearch-tool](http://omictools.com/helsearch-tool)

  This tool detects Helitron from a genome assembly.

- **MITE-Hunter** (Han and Wessler, 2010) [http://target.iplantcollaborative.org/mite_hunter.html](http://target.iplantcollaborative.org/mite_hunter.html)

  This tool detects MITE from a genome assembly.

- **SINEfinder** (Wenke et al., 2011) [http://www.plantcell.org/content/23/9/3117](http://www.plantcell.org/content/23/9/3117)

  This tool detects SINE from a genome assembly.
Approche 3: Repetitiveness-base detection

- **TEDenovo** (Flutre et al., 2011) [https://urgi.versailles.inra.fr/Tools/REPET](https://urgi.versailles.inra.fr/Tools/REPET)

  This tool belongs to the REPET package, it allows to detect repeated sequences with RECON, GROUPER and PILER, group them into cluster and create consensus sequences for each cluster.

- **RepeatScout** (Price et al., 2005) [https://bix.ucsd.edu/repeatscout/](https://bix.ucsd.edu/repeatscout/)

  This tool detects repeated sequences by using k-mer method, group them into cluster and create consensus sequences for each cluster.

Approche 4: Build repeated sequences

- **RepeatExplorer** (Novak et al., 2013) [http://repeatexplorer.umbr.cas.cz/](http://repeatexplorer.umbr.cas.cz/)

  This tool samples reads and compare them with BLAST. Overlapping read are connected with a graph-based algorithm and grouped into cluster. Read belonging to each cluster are assembled with CAP3.

- **dnaPipeTE** (Goubert et al., 2015) [https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html](https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html)

  This tool samples reads and assemble repeated elements with Trinity.

- **RepARK** (Koch et al., 2014) [https://github.com/PhKoch/RepARK](https://github.com/PhKoch/RepARK)

  This tool uses a graph-based method to detect abundant k-mers from illumina reads. Abundant k-mers were isolated and assembled using, resulting in a de novo repeat libraries.
Important:

I advised you to change the format of output files obtained from each detection tools.

- Change the headers name by the tool name to know where they come from.

Example:

```plaintext
>LTRharvest_1
>LTRharvest_2
....
```

You can rename the header of your output file with the tool “Rename headers” in the "Text Manipulation" section.

- Your FASTA sequences must be formatted with 60 pb per line.

You can do this task with the tool “FASTA within” in the “Text Manipulation” section.
STEP 2: Sort your detected sequences

- **Remove short sequences:**

  It is possible that the approach 1 and the approach 2 give a high number of short sequences. In order to be more efficient in time and decreased the high number of repeated sequences, you can choose to remove the sequences below a length of 500 pb. The tool “remove short sequences” realizes this task.
- **Concatenated outputs sequences:**
  Sequences detected from the step 1 (without those obtained with MITE-hunter and SINEfinder) can be concatenated for the “Clustering step”. You can use the tool “concat FASTA files” in the “Text Manipulation” section.

- **Clustering to remove redundant sequences:**
  In order to decrease the redundancy, PiRATE uses the tool CD-HIT-est (Li and Godzik, 2006) [http://weizhongli-lab.org/cd-hit/](http://weizhongli-lab.org/cd-hit/). We use it to cluster sequences that are 100% identical to a part of a larger sequence. This allows to remove the redundant shorter sequences which are already detected with a longest length in another sequence. If necessary, it is possible to decrease the percentage of identity.
  We used as setting: $aS= S_S=1$ and $c=\%identity=1$

Figure from [https://github.com/weizhongli/cdhit/wiki/3.-User's-Guide#CDHITEST](https://github.com/weizhongli/cdhit/wiki/3.-User's-Guide#CDHITEST)
STEP 3: Classification

Once your putative TE sequences have been clustered with CD-HIT-est to reduce the redundancy, it generate an output file that you will submitted to classification.

To realize the classification of your putative TEs, PiRATE uses PASTEC (Hoede et al., 2014)  
https://urgi.versailles.inra.fr/Tools/PASTEClassifier

This tool works with as input data a FASTA file with simple headers and with a width of 60 pb for every nucleotide line.
Currently, it is only possible to use PASTEC with the PiRATE databanks (nucleotide, protein and profil HMMs). The use of your own custom databank is still in progress and will be possible in the upgrade version of PiRATE.

**STEP 4: Manual Check**
The manual check step is advised. Here is the method that we used for the annotation of the *T. lutea* genome:

- Three libraries were manually constructed with a “Russian doll” strategy in order to perform separated annotations, a “potentially autonomous TEs library”, a “total TEs library” containing the potentially autonomous TEs and the non-autonomous TEs and a “repeated elements library” containing in addition the uncategorized repeated sequences. Sequences classified as LTR, LINE and TIR were manually sorted in superfamily (according to the evidence section produced by PASTEC).

- To facilitate their manual check, sequences belonging to the same putative superfamily were grouped into families with MCL. The percentage of identity between sequences belonging to the same family were checked with Blastn (-identity: 80%). We followed the 80-80-80 Wicker rules to form families.

- Finally, larger sequences from each TE family were checked and selected for the “potentially autonomous TEs library” according to the presence of TE domains or similarities with Pfam (http://pfam.xfam.org/), NCBI-BLASTx and Censor (http://www.girinst.org/censor/). We define as potentially autonomous LTR, sequences bearing at least a reverse transcriptase and an integrase domain and having similarity with LTR sequences in databanks. We define as potentially autonomous LINE, sequences bearing at least a reverse transcriptase domain and sharing similarity to LINE sequences in databanks. We define as potentially autonomous TIR, sequences having an evidence of a transposase domain or similarity to TIR sequences in databanks.
No manual check were performed for sequences classified as non-autonomous TEs. Sequences classified as SINE, MITE and TRIM were directly selected for the “total TEs library”. Only sequences classified as LARD, which were obtained with the repetitiveness-based approach of TE detections (TEdenovo or Repeatscout) were selected. Sequences detected by SINE-Finder and MITE-Hunter were also directly selected for the “total TEs library”. Finally, the sequences classified as noCat (uncategorized) and obtained with the repetitiveness-based approach of TE detections were selected for the “repeated elements library”.

STEP 5: Annotation

The annotation can be performed by TEannot (Flutre et al., 2011)

https://urgi.versailles.inra.fr/Tools/REPET
Here is the method that we used for the annotation of the *T. lutea* genome:

Three libraries were built a “potentially autonomous TEs library” 2) an “total TEs library” and 3) a “repeated elements library”. A first run of TEannot was performed for each library to known sequences matching with a full-length size on the genome (FLC sequences) and remove potential chimeric data. A second run of TEannot was performed with these FLC sequences for each of the final library and three annotations were obtained.
REFERENCES:


