

Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001

Made the : 26 november 2017

PiRATE:

lfremer

a Pipeline to Retrieve and Annotate TEs



2 December 2017

Wrote by :	Supervised by :
Jérémy BERTHELIER	Grégory CARRIER



Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) <u>https://wwz.ifremer.fr/pba_eng/</u>

File : PBA-A-001	Made the : 26 november 2017	

Ifremer

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	
STEP 5: Annotation	24





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) <u>https://wwz.ifremer.fr/pba_eng/</u>

File : PBA-A-001 Made the : 26 november 2017			
	File : PBA-A-001	Made the : 26 november 2017	

Requirement

- The PiRATE-Galaxy is installed on a virtual machine name PiRATE-VM.
 The PiRATE Virtual Machine (PiRATE-VM) can be download at the following URL
 http://doi.org/10.17882/51795
- To use the PiRATE-VM a virtual machine monitor need to be installed, for example VirtualBox

https://www.virtualbox.org/.

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

File	Machine Help			
<i>~</i>	Preferences	Ctrl+G		
a	Import Appliance	Ctrl+I		Details Unapshots
R	Export Appliance	Ctrl+E		Preview
07 - 5	Virtual Media Manager. Network Operations Ma Check for Updates	Ctrl+D anager	PiRATE Ubuntu (64-bit)	
	Reset All Warnings		96 MB	Max Processor Max Processor Max Processor Image: State S
	Exit	Ctrl+Q	py, Optical, Hard	Laurenteren y Regeleration and an an and an an and an an and an and an and an and an and an a
		Acceleration: VT Pa Pa	-x/AMD-V, Nested ging, KVM ravirtualization	
		🖳 Display		
		Video Memory: Remote Desktop Se Video Capture:	12 MB erver: Disabled Disabled	
		Storage		
		Controller: IDE IDE Secondary Ma Controller: SATA SATA Port 0:	aster: [Optical Drive] \ TEdetection.vdi	Vide i (Normal, 300,00 GB)
		Audio		•
Imp	ort an appliance into Virt	ualBox		





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE)

https://wwz.ifremer.fr/pba_eng/

A-A-001 Made the : 26 november 2017

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 doubleclicking on the items and disable others using the check boxes below.

Description	Configuration
Virtual System 1	
😪 Name	PiRATE_2
Product	a Pipeline to Retrieve and Annotate Transposab
Product-URL	http://doi.org/10.17882/51795
Vendor	Implemented by Jérémy Berthelier and Grégory
Vendor-URL	https://wwz.ifremer.fr/pba/
Version	1.0 (30 November 2017)
🗮 Guest OS Type	🐕 Ubuntu (64-bit)
CPU	3
RAM	10196 MB
💿 DVD	
USB Controller	
뒏 Sound Card	ICH AC97
Network Adapter	Intel PRO/1000 MT Desktop (82540EM)
🛇 Storage Controller (IDE)	PIIX4
🛇 Storage Controller (IDE)	PIIX4
4 🟈 Storage Controller (SATA)	AHCI
🗵 Virtual Disk Image	C:\Users\jberthel.IFR\VirtualBox VMs\PiRATE_2

Reinitialize the MAC address of all network cards

• Made changes according to your computer setting.

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

Oracle VM VirtualBox Ma	inager	PIRATE_1 - Settings
File Marine Help		E General System
New Settions Discard S	-	System Motherboard Processor Acceleration
PRATL 1 Provered Off	Conceral Name: PRATE_1 Concering System: Ubunk (I4-bit) System Base Nemoy: 10106 MB Processor: 3 Boot Onder: Propon, Obtical, Hard Soot Onder: Propon, Obtical, Hard Soot Onder: Propon, Obtical, Hard Conceration: 174, MAPU. Netted Paray: Tualization Video Memory: 12 MB Remote Desktop Server: Dasabled	Display Display Display Disse Memory: 416 101964 Storage Boot Order: V Proppy Post Posts Posts Posts Posts Posts Posts Posts Disset Disse
	Storage	OK Cancel Help
	Controller: IDE IDE Secondary Master: [Optical Drive] Em Controller: SATA SATA Port 0: PRATE-disk1.vmd	y (tormal, 300,00 68)
	🚱 Audio	
	Host Driver: Windows DirectSound	•

• Open the PiRATE-VM and the "Jeremy" account, the password is: jeremy07





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) <u>https://wwz.ifremer.fr/pba_eng/</u>

File : PBA-A-001

Made the : 26 november 2017



2. How to run your PiRATE-Galaxy

Now the PiRATE-Galaxy can be launched.







Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) <u>https://wwz.ifremer.fr/pba_eng/</u>

File : PBA-A-001	Made the : 26 november 201	7	
	jeremy:bash- Fichen [™] Edition Affichage Speets Configuration Aide Jerang/Qitraum/Altisticates¶	Konsole v e e	
		2. Enter: sh /home/jeremy/galaxy/	run.sh
	jeremy: bash i gremy: bash – Konzole	S S S S S S S S S S S S S S S S S S S	
	ieremy : bash	– Konsole	\odot \otimes \otimes
File Edit View S	ignets Settings Help		000
jeremy@jeremy-Virtu	ualBox:~\$ sh /home/jeremy/galax	//run.sh	Â
	💌 🕢 galaxy:sh-K Fichier Édition Affichage Signets Configuration Aide	onsole 🏾 🕲 🕲 🕲	
	Controller galaxy.eeb.framework.base DEBUG 2017-08-23 10:09:41,534 Enabling 'workflow z galaxy.eeb.framework.base DEBUG 2017-08-23 10:09:41,534 Enabling 'genomes' galaxy.eeb.framework.base DEBUG 2017-08-23 10:09:41,534 Enabling historize' galaxy.eeb.framework.base DEBUG 2017-08-23 10:09:41,535 Enabling historize' galaxy.eeb.framework.base DEBUG 2017-08-23 10:09:41,535 Enabling historize' galaxy.eebaps galaxy.buildap DEBUG 2017-08-23 10:09:41,735 Enabling 'http: galaxy.eebaps galaxy.buildap DEBUG 2017-08-23 10:09:41,735 Enabling 'tror galaxy.eebaps galaxy.buildap DEBUG 2017-08-23 10:09:41,735 Enabling 'tror	<pre>nnotations' API controller, class: WorkflowAnnotationsController Pf controller, class: BenomesController API controller, class: PapedevisionsController API controller, class: PhilorizeController sciellandleware sciellandleware 'siddleware' 'siddleware logger middleware wordenbast; moldleware</pre>	
	pinasy, weekspis. galaxy, buildapp terms carries ca 100 or 41,72 stadiuting requires trafpolignics, biological actions, device and the second carries can be applied of representation palaxy weekspis galaxy, buildapp DEBMG 2017-08-23 10:09:41,728 added url, patt /puignis/visial/actions/csg/stafi galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 added url, patt /puignis/visial/actions/csg/stafi galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 added url, patt /ortig/puignis/visial/actions/csg/stafic palaxy, queek, wrker IMF0 2017 db:23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEBMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEMG 2017-08-23 10:09:41,729 binding and starting galaxy galaxy, weekspis, galaxy, buildapp DEMG 2017-08-23 10:09:41,729 binding galaxy galaxy, weekspis, galaxy, buildapp DEMG 2017-08-23 binding galaxy, buildapp DEMG 2017-08-23 binding galaxy, buildapp DEMG 2017-08-23 binding galaxy, buildapp DEMG 2017-08-24,729 binding galaxy, buildapp DEMG 2017,929 30,920,720,720,720,720,720,720,720,720,720,7	5. LD ministemare to static middleware: /plugins/visualizations/charts/static, ./con to static middleware: /plugins/visualizations/csg/static, ./con to static middleware: /plugins/visualizations/graphviz/static, ./c to static middleware: /plugins/visualizations/scatterplot/static, control, worker for main return, Galaxy thewad <_MsinThread(MainThread, started 14067944780	
	pplasy webspos galaxy.buildapp DEB06 2017-08-23 10:09:41,746 Prior to webapp amon 14670712069140) is alive galaxy webspos, galaxy.buildapp DEB06 2017-08-23 10:09:41,746 Prior to webapp started daemon (4467500251664) is alive. galaxy.webspos.galaxy.buildapp DEB06 2017-08-23 10:09:41,746 Prior to webapp semon 1460790135466) is alive. galaxy.webspos.galaxy.buildapp DEB06 2017-08-23 10:09:41,746 Prior to webapp semon 146079017707) is alive. galaxy.webspos.galaxy.buildapp DEB06 2017-08-23 10:09:41,746 Prior to webapp semon 146079017707) is alive.	return, Galaxy thread <thread(localrunner.work_thread-0, d<br="" started="">return, Galaxy thread <thread(jobhandlerstopqueue.monitor_thread, return, Galaxy thread <thread(localrunner.work_thread-4, d<br="" started="">return, Galaxy thread <thread(localrunner.work_thread-3, d<br="" started="">return Galaxy thread <thread(localrunner.work_thread-3, d<="" started="" td=""><td></td></thread(localrunner.work_thread-3,></thread(localrunner.work_thread-3,></thread(localrunner.work_thread-4,></thread(jobhandlerstopqueue.monitor_thread, </thread(localrunner.work_thread-0,>	
	μιανγ καναγρίς, ήμιανγ καιτισμής μεθώς «20/7.48.2.3 (20/97.47.46 PT) 3. 3. μιανγ καναγρίς, ήμιανγ καιτισμής μεθώς «20/7.48.2.3 (20/97.47.46 PT) 3. 4. 3. μιανγ καναγρίς, ήμιανγ καιτισμής μεθώς 20/7.68.23 (20/97.47.46 PT) 3. 4. 4. 4. μιανγ καναγρίς, ήμιανγ καιτισμής μεθώς 20/7.68.23 (20/97.47.46 PT) 3. 4. <td>Right click on the URL</td> <td></td>	Right click on the URL	
	galaxy:sh	E M 2 4 (1) 早 ・ 10:12 (2)	

The PiRATE-Galaxy is alive!

If not... Check that the network setting of your VM is correctly configured.





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001Made the : 26 november 2017			
	File : PBA-A-001	Made the : 26 november 2017	

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/





lfremer

PiRATE tutorial

Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001

Made the : 26 november 2017

4. Started with the PiRATE pipeline





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) <u>https://wwz.ifremer.fr/pba_eng/</u>

File : PBA-A-001 Made th	he : 26 november 2017	

lfremer

STEP 0: Load and prepare your dataset

• Connect you as administrator to the PiRATE-Galaxy

🥑 💽	Galaxy - Mozilla Firefox) () (X)
📃 Gal y	× +	
€ € 127.0.0.1:8080	マ C Q Search ☆ 自 マ ◆ 合 タ	
& Galaxy	Analyze Data Workflow Shared Data - Visualization - Help - User - 🖬 Using () bytes
Tools	History	<i>C</i> ¢
search tools	Welcome to the PiRATE-Galaxy!!	8
C+1D+1-	Unnamed bistory	
	Login	
	Usernam Email Address:	
	administrator	
	Password:	
	•••••	
	Forgot password? Reset here	
	Login	

Username: administrator **Password:** administrator





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

- 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?

& Galaxy	
Tools	\rightarrow \pm
search tools	8
Get 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".

			lf	remer	PBA
	PiRATE	tutorial			
Physiology a	nd Biotechnolog	y of Algae Laboratoty (PBA)) – IFREMER	Nantes (FRANCE)	
https://wwz.if	remer.fr/pba_eng	<u>1/</u>			
File : PBA-A	-001	Made the : 26 november 2017	7		
rch ita ani iubi iubi rt F Alic E 1.1: 2.1: 2.2: 3: A oww vori	Download from web or Regular Composite	upload from disk	es here		A A B B B B B B B B B B B B B B B B B B
	Type (set all):	Auto-detect	Genome (set all):	unspecified (?) v	

However, your Illumina raw data should have a weight of more than 1 Go. Thus it will be necessary to import it with FTP:

□ Choose local file 🕞 Choose FTP file

 You need to copy-past your files in the directory "administrator@pba.fr" of the PiRATE-VM /home/Jeremy/Documents/administrator@pba.fr

Paste/Fetch data Pause Reset Start Close





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

: PBA-A-001	Made the : 26 november 2017	
	administrator@pba.fr - Dolphin ils	
 Starty Starty Périphériques Z89.6 GIB Hard Drive Jeremy_PBA_JFREMER 	NANTES	administrator@pba.fr Type: folder Size: 2 klements Tags: AddTags_ Rating: CCCCCC Comment: AddComment_
	2 fichiers (63.6 MiB)	

2) You can now load them into the PiRATE-Galaxy

Regular	<u>Composite</u>								
			You a	dded 2 file(s) to the queue	. Add more files (pr click 'Start' to p	roceed.		
	Name		Size	Туре	Gei	nome	Settings	Status	
ß	1.fq	FTP fi	les	•••••			~ 0	0%	⑪
ß	2.fq	This (serve	Galaxy server al er at 134.246.55 .	lows you to upload files 42 using your Galaxy c	via FTP. To up redentials (ema	load some files, ail address and	log in to the FTP password).	0%	⑪
		Avail	Name		Size	Created	2 files 🖽 63.6 MB		
		¥	1.fq			11/29/2017	02:09:46 PM		
		Ľ	2.fq			11/29/2017	02:09:46 PM		
	Type (set all):) v	





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

 File : PBA-A-001
 Made the : 26 november 2017

1) <u>The genome assembly (FASTA)</u>

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)

>Chromomose1 AAATTTTAAAATTTGGGCCCAAAACCCCAAACCCCAAACCCCAAACCCCAATTTTTAA ... >Chromosome2 AAATTTTAAAATTTGGGCCCAAAACCCCAAACCCCAAACCCCAACCCCAATTTTTAA

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





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

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 AAATTTTAAAATTTGGGCCCAAAACCCCAAACCCCAAACCCCAACCCCAATTTTTAA

>readname/2 AAATTTTAAAATTTGGGCCCAAAACCCCAAACCCCAAACCCCAACCCCAATTTTTAA

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!



Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) <u>https://wwz.ifremer.fr/pba_eng/</u>

File : PBA-A-001 Ma	lade the : 26 november 2017	

lfremer

STEP1: Detection of putative TEs

Sealaxy 😪	Analyze Data	Workflow	Shared Data -	Visualization 🗸	Help∓ User∓ 🚦	
Tools search tools	✓ Wel	come	to the Pi	RATE-Ga	laxy!!	
<u>Get Data</u> <u>Text Manipulation</u> Join, Subtract and Group						
Filter and Sort Convert Formats	0: Input data		Genome		Illumina raw data	
PIRATE	1.1: Detection Similitary-based		Structural-based	Repetitiveness-base	Build repeated elements	
STEP 1.1: Detection STEP 1.2: Clustering STEP 2.1: Classification	Repeat TE- Masker HMME Nucl HMM databank databar	R Hunter SINE Finder	Hel LTR Search harvest MGE Scan	TEdenovo Repeat Scout	dna PipeTE RepArk Repeat Explorer	
STEP 2.2: Manual validation	>500 bp			$\overline{\mathcal{I}}$	>500 bp	
Workflows All workflows	1.2: Clustering 2.1: Classification		CD-HIT-e PASTEC Nucl, da	prot, HMM atabanks	PIRATE	
	2.2: Manual Check	Autonom	MCL	ious TEs Uncategoriz	red Keys:	
<	3: Annotation	E	TEannot	ibraries x2	Tools Filter	

The detection step of PiRATE is flexible, you can use every tools one after one or only select your favorite ones.





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001 Made the : 26 november 2017				
	I	File : PBA-A-001	Made the : 26 november 2017	

Here is the list of the available tools with their authors and the URL:

Approach 1: Similarity-based detection

• RepeatMasker (Smit, A. F., Hubley, R., & Green, P. (1996).) www.repeatmasker.org/

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.

Approche 2: Structural-base detection

• LTRharvest (Ellinghaus et al., 2008) <u>http://www.zbh.uni-hamburg.de/?id=206</u>

This tool detects LTR from a genome assembly.

• MGEScan non-LTR (Rho and Tang, 2009) http://mgescan.readthedocs.io/en/latest/nonltr.html

This tool detects LINE from a genome assembly.

• Helsearch (Yang and Bennetzen, 2009) <u>http://omictools.com/helsearch-tool</u>

This tool detects Helitron from a genome assembly.

• MITE-Hunter (Han and Wessler, 2010) <u>http://target.iplantcollaborative.org/mite_hunter.html</u>

This tool detects MITE from a genome assembly.

• SINEfinder (Wenke et al., 2011) <u>http://www.plantcell.org/content/23/9/3117</u>

This tool detects SINE from a genome assembly.





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File · PBA-A-001 Made the · 26 november 2017			
	File : PBA-A-001	Made the : 26 november 2017	

Approche 3: Repetitiveness-base detection

• TEdenovo (Flutre et al., 2011) <u>https://urgi.versailles.inra.fr/Tools/REPET</u>

This tool belongs to the REPET pakage, it allow 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) <u>https://bix.ucsd.edu/repeatscout/</u>

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) <u>http://repeatexplorer.umbr.cas.cz/</u>

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) <u>https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html</u>

This tool samples reads and assemble repeated elements with Trinity.

• **RepARK** (Koch et al., 2014) <u>https://github.com/PhKoch/RepARK</u>

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.





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

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:

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.



Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

lfremer

STEP 2: Sort your detected sequences

& Galaxy	Analyze Data	Workflow	Shared Data -	Visualization ~	Help - User - ∎
Tools search tools	✓ We	lcome	to the Pi	RATE-Ga	laxy!!
<u>Get Data</u> <u>Text Manipulation</u>					
Join, Subtract and Group Filter and Sort Convert Formats Fetch Alignments	0: Input data		Genome assembly		Illumina raw data
PIRATE STEP 1.1: Detection STEP 1.2: Clustering	Similitary-base Repeat Masker Nucl HMM databank datab	ed Since Sin	itructural-based Hel LTR Search harvest	Repetitiveness-base TEdenovo Repeat	d Build repeated elements dna PipeTE Repeat
STEP 2.2: Manual validation STEP 3: Annotation Workflows	>500 bp 1.2: Clustering	Finder	CD-HIT-e	est	>500 bp
<u>All workflows</u>	2.1: Classification	Autonom	PASTEC Nucl, da pus TES Non-autonom	prot, HMM atabanks ous TEs Uncategoriz	PIRATE
	2.2: Manual Cheo	k B	MCL	Repeated elements	Keys: Library Manual
<	3: Annotation		TEannot	braries x2	Filter

• 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.





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

• 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) <u>http://weizhongli-lab.org/cd-hit/</u>. 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= Sa/S=1 and c=%identity=1



Figure from https://github.com/weizhongli/cdhit/wiki/3.-User's-Guide#CDHITEST



Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

lfremer

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.





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

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

⊗ Galaxy	Analyze Data	Workflow	Shared Data -	Visualization -	Help▼ User▼	
Tools		como	to the Di	DATE CO	lovull	Â
Get Data Text Manipulation Join. Subtract and Group		come		RAIE-Ga		
Filter and Sort Convert Formats Fetch Alignments	0: Input data 1.1: Detection		Genome assembly		Illumina raw data	
PIRATE <u>STEP 1.1: Detection</u> <u>STEP 1.2: Clustering</u> <u>STEP 2.1: Classification</u> <u>STEP 2.2: Manual validation</u> <u>STEP 3: Annotation</u>	Similitary-base Repeat Masker Nucl databank >500 bp 1.2: Clustering	MITE R Hunter SINE Finder	Structural-based Hel LTR Search harvest MGE Scan CD-HIT-e	Repetitiveness-based TEdenovo Repeat Scout	Build repeated elements dna PipeTE RepArk Explorer >500 bp	
Norkflows • <u>All workflows</u>	2.1: Classification	Autonom	PASTEC Nucl, da ous TES Non-autonom	prot, HMM atabanks ous TEs Uncategorize	PIRATE	
	2.2: Manual Chec	k a	MCL	Repeated elements	Keys: Library Manual	
<	3: Annotation		TEannot	braries x2	Filter	~ >





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

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.



Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

$File \cdot PBA_A_001$ Made the $\cdot 26$ povember 2017	

lfremer

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





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

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.





Physiology and Biotechnology of Algae Laboratoty (PBA) – IFREMER Nantes (FRANCE) https://wwz.ifremer.fr/pba_eng/

File : PBA-A-001	Made the : 26 november 2017	

REFERENCES:

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410.

Eddy, S.R., and others (1995). Multiple alignment using hidden Markov models. Ismb 3, 114–120.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. PLoS ONE *6*, e16526.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., and Taylor, J. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Research *15*, 1451–1455.

Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., and Boulesteix, M. (2015). De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). Genome Biology and Evolution *7*, 1192–1205.

Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Research *38*, e199–e199.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: An Automatic Transposable Element Classification Tool. PLoS ONE *9*, e91929.

Koch, P., Platzer, M., and Downie, B.R. (2014). RepARK--de novo creation of repeat libraries from wholegenome NGS reads. Nucleic Acids Research *42*, e80–e80.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics *29*, 792–793.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21*, i351–i358.

Rho, M., and Tang, H. (2009). MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Research *37*, e143–e143.

Smit, A. F., Hubley, R., & Green, P. (1996). RepeatMasker.

Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B., and Schmidt, T. (2011). Targeted Identification of Short Interspersed Nuclear Element Families Shows Their Widespread Existence and Extreme Heterogeneity in Plant Genomes. THE PLANT CELL ONLINE *23*, 3117–3128.

Yang, L., and Bennetzen, J.L. (2009). Structure-based discovery and description of plant and animal Helitrons. Proceedings of the National Academy of Sciences *106*, 12832–12837.