

PIRATE: a Pipeline to Retrieve and Annotate TEs of non-model organisms



Jérémy BERTHELIER, Ph.D. student

N. Casse, N. Daccord, V. Jamilloux, B. Saint-Jean, G. Carrier

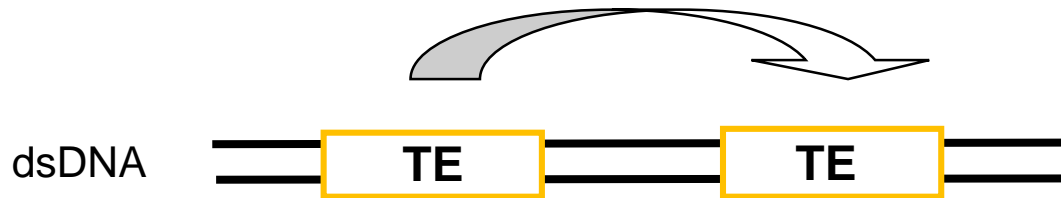
Transposable elements (TEs)



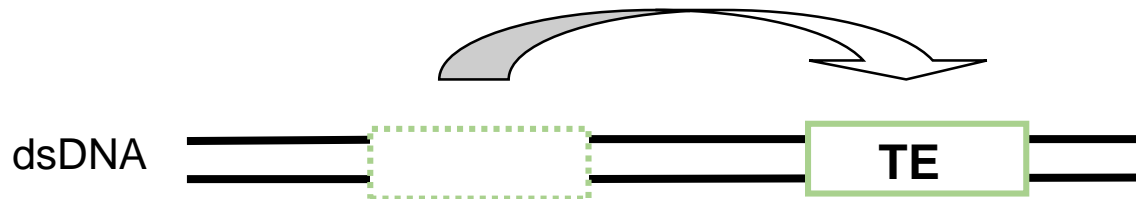
TEs are DNA sequences able to move (= transposition) into the host genome of eucaryotic and procaryotic organisms

Two classes:

Classe I named retrotransposons move by a “copy-paste” mechanism



Classe II named DNA transposon move by a “cut-paste” mechanism



A huge diversity

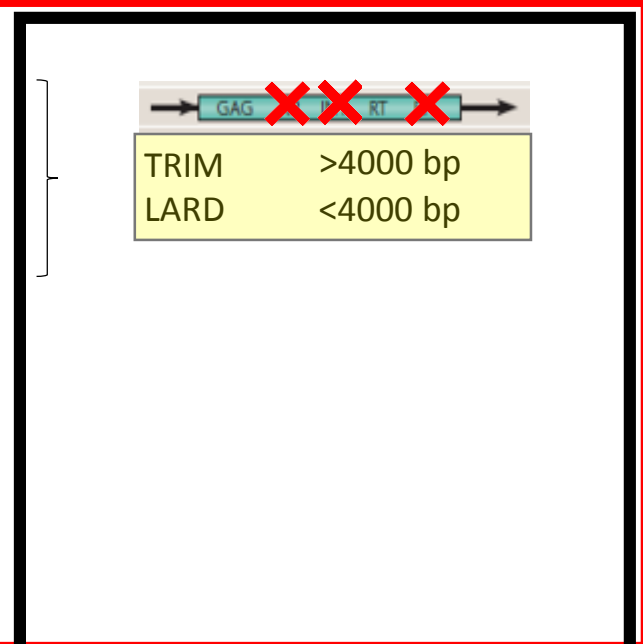
Wicker et al., 2007

Autonomous TEs

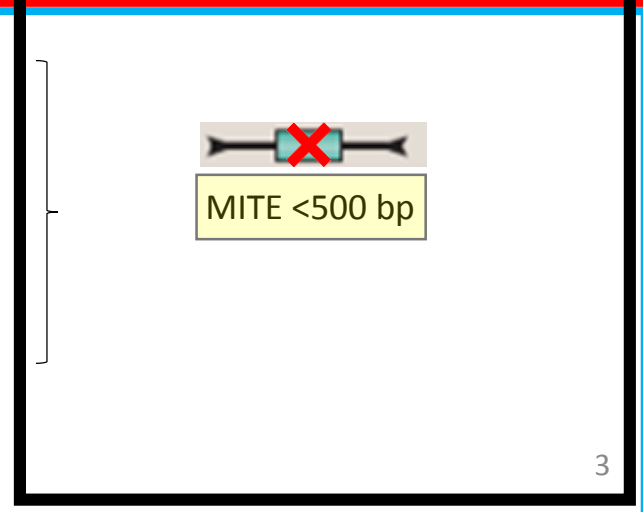
Non-autonomous TEs

Classification		Structure
Order	Superfamily	
Class I (retrotransposons)		
LTR	Copia	GAG AP INT RT RH
	Gypsy	GAG AP RT RH INT
	Bel-Pao	GAG AP RT RH INT
	Retrovirus	GAG AP RT RH INT ENV
	ERV	GAG AP RT RH INT ENV
DIRS	DIRS	GAG AP RT RH YR
	Ngaro	GAG AP RT RH YR
	VIPER	GAG AP RT RH YR
PLE	Penelope	RT EN
LINE	R2	RT EN
	RTE	APE RT
	Jockey	ORF1 APE RT
	L1	ORF1 APE RT
	I	ORF1 APE RT RH
SINE	tRNA	
	7SL	

(Non-autonomes)

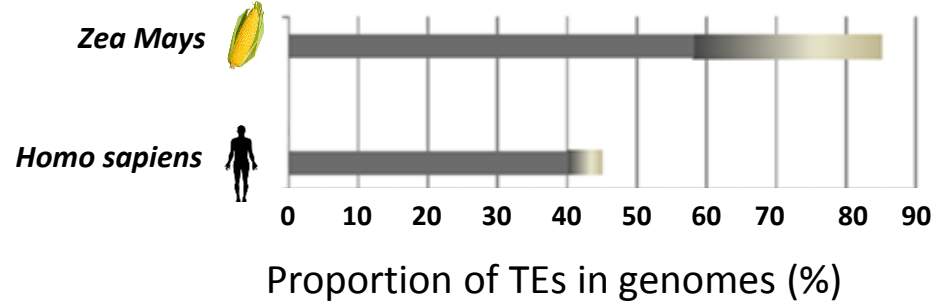


Class II (DNA transposons) - Subclass 1		
TIR	Tc1-Mariner	Tase*
	hAT	Tase*
	Mutator	Tase*
	Merlin	Tase*
	Transib	Tase*
	P	Tase
	PiggyBac	Tase
	PIF-Harbinger	Tase* ORF2
	CACTA	Tase ORF2
Crypton	Crypton	YR
Class II (DNA transposons) - Subclass 2		
Helitron	Helitron	RPA Y2 HEL
Maverick	Maverick	C-INT ATP CYP POL B



Impact of TEs on genomes

TEs impact genome size



Chénais et al., 2012

TEs can generate phenotype changes



Vitis vinifera

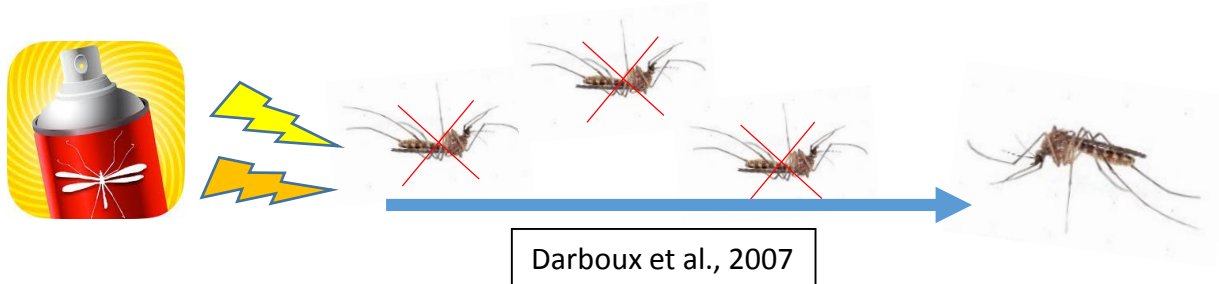
Kobayashi et al., 2004



Citrus sinensis

Butelli et al., 2012

TE mutations can promote species adaptation



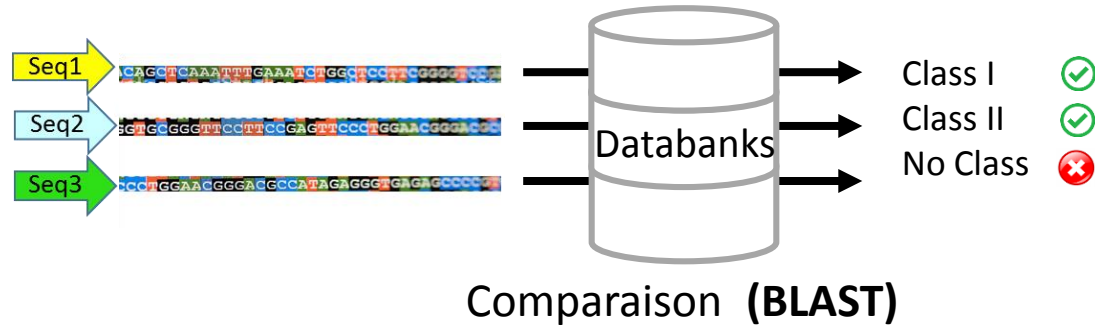
Darboux et al., 2007

How to conduct a de novo TE annotation

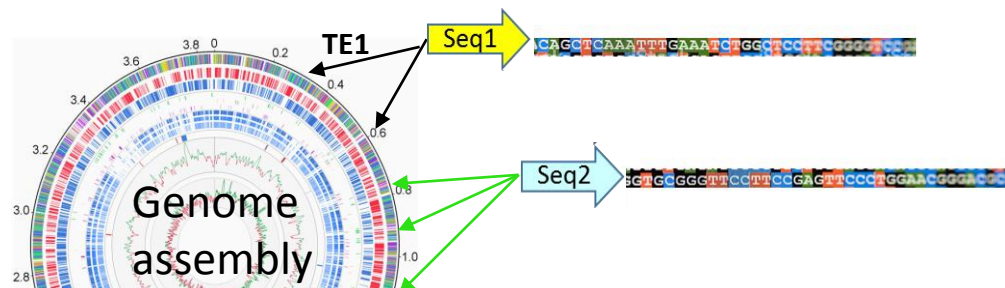
1. Detection of TEs



2. Classification of detected TEs



3. TE annotation



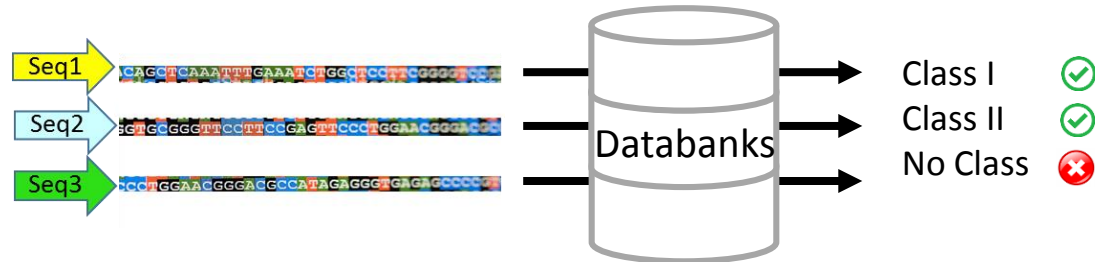
Limitations with non-model organisms

1. Detection of TEs



Genome assembly is usually fragmented
→ missassembly of repeated elements

2. Classification of detected TEs

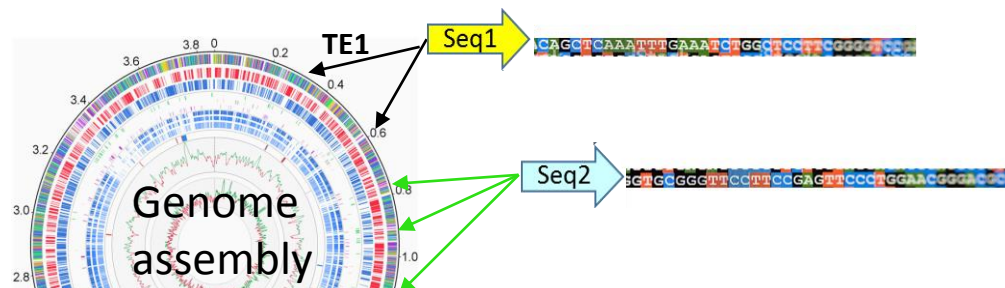


Comparison (BLAST)



Few knowledge in TE databanks
→ difficulty to classify/recognize TEs

3. TE annotation



Limitations with non-model organisms

1. Detection of TEs

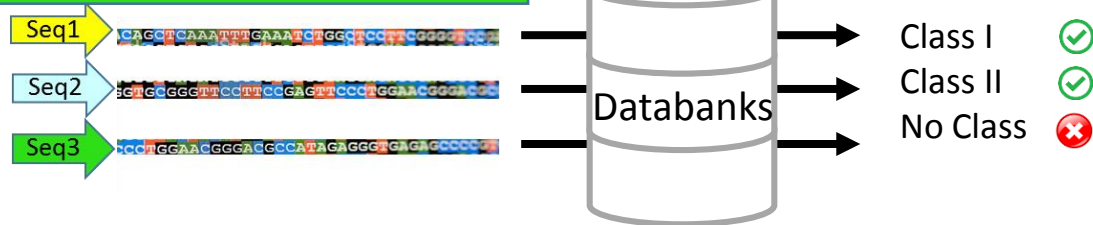


Genome assembly is usually fragmented
→ missassembly of repeated elements



Improve the detection step

2. Classification of detected TEs



Comparison (BLAST)

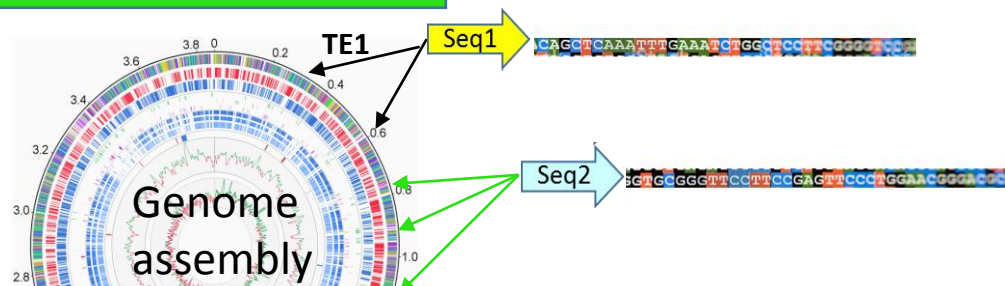


Few knowledge in TE databanks
→ difficulty to classify/recognize TEs



Improve the classification step

3. TE annotation



PIRATE-Galaxy distribution

→ PiRATE is implemented into a stand-alone Galaxy:

The screenshot displays the Galaxy web interface for the PiRATE-Galaxy distribution. The browser window title is "Galaxy - Mozilla Firefox" and the address bar shows "127.0.0.1:8080". The interface includes a top navigation bar with "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User" menus, and a user status of "Using 1.8 GB".

The main content area features a green banner with a checkmark and the text "Welcome to the PiRATE-Galaxy!!". Below this is a detailed workflow diagram:

- 0: Input data**: Genome assembly and Illumina raw data.
- 1.1: Detection**:
 - Similarity-based**: Repeat Masker, TE-HMMER, Nucl databank, HMM databank.
 - Structural-based**: MITE Hunter, Hel Search, LTR harvest, SINE Finder, MGE Scan.
 - Repetitiveness-based**: TEde novo, Repeat Scout.
 - Build repeated elements**: dna PipeTE, Repeat Explorer.
- 1.2: Clustering**: CD-HIT-est (receiving input from >500 bp).
- 2.1: Classification**: PASTEC (Nucl, prot, HMM databanks) leading to Autonomous TEs, Non-autonomous TEs, and Uncategorized.
- 2.2: Manual check**: MCL, BLASTn, and Repeated elements.
- 3: Annotation**: TEannot and Libraries (x2).

A "PIRATE" logo is visible in the bottom right of the workflow diagram. A "Keys" legend indicates: Tool (blue), Library (green), Filter (yellow), and Manual (orange).

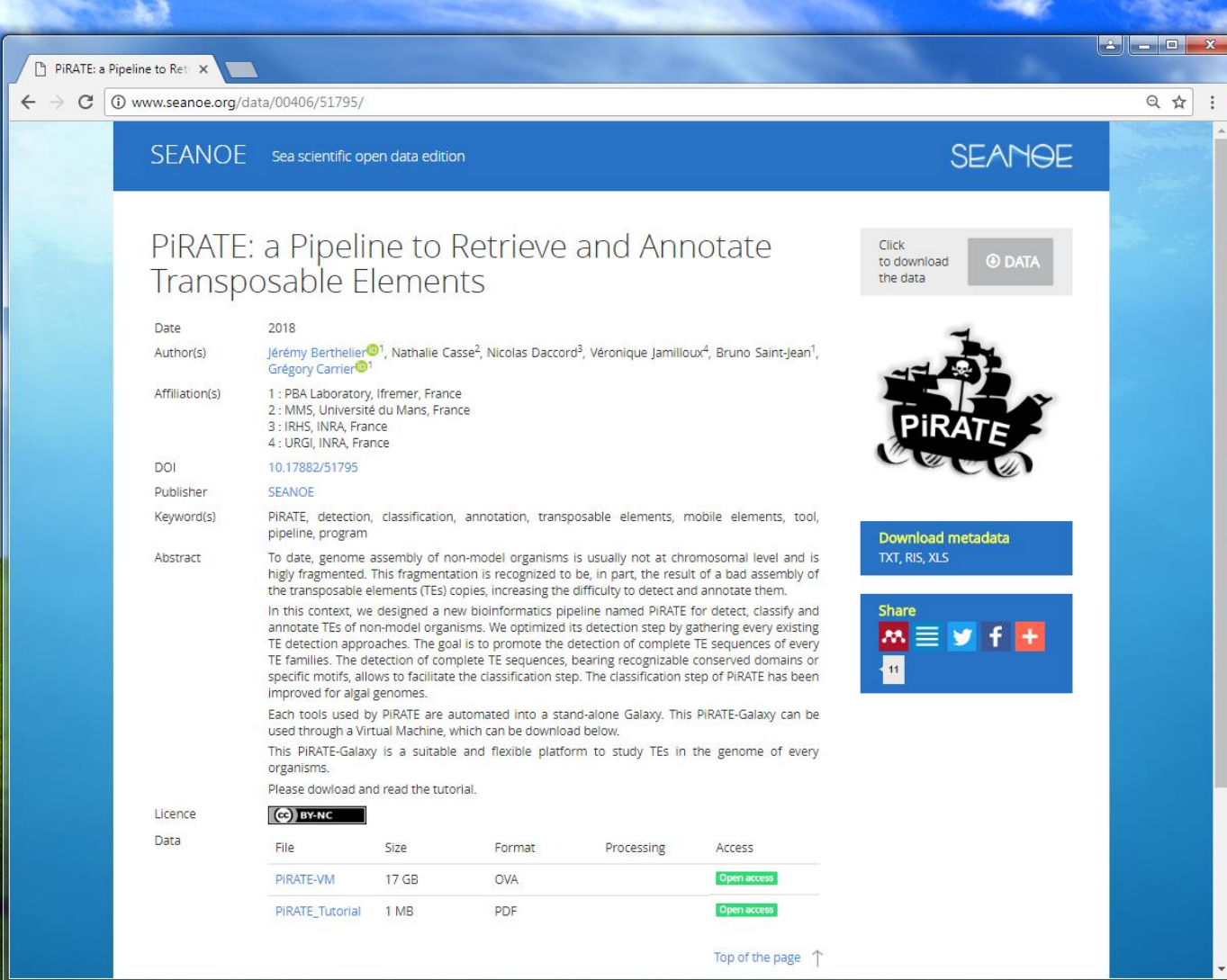
The right sidebar shows a "History" panel with a search bar and a list of datasets:

- 14: RepeatScout log
- 13: RepeatScout output
- 12: log
- 11: Helsearch output
- 9: MGEScan-nonLTR output
- 6: LTRharvest output
- 2: Rename headers
- 1: symbB.v1.0.genome.fa

The left sidebar contains a "Tools" section with a search bar and categories: "Get Data", "Text Manipulation", "Join, Subtract and Group", "Filter and Sort", "Convert Formats", "Fetch Alignments", "PIRATE", and "Workflows". Under "PIRATE", the following steps are listed: STEP 1.1: Detection, STEP 1.2: Clustering, STEP 2.1: Classification, STEP 2.2: Manual check, and STEP 3: Annotation. Under "Workflows", there is a link for "All workflows".

PiRATE-Galaxy distribution

1. Download the PiRATE Virtual Machine (Virtual Linux)




The screenshot shows a web browser window displaying the SEANOE website. The page title is "PiRATE: a Pipeline to Retrieve and Annotate Transposable Elements". The browser address bar shows the URL "www.seanoe.org/data/00406/51795/". The page features a blue header with the SEANOE logo and navigation links. The main content area includes a metadata section with fields for Date, Author(s), Affiliation(s), DOI, Publisher, and Keyword(s). An abstract section provides a detailed description of the PiRATE pipeline. A "Download metadata" button is visible, along with social media sharing options. At the bottom, a table lists downloadable files: "PIRATE-VM" (17 GB, OVA format) and "PIRATE_Tutorial" (1 MB, PDF format), both with "Open access" links.

SEANOE Sea scientific open data edition

SEANOE

PiRATE: a Pipeline to Retrieve and Annotate Transposable Elements

Click to download the data **DATA**



Download metadata
TXT, RIS, XLS

Share

Abstract


To date, genome assembly of non-model organisms is usually not at chromosomal level and is highly fragmented. This fragmentation is recognized to be, in part, the result of a bad assembly of the transposable elements (TEs) copies, increasing the difficulty to detect and annotate them.

In this context, we designed a new bioinformatics pipeline named PiRATE for detect, classify and annotate TEs of non-model organisms. We optimized its detection step by gathering every existing TE detection approaches. The goal is to promote the detection of complete TE sequences of every TE families. The detection of complete TE sequences, bearing recognizable conserved domains or specific motifs, allows to facilitate the classification step. The classification step of PiRATE has been improved for algal genomes.

Each tools used by PiRATE are automated into a stand-alone Galaxy. This PiRATE-Galaxy can be used through a Virtual Machine, which can be download below.

This PiRATE-Galaxy is a suitable and flexible platform to study TEs in the genome of every organisms.

Please download and read the tutorial.

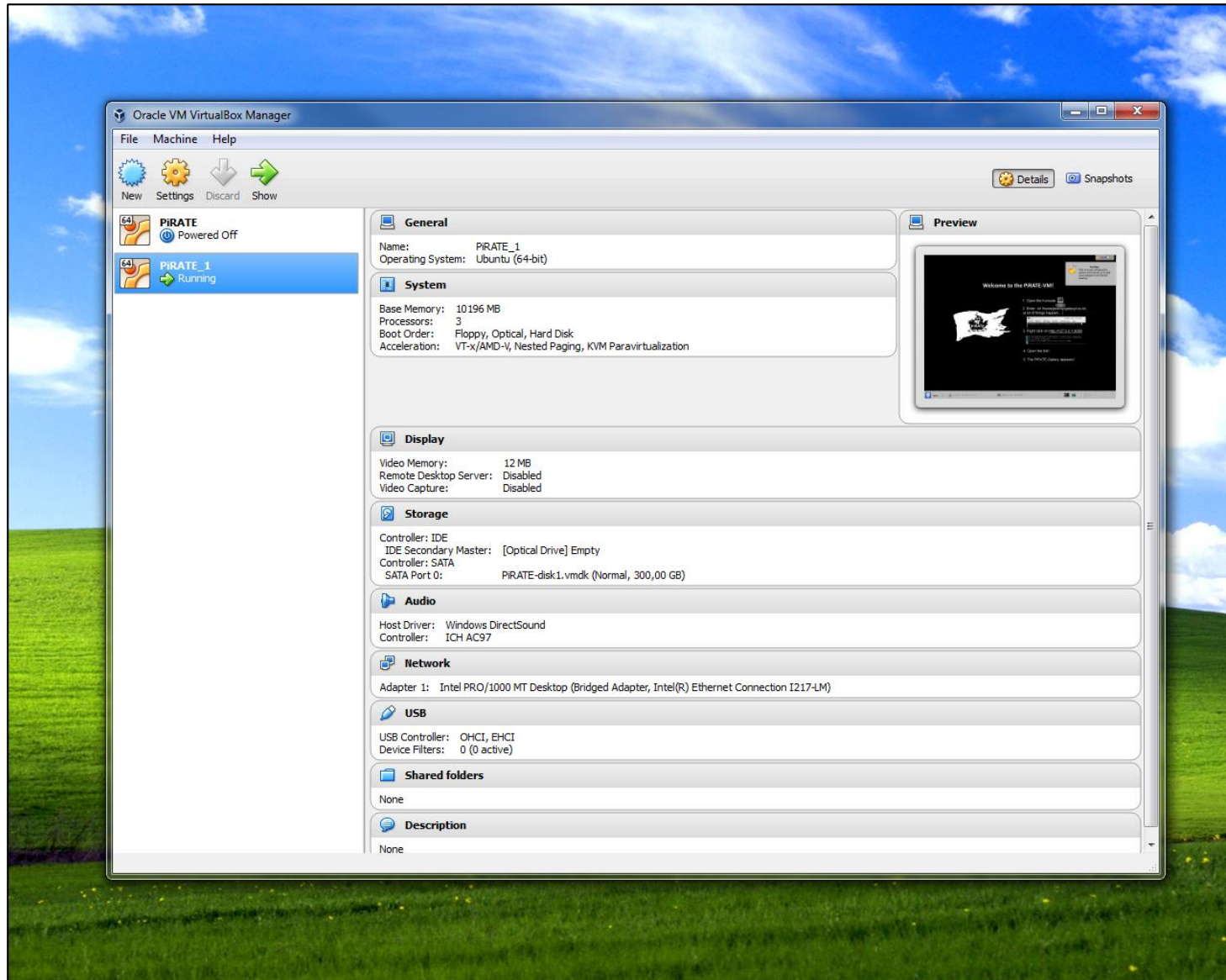
Licence 

Data	File	Size	Format	Processing	Access
	PIRATE-VM	17 GB	OVA		Open access
	PIRATE_Tutorial	1 MB	PDF		Open access

Top of the page ↑

PIRATE-Galaxy distribution


2. Import the PIRATE Virtual Machine into a Virtual Machine Monitor (VirtualBox)




PIRATE-Galaxy distribution

3. Launch the PiRATE-Galaxy

Welcome to the PiRATE-VM!



1. Open the Konsole 
2. Enter: `sh /home/jeremy/galaxy/run.sh`
(a lot of things happen...)
3. Right click on <http://127.0.0.1:8080>
4. Open the link!
5. The PiRATE-Galaxy appears!

```
jeremy@jeremy-VirtualBox:~$ sh /home/jeremy/galaxy/run.sh
```

```
Starting server in PID 3077.
Serving on 0.0.0.0:8080 view at http://127.0.0.1:8080
```

Galaxy - Mozilla Firefox galaxy: sh - Konsole 15:49

PIRATE-Galaxy distribution

It's ready to use !!!

The screenshot displays the Galaxy web interface in a Mozilla Firefox browser window. The address bar shows the URL 127.0.0.1:8080. The Galaxy logo and navigation menu are visible at the top. The main content area features a large green banner with a checkmark and the text "Welcome to the PiRATE-Galaxy!!". Below the banner is a detailed workflow diagram for PiRATE. The workflow starts with "0: Input data" (Genome assembly and Illumina raw data) and proceeds through several steps: "1.1: Detection" (divided into Similarity-based, Structural-based, and Repetitiveness-based methods), "1.2: Clustering" (using CD-HIT-est), "2.1: Classification" (using PASTEC), "2.2: Manual check" (using MCL and BLASTn), and "3: Annotation" (using TEannot and Libraries). The diagram also includes a "Build repeated elements" section with tools like dna PipeTE and RepArk. A "PIRATE" logo is present in the bottom right of the workflow diagram. The left sidebar contains a "Tools" section with a search bar and a list of categories: "Get Data", "Text Manipulation", "Join, Subtract and Group", "Filter and Sort", "Convert Formats", "Fetch Alignments", "PIRATE", and "Workflows". The right sidebar shows a "History" section with a search bar and a list of datasets, including "14: RepeatScout_log", "13: RepeatScout output", "12: log", "11: Helsearch output", "9: MGEScan-nonLTR output", "6: LTRharvest output", "2: Rename headers", and "1: symbB.v1.0.genome.fa". The bottom status bar shows the system tray with icons for network, volume, and time (15:49).

PiRATE-Galaxy: Detection

Tools

search tools

Get Data

Text Manipulation

Join, Subtract and Group

Filter and Sort

Convert Formats

Fetch Alignments

PIRATE

STEP 1.1: Detection

STEP 1.2: Clustering

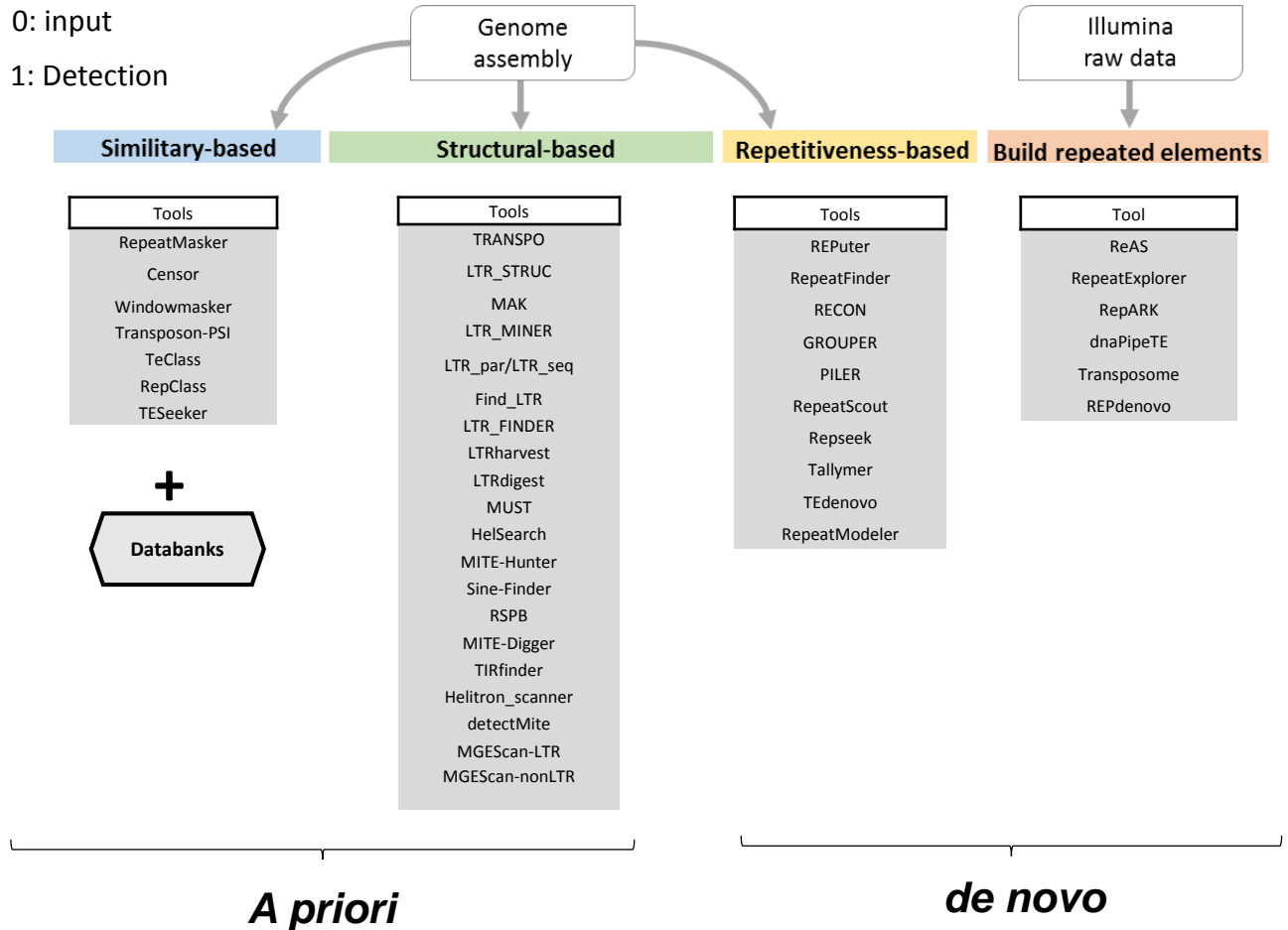
STEP 2.1: Classification

STEP 2.2: Manual check

STEP 3: Annotation

Workflows

- All workflows



PiRATE-Galaxy: Detection

Tools

search tools

Get Data

Text Manipulation

Join, Subtract and Group

Filter and Sort

Convert Formats

Fetch Alignments

PIRATE

STEP 1.1: Detection

STEP 1.2: Clustering

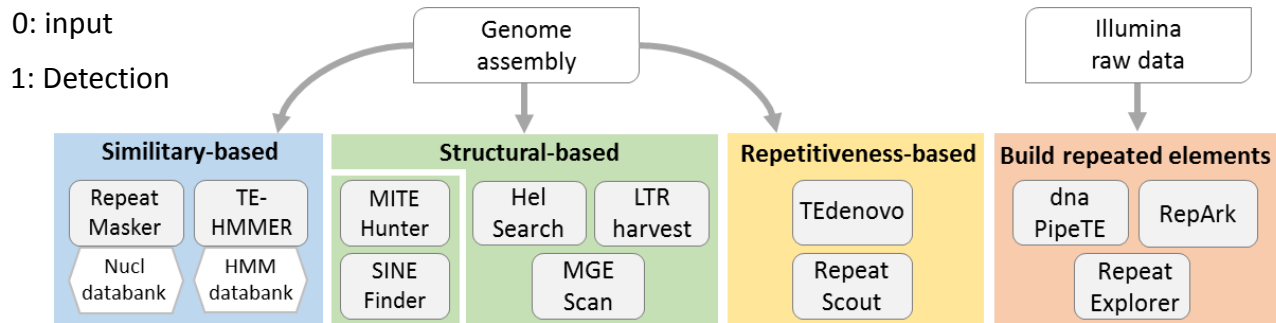
STEP 2.1: Classification

STEP 2.2: Manual check

STEP 3: Annotation

Workflows

- All workflows



Group every existing TE detection approach to:

- promote the detection of every TE families
- promote the detection of full-length TEs

→ To facilitate the classification step

PiRATE-Galaxy: Detection

Tools

search tools

Get Data
Text Manipulation
Join, Subtract and Group
Filter and Sort
Convert Formats
Fetch Alignments

PIRATE

STEP 1.1: Detection

BUILD REPEATED ELEMENTS

- [RepeatExplorer](#) De novo detection of repeated sequences
- [dnaPipeTE](#) De novo detection of repeated sequences
- [RepARK](#) De novo detection of repeated sequences

REPETITIVENESS-BASED

- [TEdenovo \(REPET\)](#) De novo detection of repeated sequences
- [RepeatScout](#) De novo detection of repeated sequences

STRUCTURAL-BASED

- [Helsearch](#) Detection of Helitrons
- [MITE-Hunter](#) Detection of MITES
- [SINE-Finder](#) Detection of SINEs
- [LTRharvest](#) Detection of LTRs
- [MGEScan-nonLTR](#) Detection of non-LTRs

SIMILARITY-BASED

- [TE-HMMER](#) Detection of TEs from profile HMMs
- [RepeatMasker](#) Detection of TEs using a nucleotide databank

STEP 1.2: Clustering

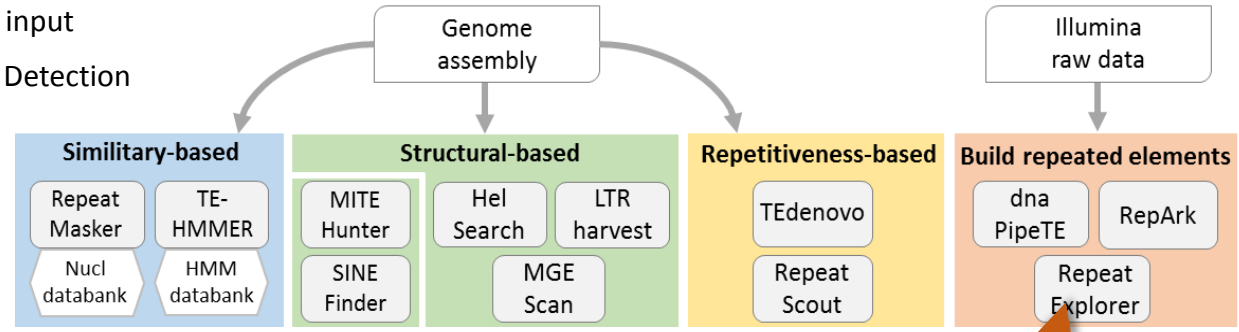
STEP 2.1: Classification

STEP 2.2: Manual check

STEP 3: Annotation

0: input

1: Detection



Build repeated sequences

PiRATE-Galaxy: Detection

Tools

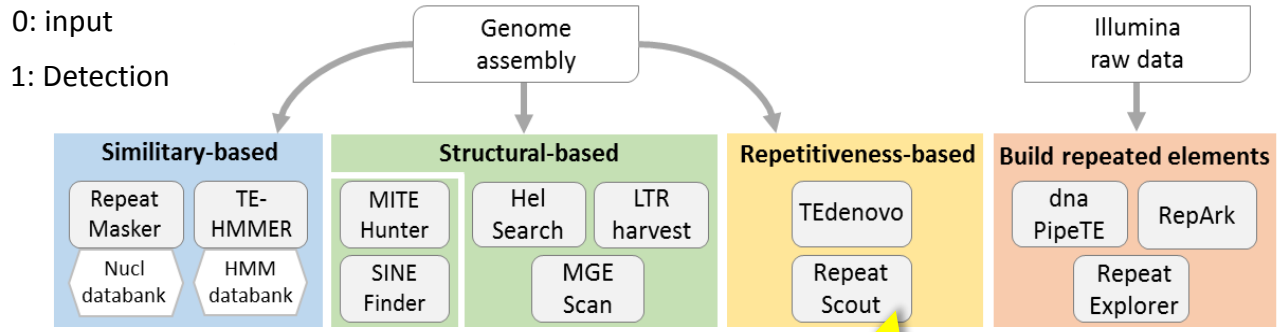
- search tools
- Get Data
- Text Manipulation
- Join, Subtract and Group
- Filter and Sort
- Convert Formats
- Fetch Alignments

PIRATE

STEP 1.1: Detection

- BUILD REPEATED ELEMENTS**
 - [RepeatExplorer](#) De novo detection of repeated sequences
 - [dnaPipeTE](#) De novo detection of repeated sequences
 - [RepARK](#) De novo detection of repeated sequences
- REPETITIVENESS-BASED**
 - [TEdenovo \(REPET\)](#) De novo detection of repeated sequences
 - [RepeatScout](#) De novo detection of repeated sequences
- STRUCTURAL-BASED**
 - [Helsearch](#) Detection of Helitrons
 - [MITE-Hunter](#) Detection of MITES
 - [SINE-Finder](#) Detection of SINEs
 - [LTRharvest](#) Detection of LTRs
 - [MGEScan-nonLTR](#) Detection of non-LTRs
- SIMILARITY-BASED**
 - [TE-HMMER](#) Detection of TEs from profile HMMs
 - [RepeatMasker](#) Detection of TEs using a nucleotide databank

- STEP 1.2: Clustering
- STEP 2.1: Classification
- STEP 2.2: Manual check
- STEP 3: Annotation

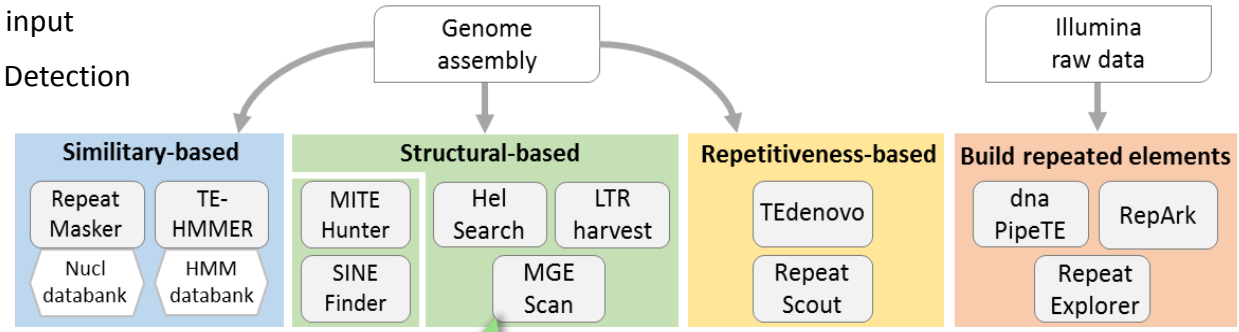


Detect repeated sequences

PiRATE-Galaxy: Detection

0: input

1: Detection



Detect specific structure of TEs

PiRATE-Galaxy: Detection

Tools

- search tools
- Get Data
- Text Manipulation
- Join, Subtract and Group
- Filter and Sort
- Convert Formats
- Fetch Alignments

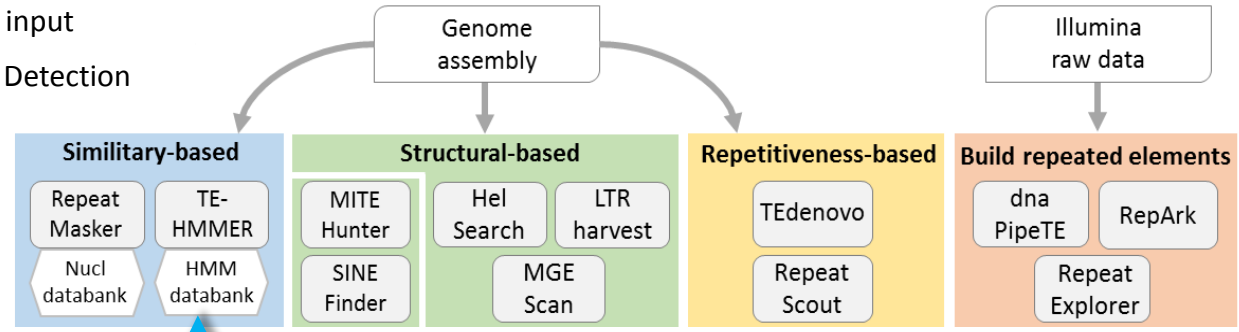
PIRATE

STEP 1.1: Detection

- BUILD REPEATED ELEMENTS**
 - [RepeatExplorer](#) De novo detection of repeated sequences
 - [dnaPipeTE](#) De novo detection of repeated sequences
 - [RepARK](#) De novo detection of repeated sequences
- REPETITIVENESS-BASED**
 - [TEdenovo \(REPET\)](#) De novo detection of repeated sequences
 - [RepeatScout](#) De novo detection of repeated sequences
- STRUCTURAL-BASED**
 - [Helsearch](#) Detection of Helitrons
 - [MITE-Hunter](#) Detection of MITES
 - [SINE-Finder](#) Detection of SINEs
 - [LTRharvest](#) Detection of LTRs
 - [MGEScan-nonLTR](#) Detection of non-LTRs
- SIMILARITY-BASED**
 - [TE-HMMER](#) Detection of TEs from profile HMMs
 - [RepeatMasker](#) Detection of TEs using a nucleotide databank

0: input

1: Detection

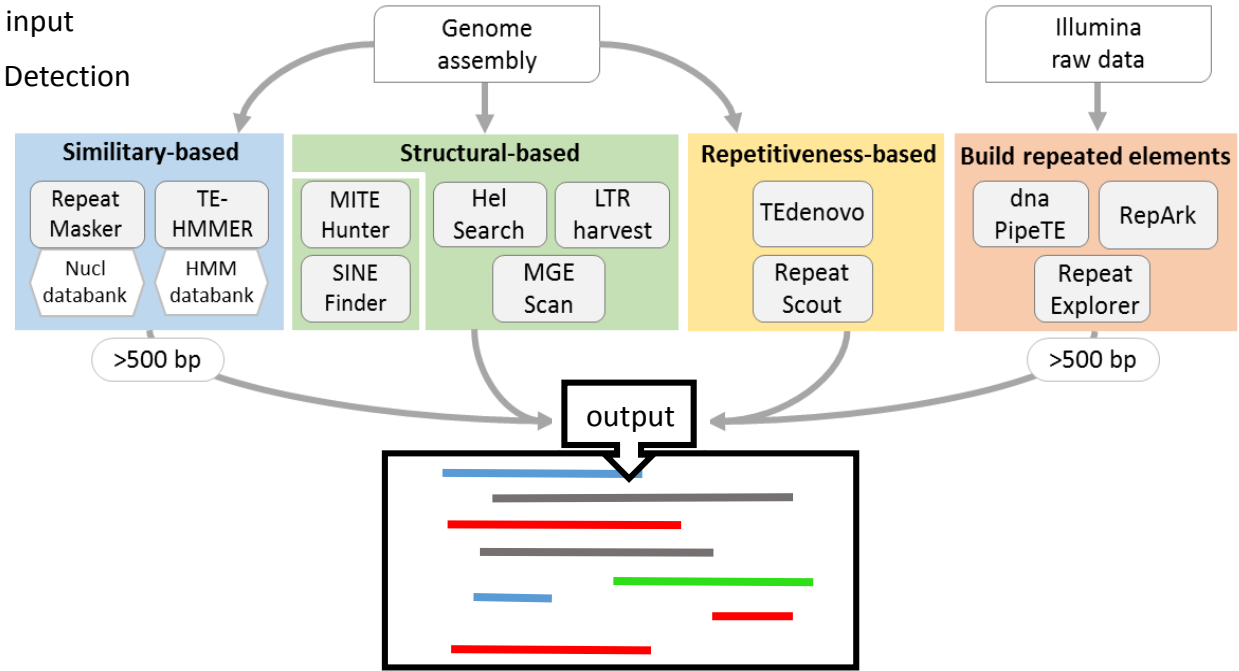


Detect similare sequences

PiRATE-Galaxy: Detection

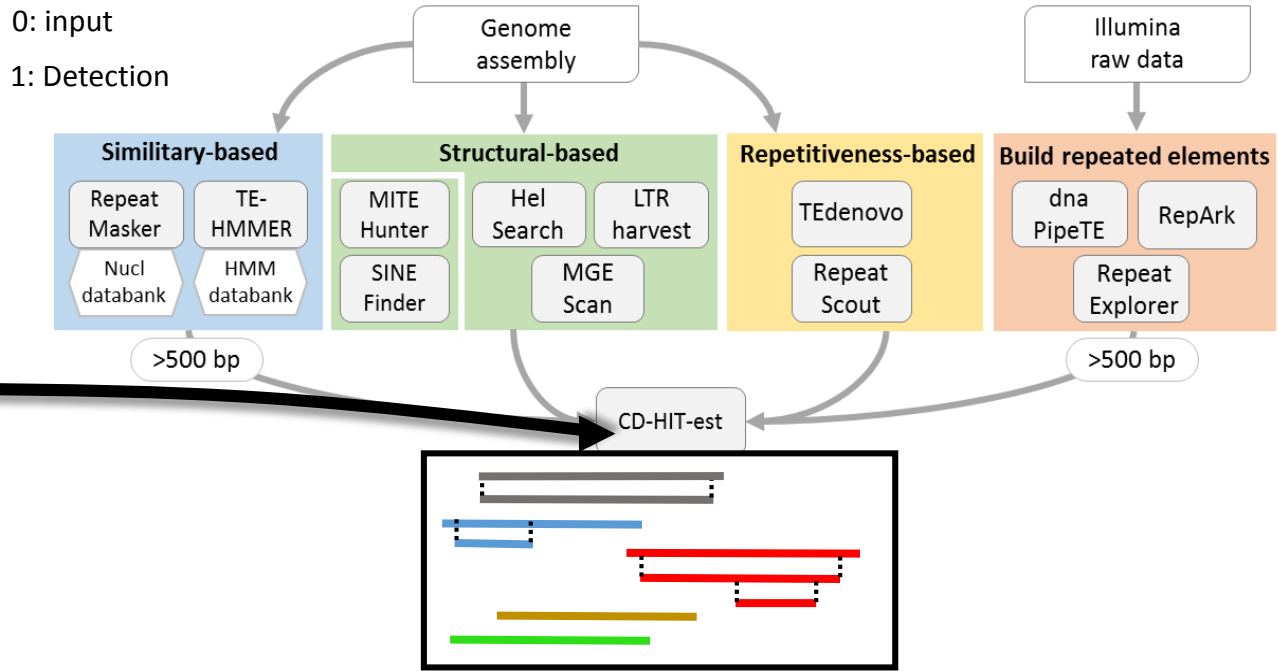
0: input

1: Detection



Concatenate the detected sequences

PiRATE-Galaxy: Detection



Cluster identical sequences
→ Remove redundancy
→ Only keep the larger putative TEs

PiRATE-Galaxy: Detection

Tools

search tools

Get Data

Text Manipulation

Join, Subtract and Group

Filter and Sort

Convert Formats

Fetch Alignments

PIRATE

STEP 1.1: Detection

STEP 1.2: Clustering

[remove short](#) Removal of short sequences

[CD-HIT-est](#) Clusters sequences to reduce the redundancy

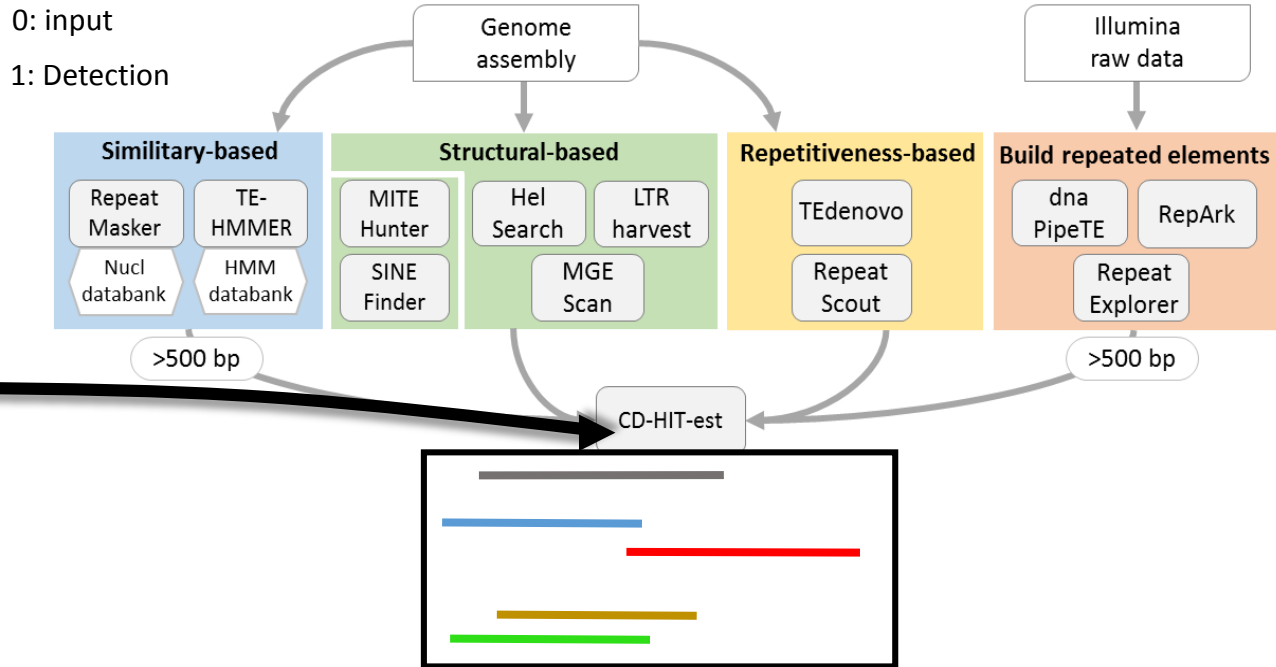
STEP 2.1: Classification

STEP 2.2: Manual check

STEP 3: Annotation

Workflows

- All workflows



Cluster identical sequences
→ Remove redundancy
→ Only keep the larger putative TEs

PiRATE-Galaxy: Classification

Tools

search tools

Get Data

Text Manipulation

Join, Subtract and Group

Filter and Sort

Convert Formats

Fetch Alignments

PIRATE

STEP 1.1: Detection

STEP 1.2: Clustering

STEP 2.1: Classification

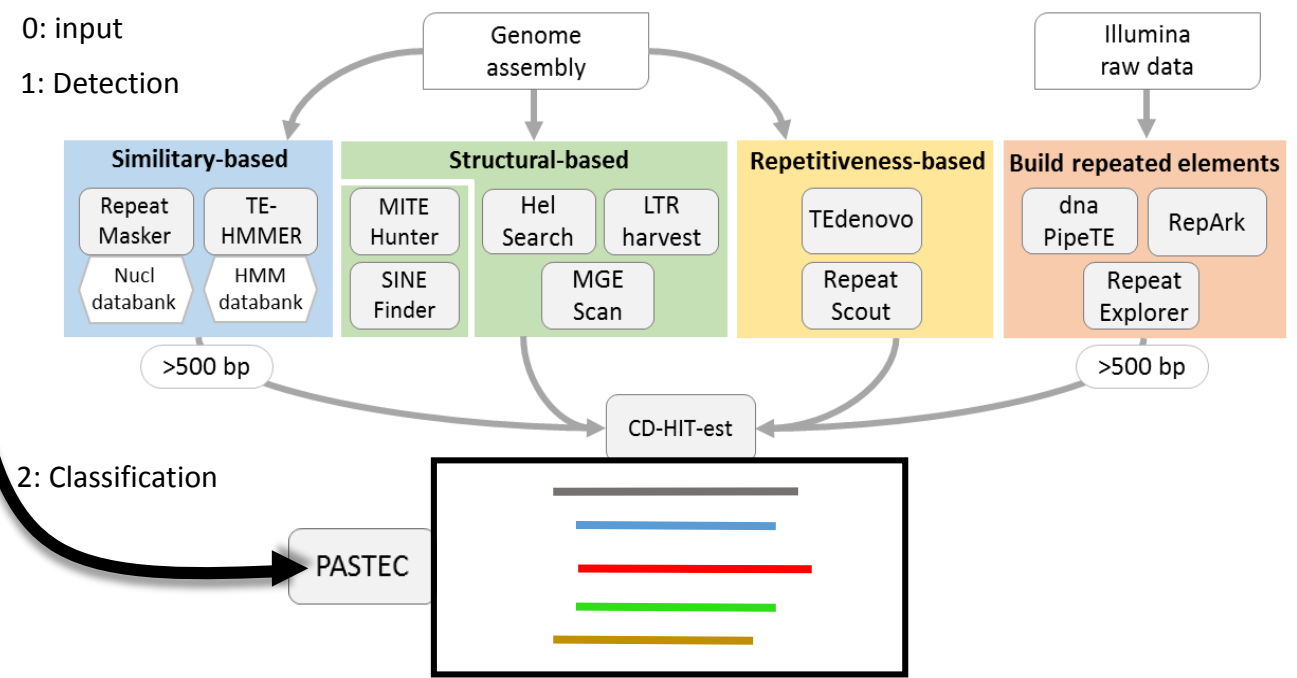
PASTEClassifier Classification of putative TEs

STEP 2.2: Manual check

STEP 3: Annotation

Workflows

- All workflows



Automated classification

PiRATE-Galaxy: Classification

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools

search tools

Get Data
Text Manipulation
Join, Subtract and Group
Filter and Sort
Convert Formats
Fetch Alignments

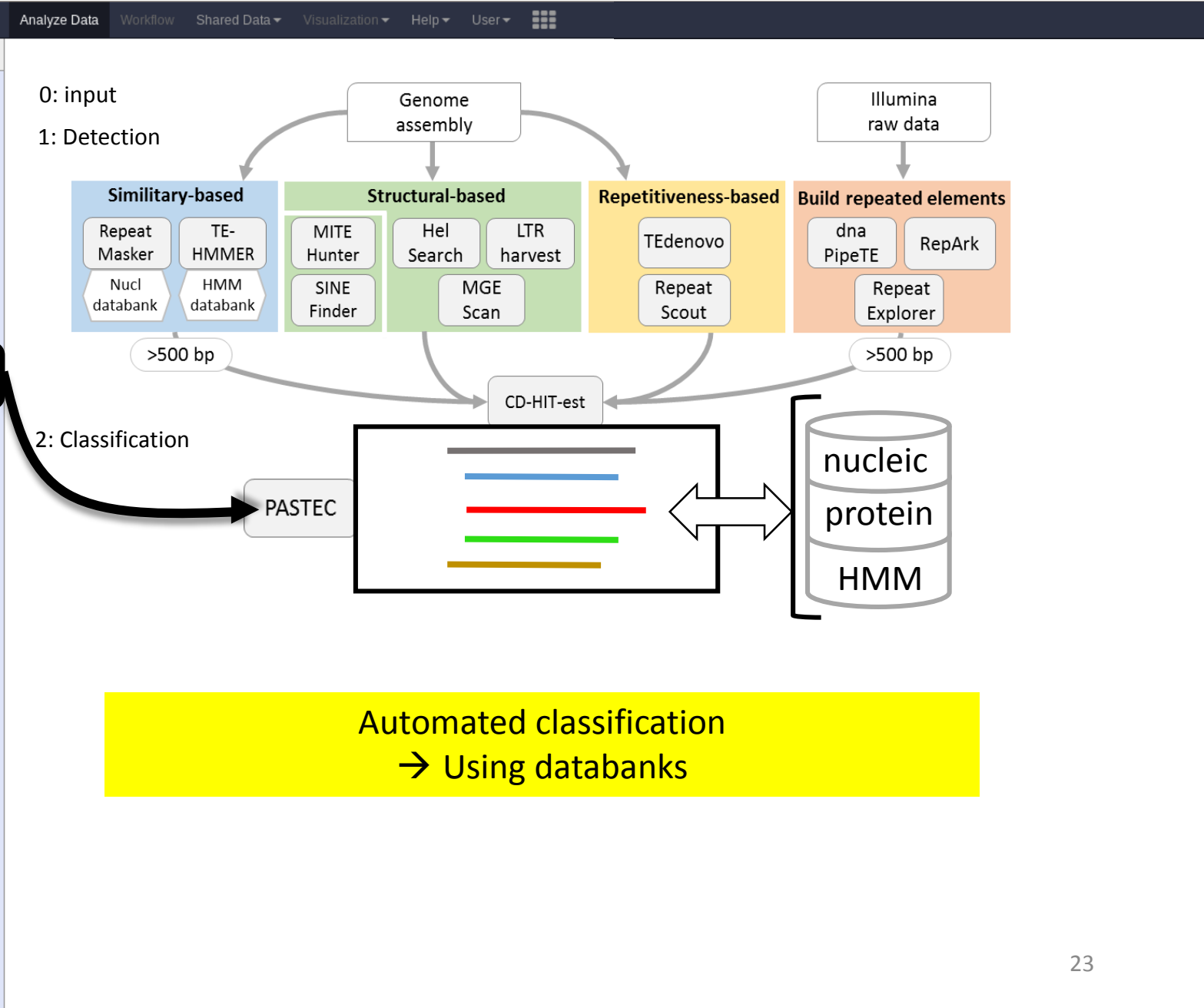
PIRATE

STEP 1.1: Detection
STEP 1.2: Clustering
STEP 2.1: Classification
STEP 2.2: Manual check
STEP 3: Annotation

PASTEClassifier Classification of putative TEs

Workflows

- All workflows



Automated classification
→ Using databanks

PiRATE-Galaxy: Classification

Tools

search tools

Get Data

Text Manipulation

Join, Subtract and Group

Filter and Sort

Convert Formats

Fetch Alignments

PIRATE

STEP 1.1: Detection

STEP 1.2: Clustering

STEP 2.1: Classification

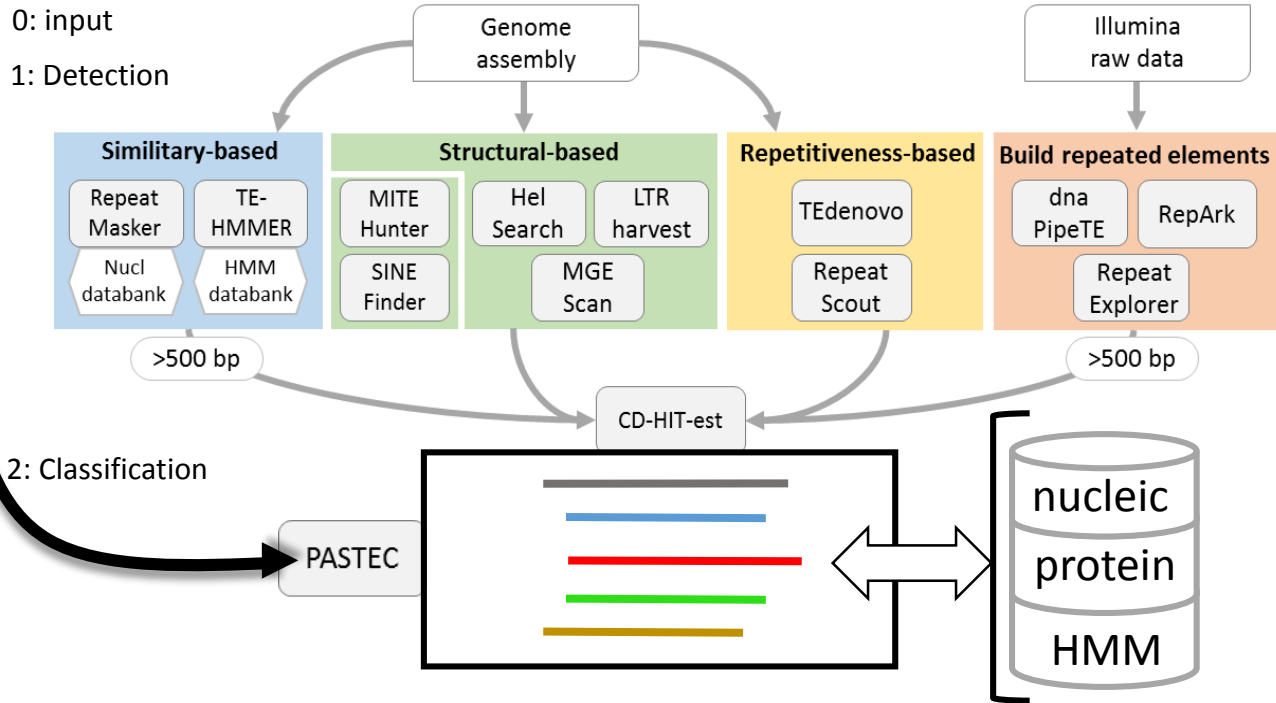
PASTEC Classifier Classification of putative TEs

STEP 2.2: Manual check

STEP 3: Annotation

Workflows

- All workflows



Automated classification
→ Using databanks

- Improved by:
- combining public TE databanks
 - adding non-inverted TEs sequences
 - Creating new HMM profiles

PiRATE-Galaxy: Classification

Tools

search tools

Get Data

Text Manipulation

Join, Subtract and Group

Filter and Sort

Convert Formats

Fetch Alignments

PIRATE

STEP 1.1: Detection

STEP 1.2: Clustering

STEP 2.1: Classification

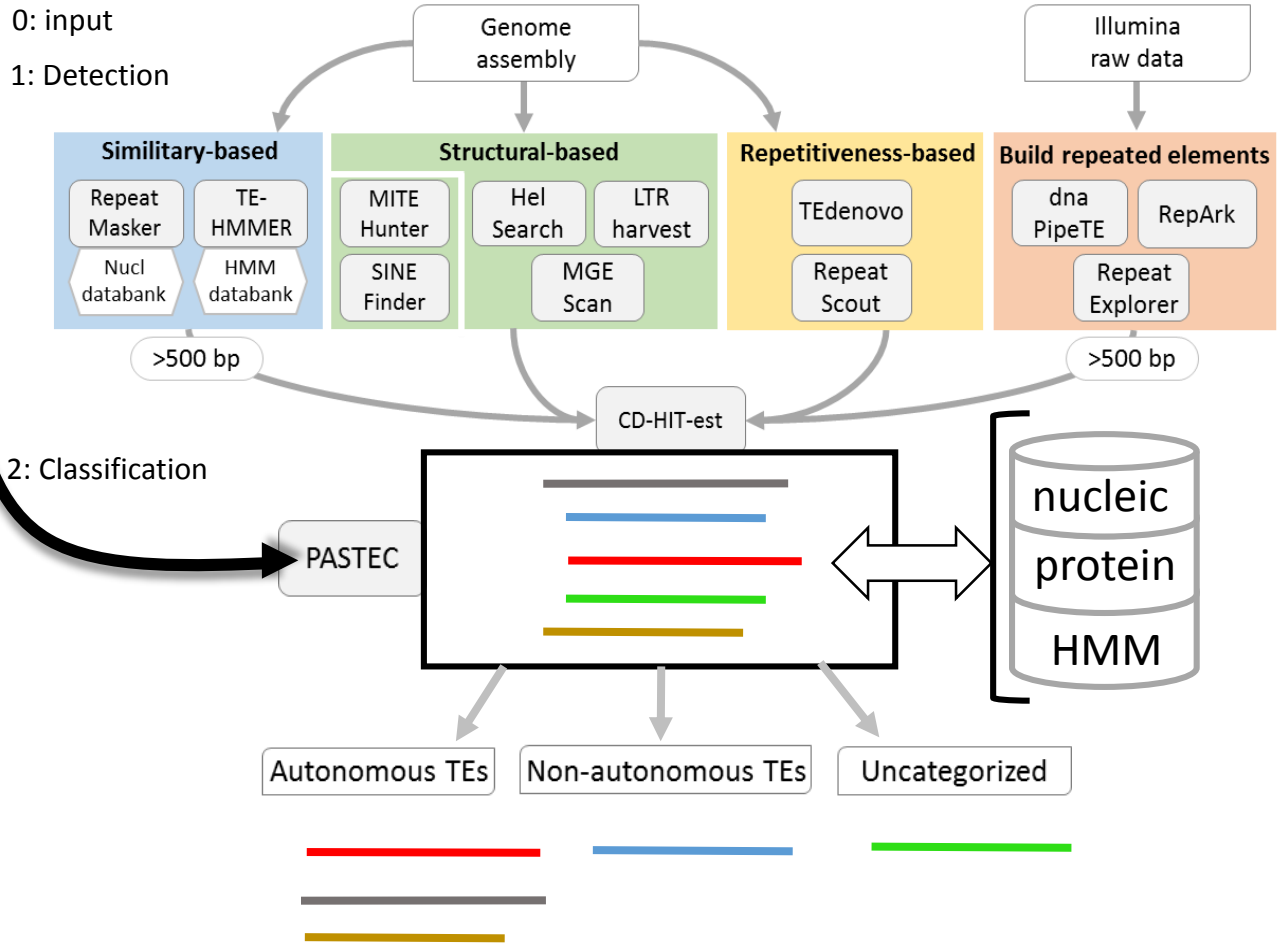
PASTEC Classifier Classification of putative TEs

STEP 2.2: Manual check

STEP 3: Annotation

Workflows

- All workflows



PiRATE-Galaxy: Classification

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools

search tools

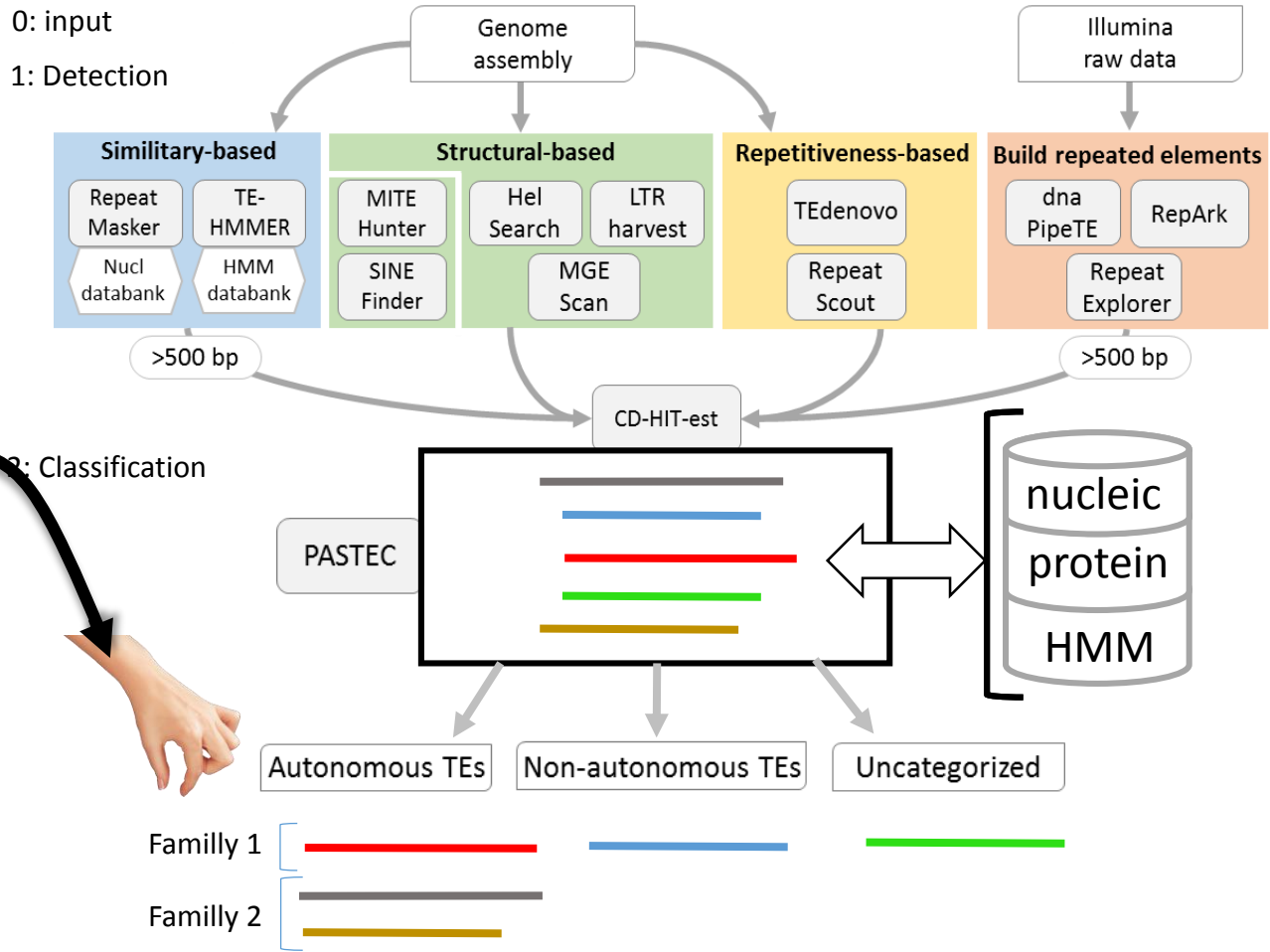
Get Data
Text Manipulation
Join, Subtract and Group
Filter and Sort
Convert Formats
Fetch Alignments

PIRATE

STEP 1.1: Detection
STEP 1.2: Clustering
STEP 2.1: Classification
STEP 2.2: Manual check
STEP 3: Annotation

MCL Groups TE sequences into TE families
BLASTn Compares the TE sequences

Workflows
All workflows



Group TE sequences into TEs families

PiRATE-Galaxy: Annotation

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Tools

search tools

Get Data
Text Manipulation
Join, Subtract and Group
Filter and Sort
Convert Formats
Fetch Alignments

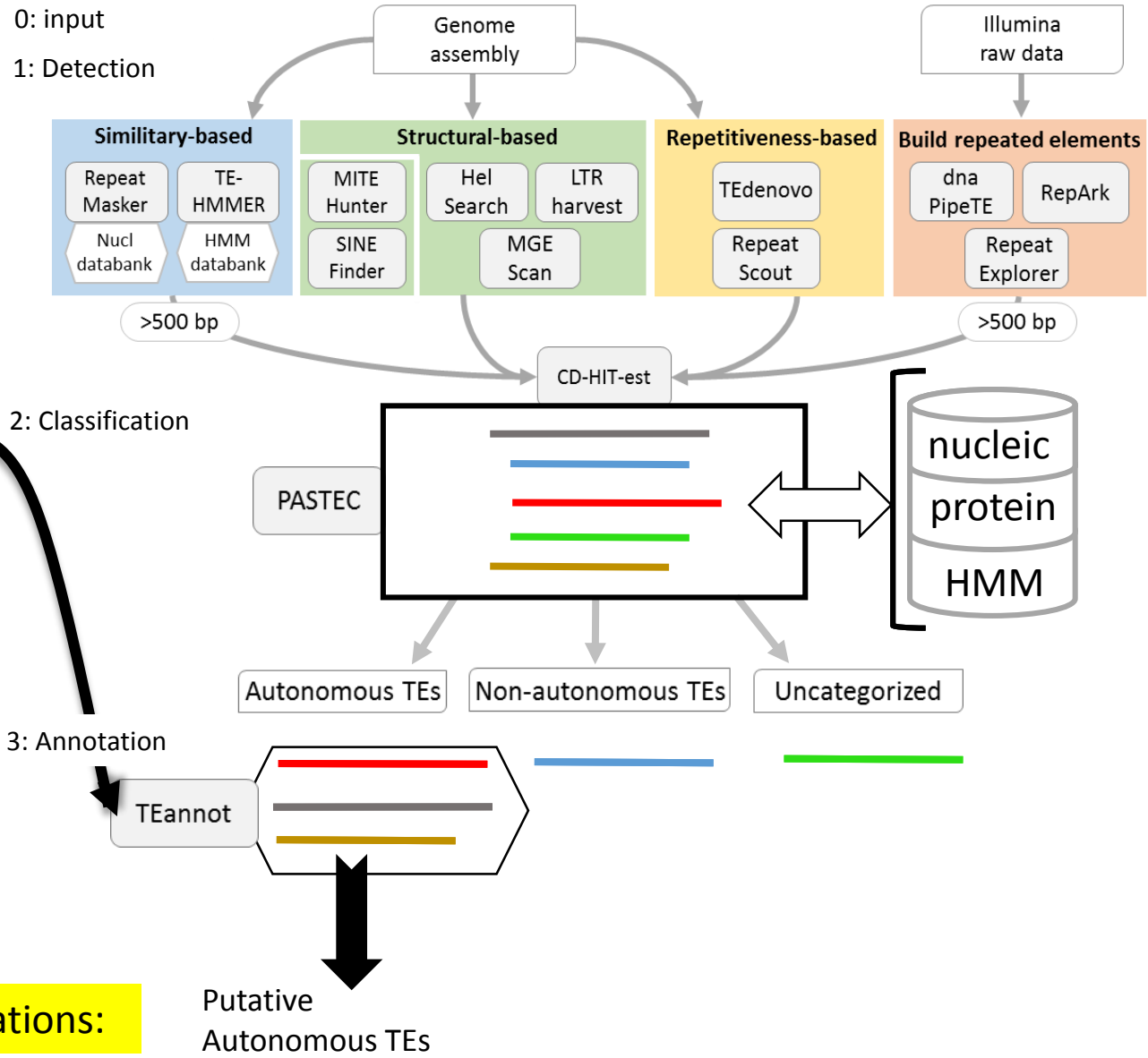
PIRATE

STEP 1.1: Detection
STEP 1.2: Clustering
STEP 2.1: Classification
STEP 2.2: Manual check
STEP 3: Annotation

Run 1 TEannot (REPET) Annotation of detected TE families
Run 2 TEannot (REPET) Annotation of detected TE families

Workflows

- All workflows



3 annotations:

PiRATE-Galaxy: Annotation

Tools

search tools

Get Data

Text Manipulation

Join, Subtract and Group

Filter and Sort

Convert Formats

Fetch Alignments

PIRATE

STEP 1.1: Detection

STEP 1.2: Clustering

STEP 2.1: Classification

STEP 2.2: Manual check

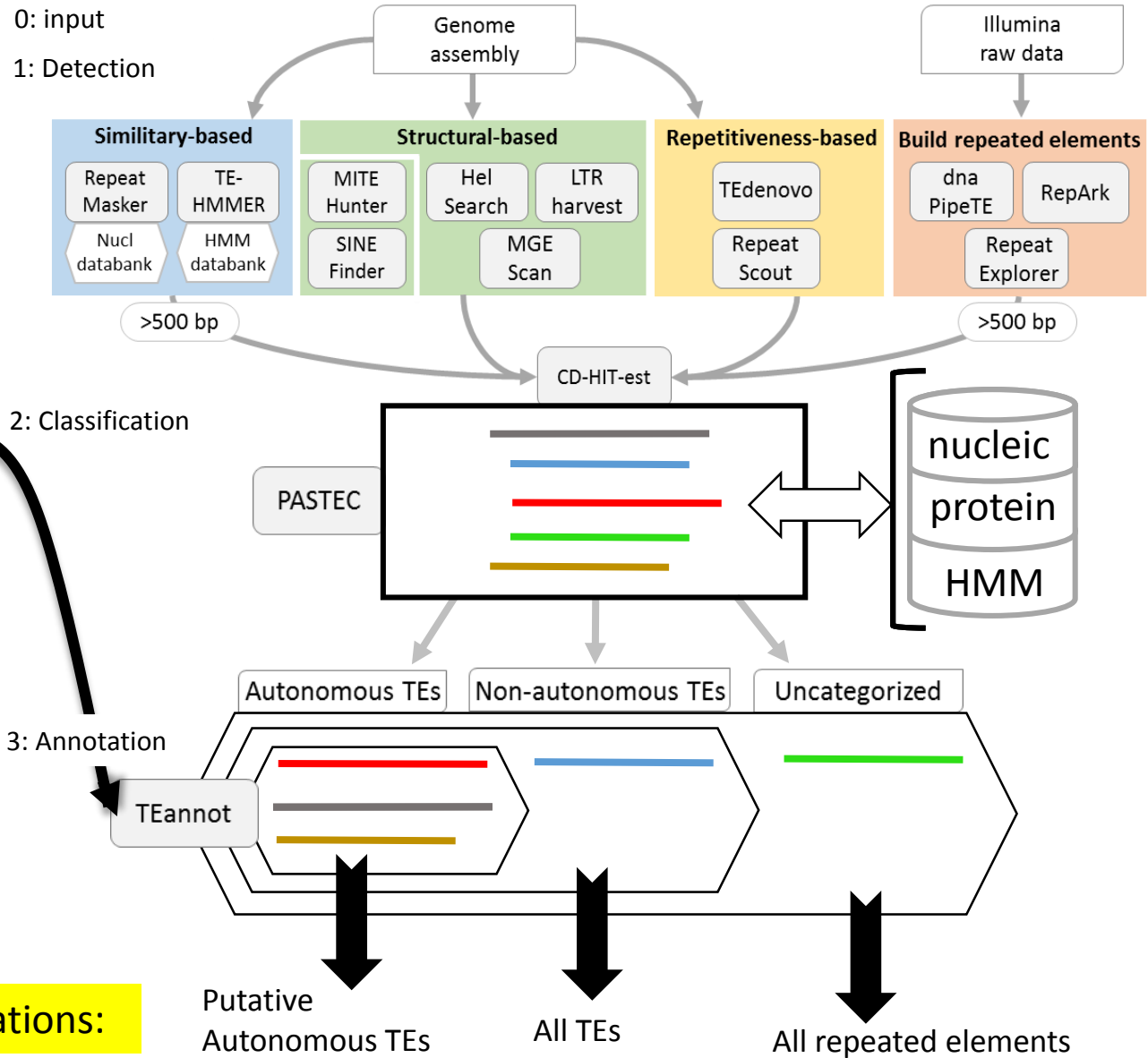
STEP 3: Annotation

Run 1 TEannot (REPET) Annotation of detected TE families

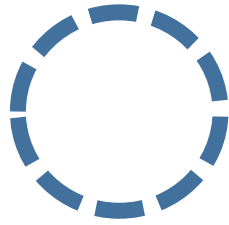
Run 2 TEannot (REPET) Annotation of detected TE families

Workflows

- All workflows



Control with *Arabidopsis thaliana*



Genome size: **157 Mb**



259 TE families

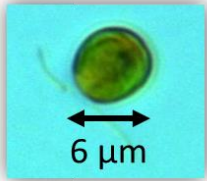
Arabidopsis thaliana

- PiRATE detects **81%** of the TE families of *A. thaliana*
(difficulty to detect non-autonomous TEs)
- **75%** of these detected TEs were correctly classified
(difficulty to classify non-autonomous TEs)

Use of PiRATE with *T. lutea* (Haptophyta)

Example with the microalga *Tisochrysis lutea*

Phylum: Haptophyte



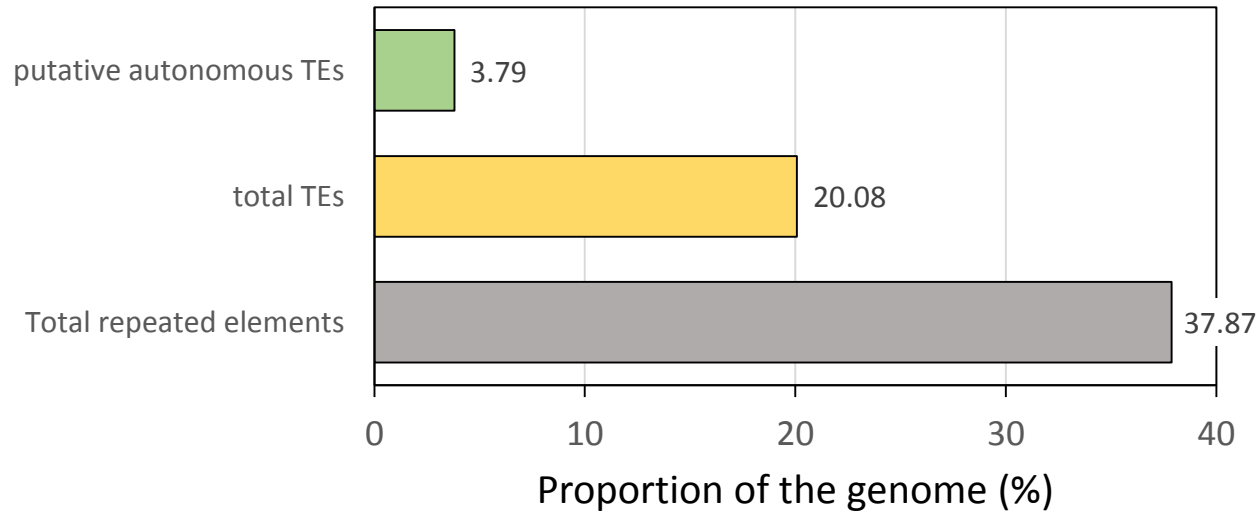
Tisochrysis lutea



Genome size: **82 Mb**
193 contigs



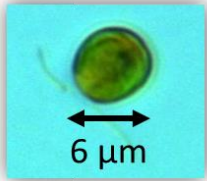
17 TE families
of Haptophytes



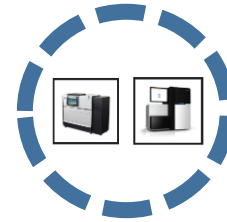
Use of PiRATE with *T. lutea* (Haptophyta)

Example with the microalga *Tisochrysis lutea*

Phylum: Haptophyte



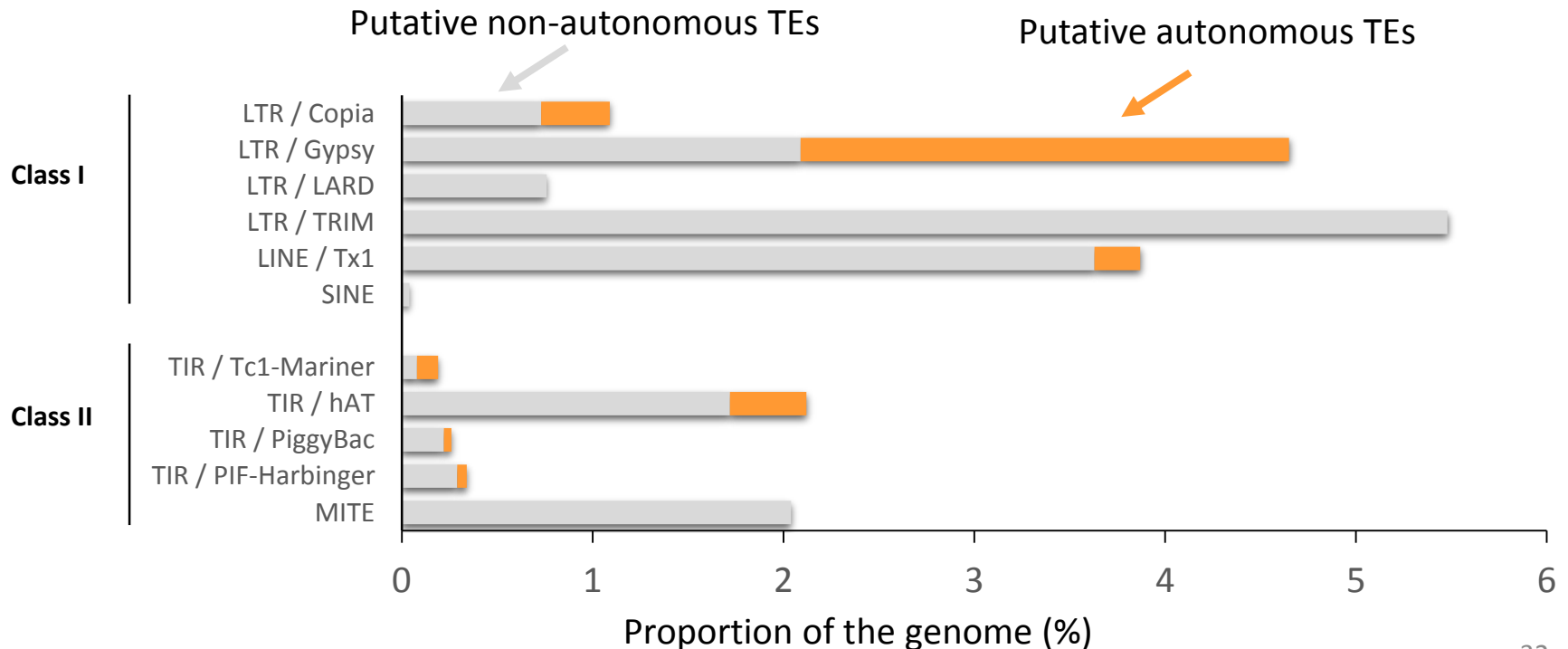
Tisochrysis lutea



Genome size: **82 Mb**
193 contigs



17 TE families
of Haptophytes



Thanks for your attention!

Acknowledgements:



Grégory
Carrier



Bruno
Saint-Jean



Nathalie
Casse

- Nicolas Daccord

- Véronique Jamilloux

Funding:



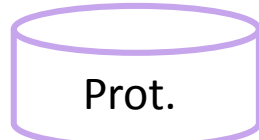
How improve TE classification ?



PASTEClassifier

To improve: Combine public databanks and adding non-inventoried TEs

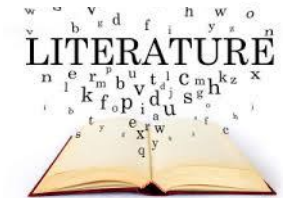
Databanks



Public TE databanks



Non-inventoried TEs



Public profil HMM



New profiles HMM



→ done

- - - - -> done for microalgal genomes