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

PAG XXVI

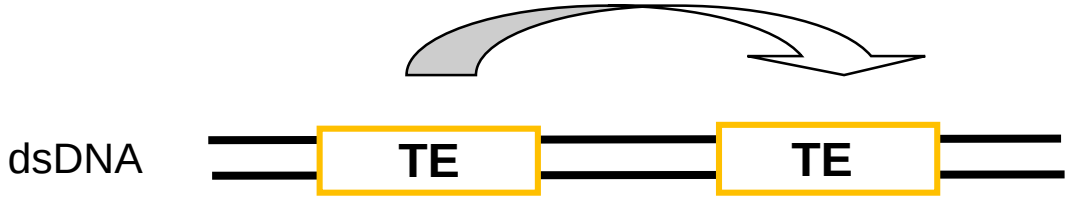
Transposable elements (TEs)

TEs are DNA sequences able to move (= transposition)

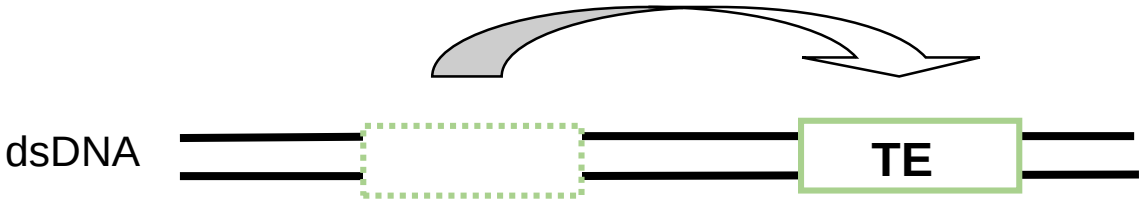
into the host genome of eucaryotic and procaryotic organisms

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

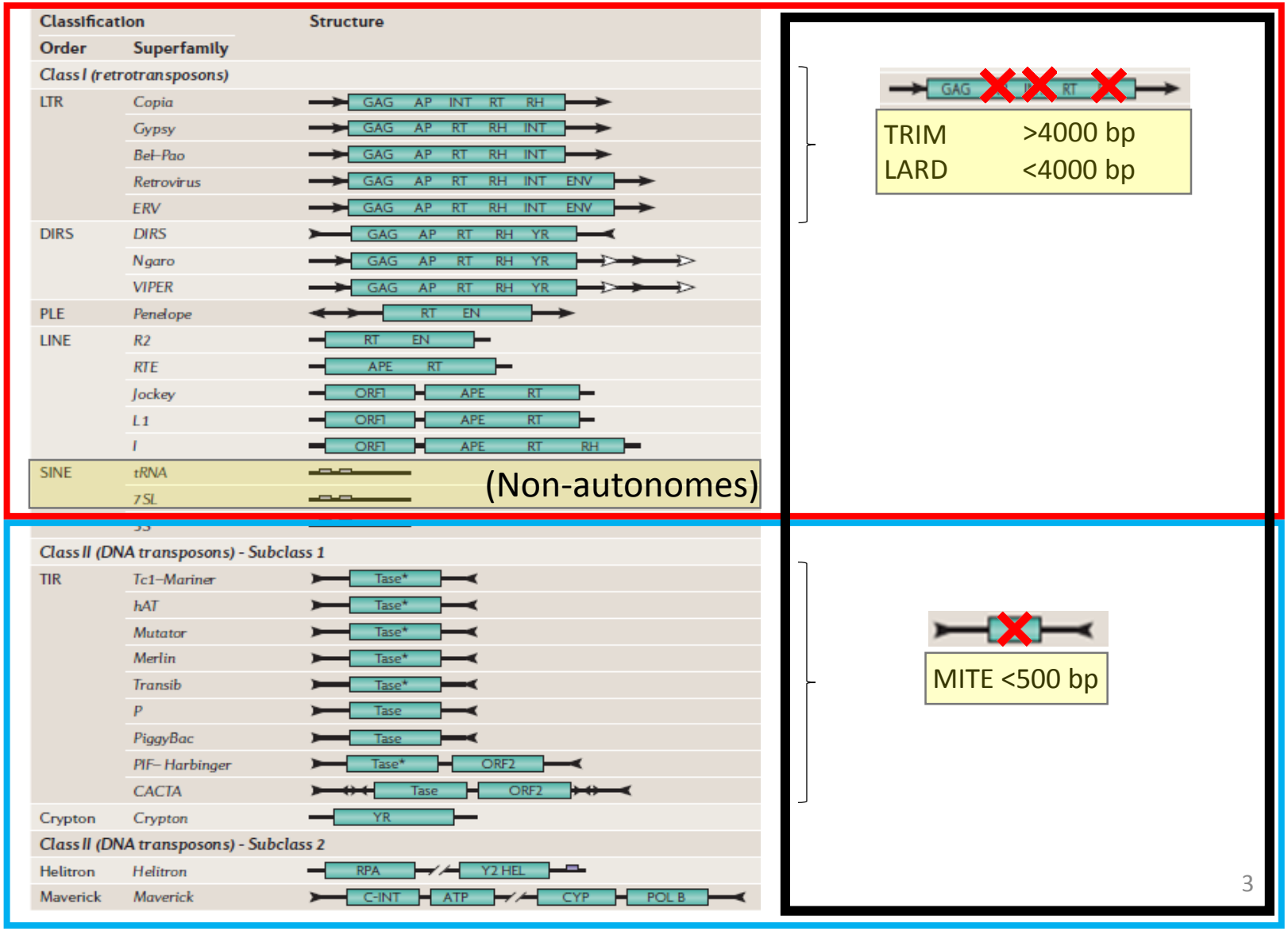
Wicker et al., 2007

Autonomous TEs

Non-autonomous TEs

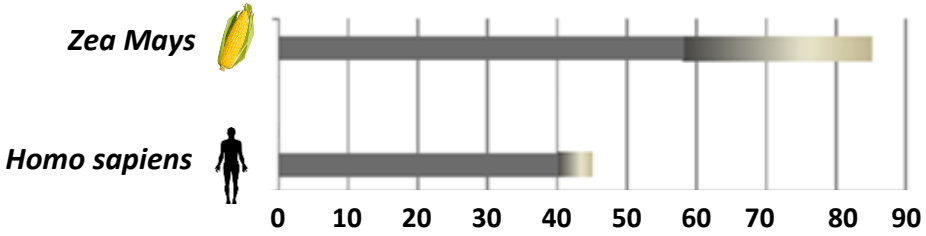
Classe I

Classe II



Impact of TEs on genomes

TEs impact genome size



Chénais et al., 2012

TEs can generate phenotype changes

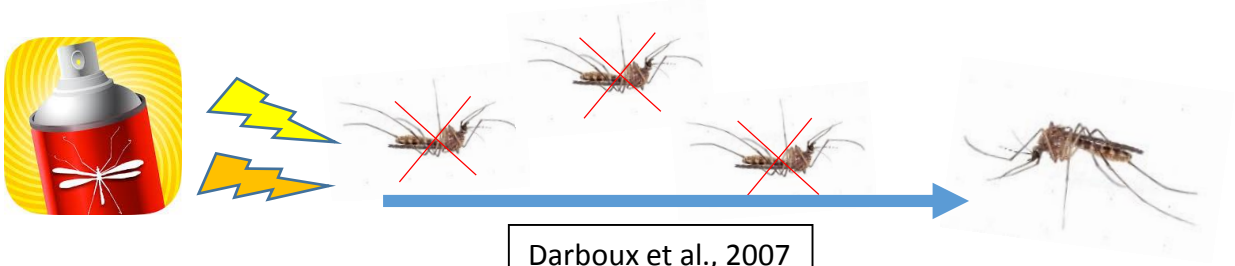
Vitis vinifera

Kobayashi et al., 2004

Citrus sinensis

Butelli et al., 2012

TE mutations can promote species adaptation

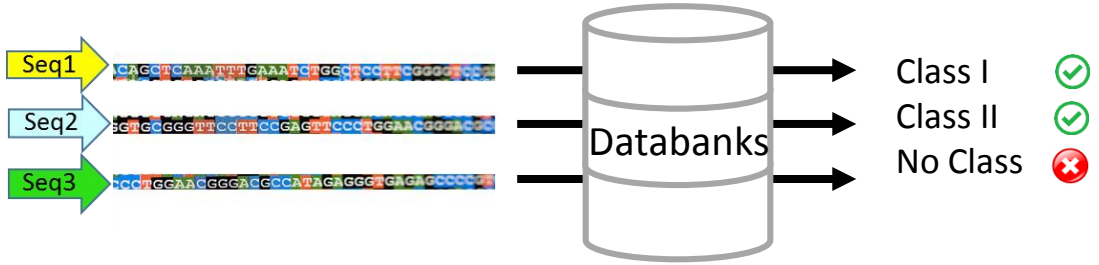


Darboux et al., 2007

How to conduct a de novo TE annotation

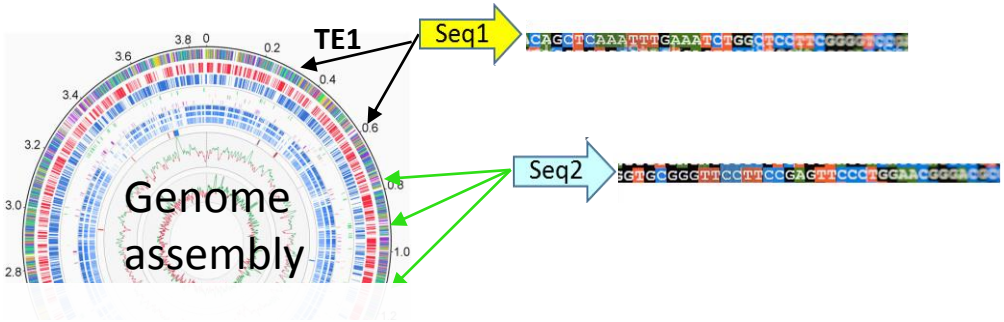
1. Detection of TEs

2. Classification of detected TEs



Comparaison **(BLAST)**

3. TE annotation

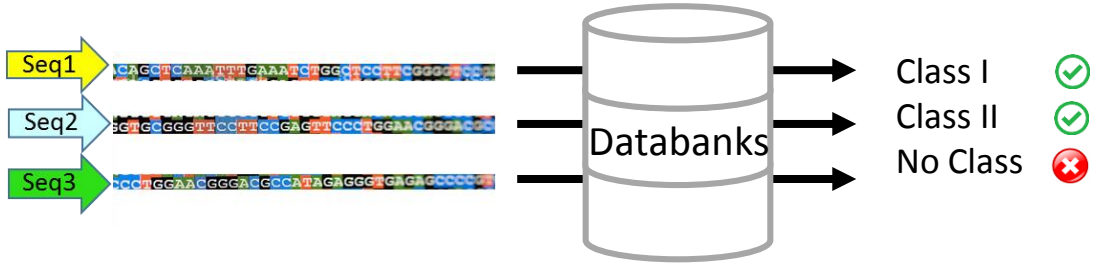


Limitations with non-model organisms

1. Detection of TEs

Genome assembly is usually fragmented
→ missassembly of repeated elements

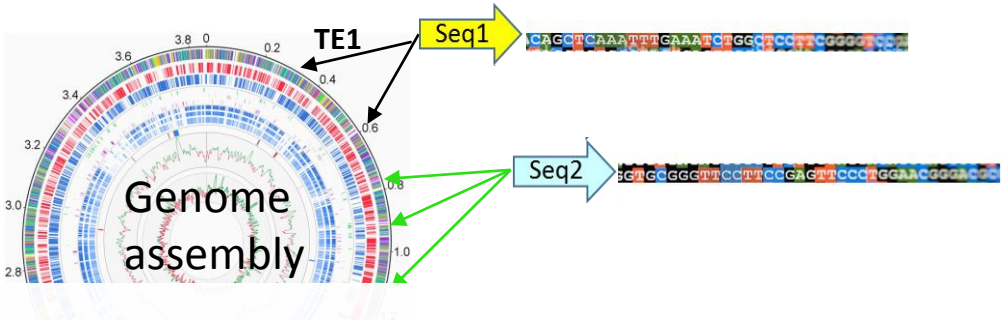
2. Classification of detected TEs



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

Comparaison **(BLAST)**

3. TE annotation



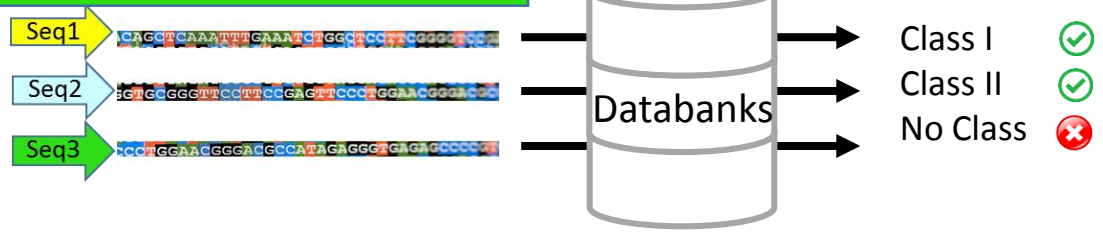
Limitations with non-model organisms

1. Detection of TEs

Genome assembly is usually fragmented
→ missassembly of repeated elements

Improve the detection step

2. Classification of detected TEs

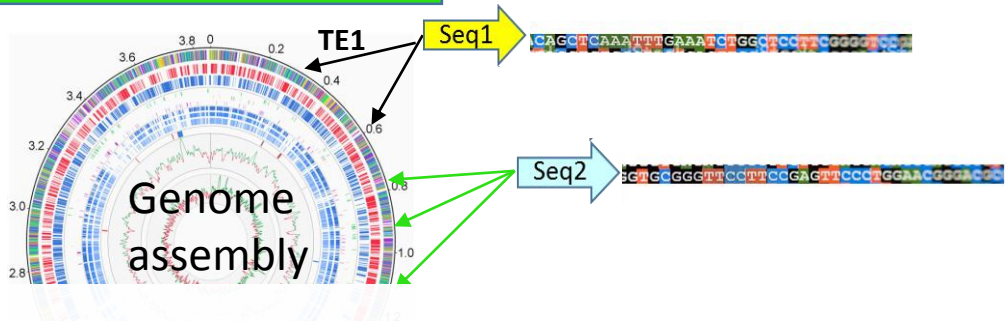


Few knowledge in TE databanks
→ difficulty to classify/recognize TEs

Comparaison **(BLAST)**

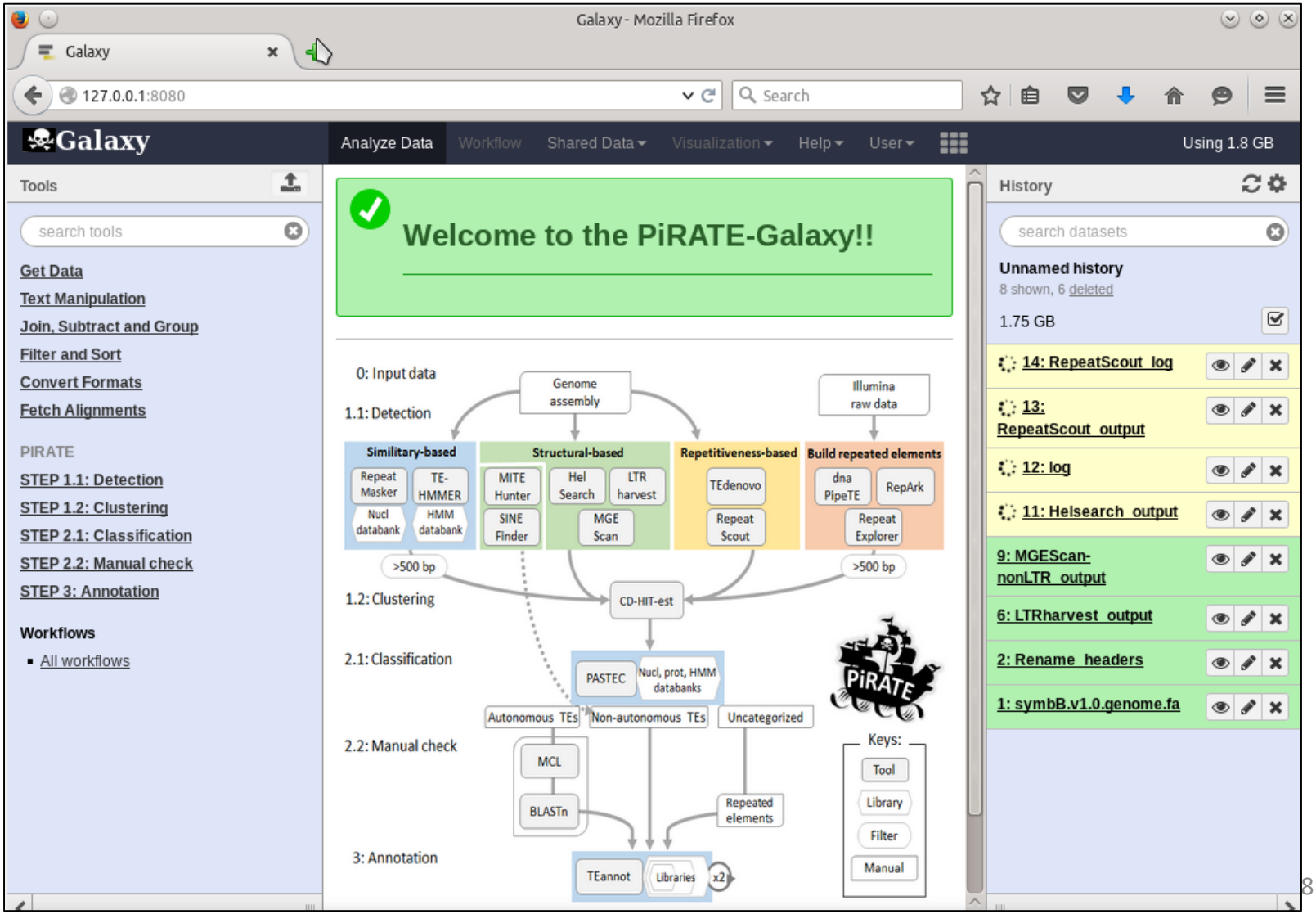
Improve the classification step

3. TE annotation



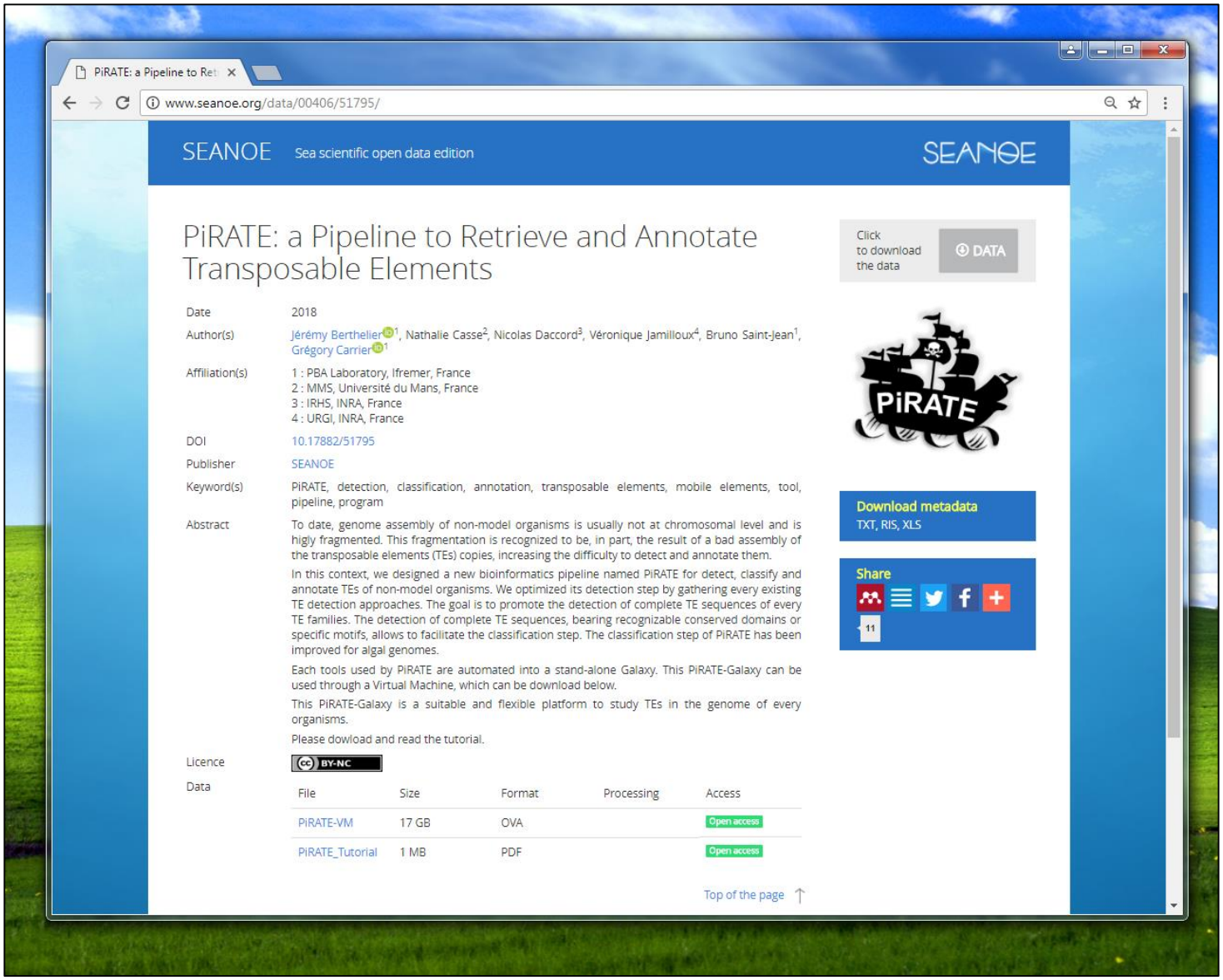
PiRATE-Galaxy distribution

→ PiRATE is implemented into a stand-alone Galaxy:



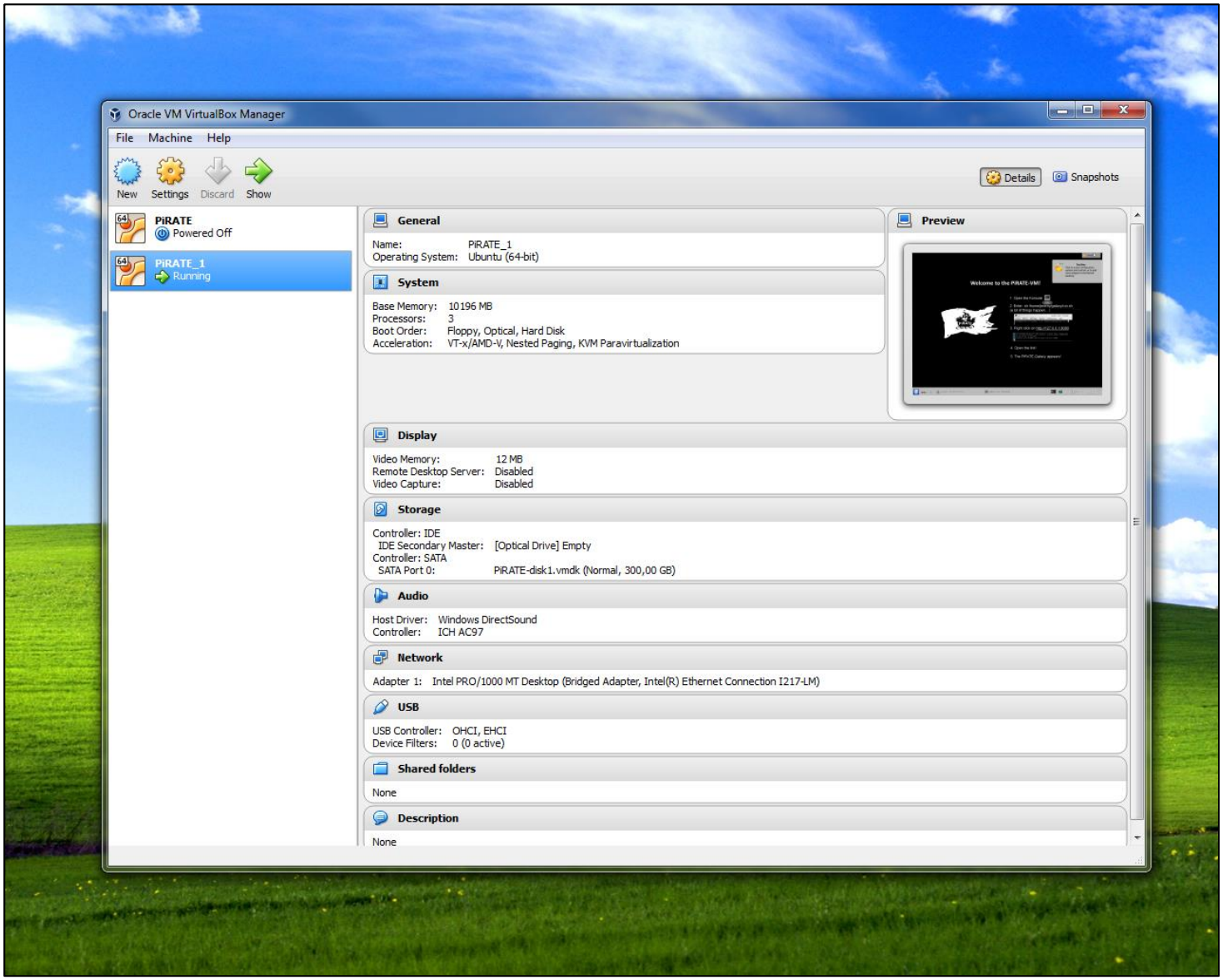
PiRATE-Galaxy distribution

1. Download the PiRATE Virtual Machine (Virtual Linux)



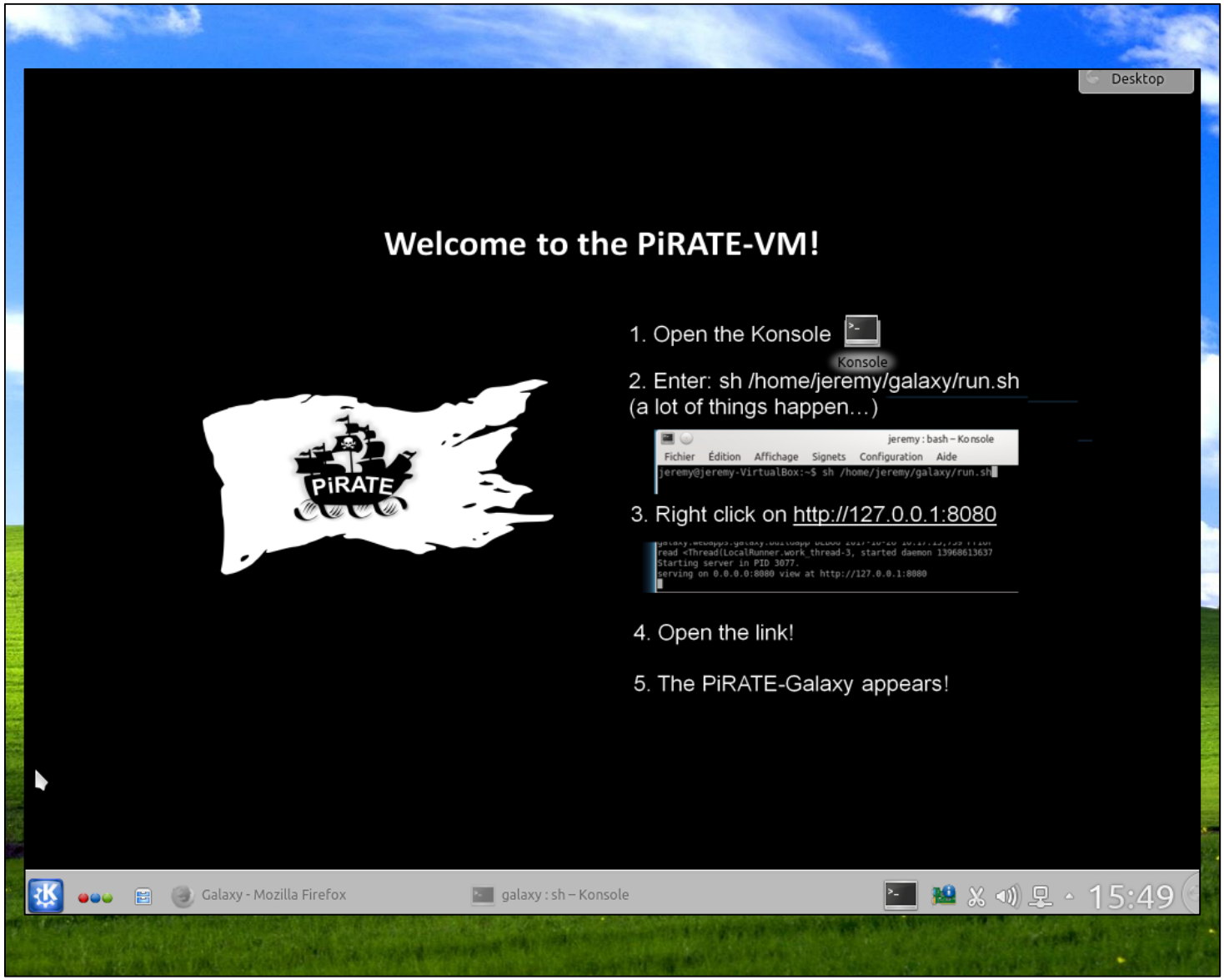
PiRATE-Galaxy distribution

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



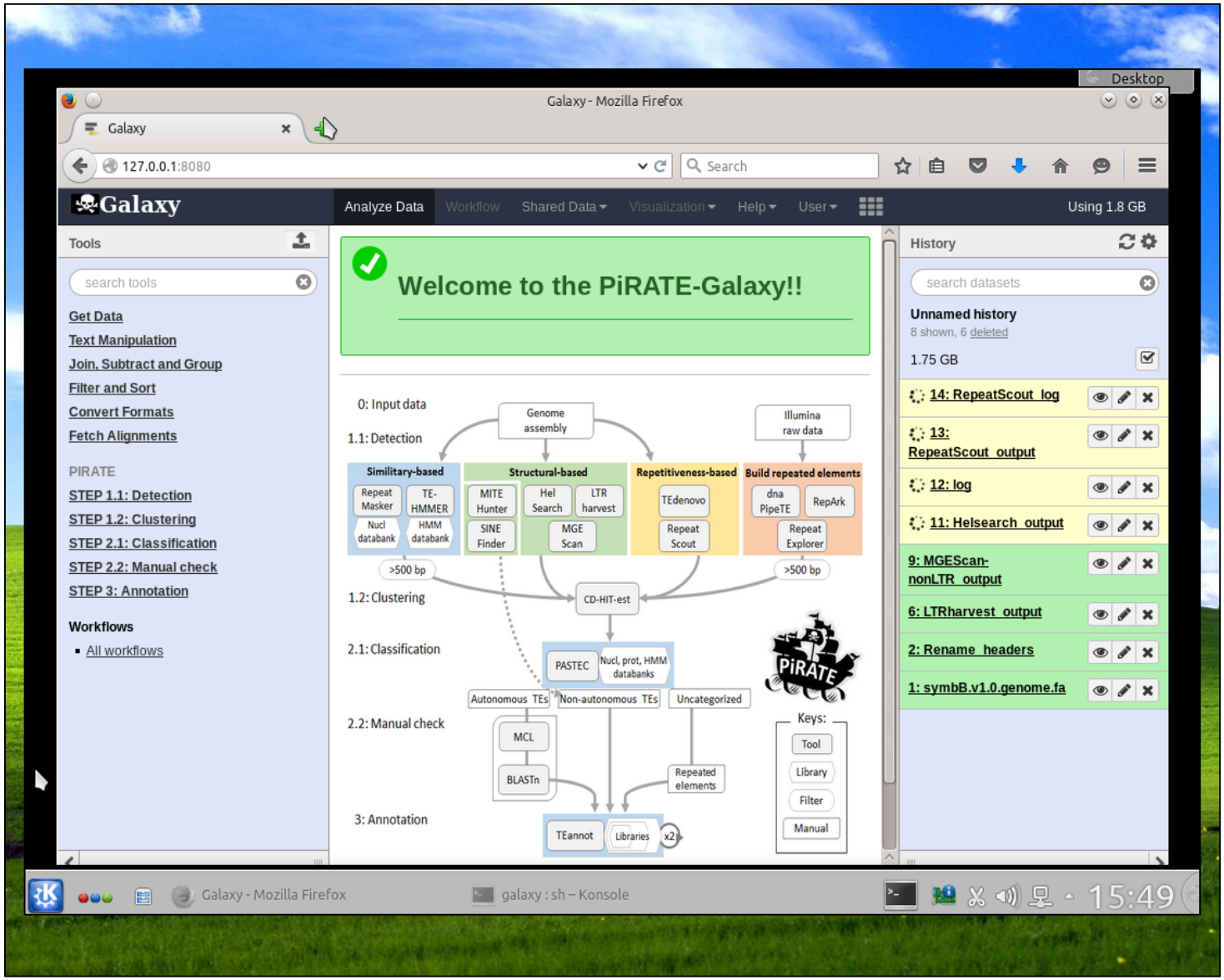
PiRATE-Galaxy distribution

3. Launch the PiRATE-Galaxy

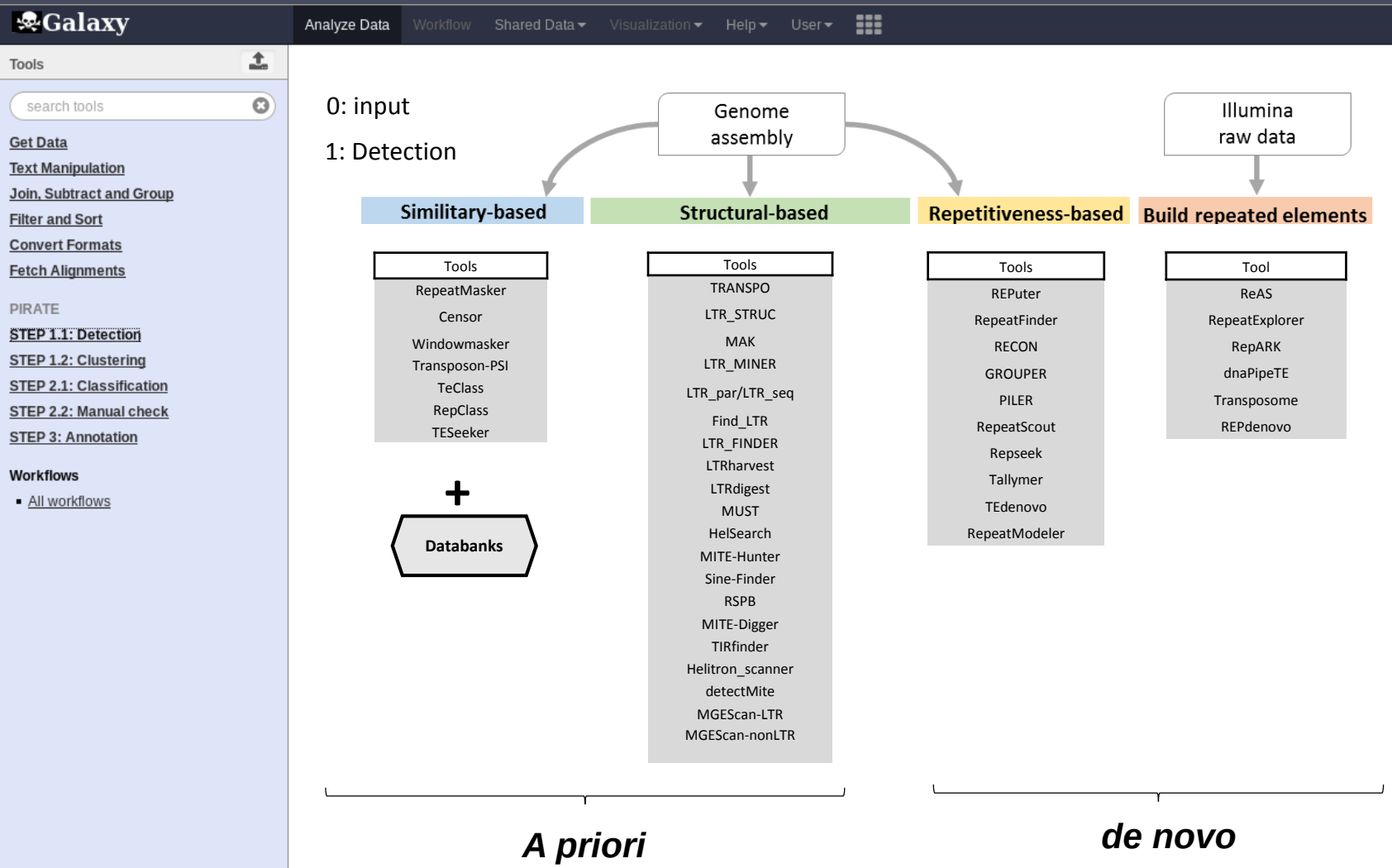


PiRATE-Galaxy distribution

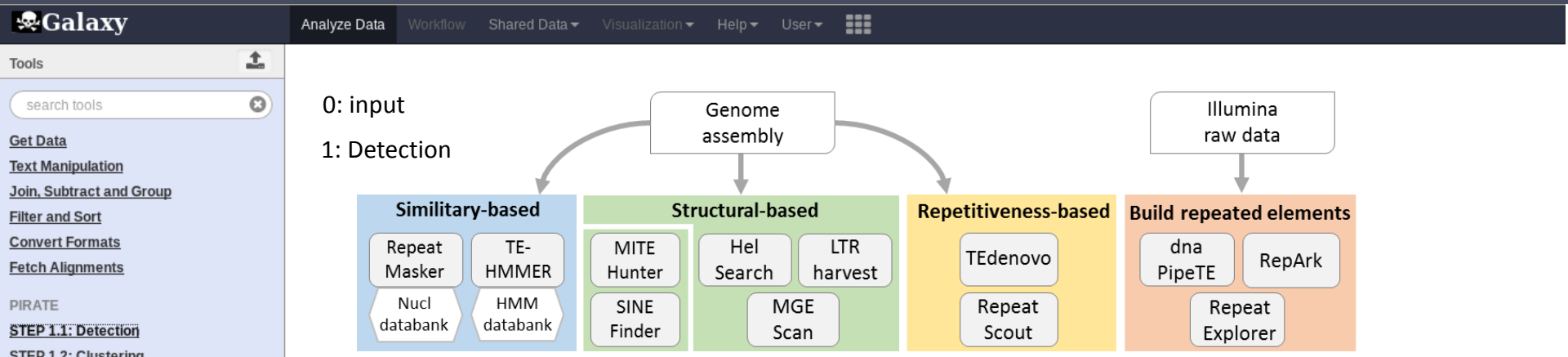
It’s ready to use !!!



PiRATE-Galaxy: Detection



PiRATE-Galaxy: Detection

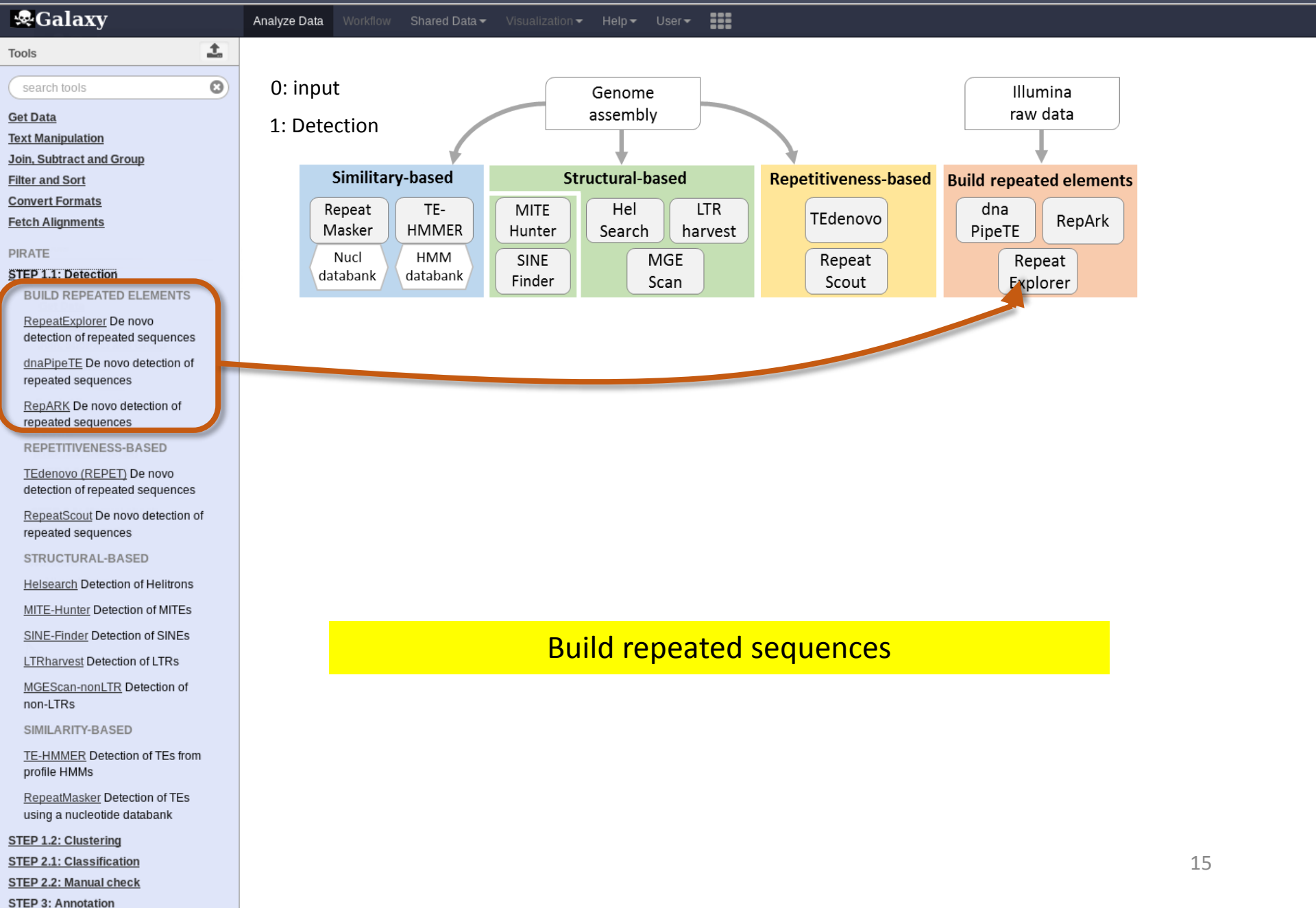


Group every existing TE detection approach to:

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

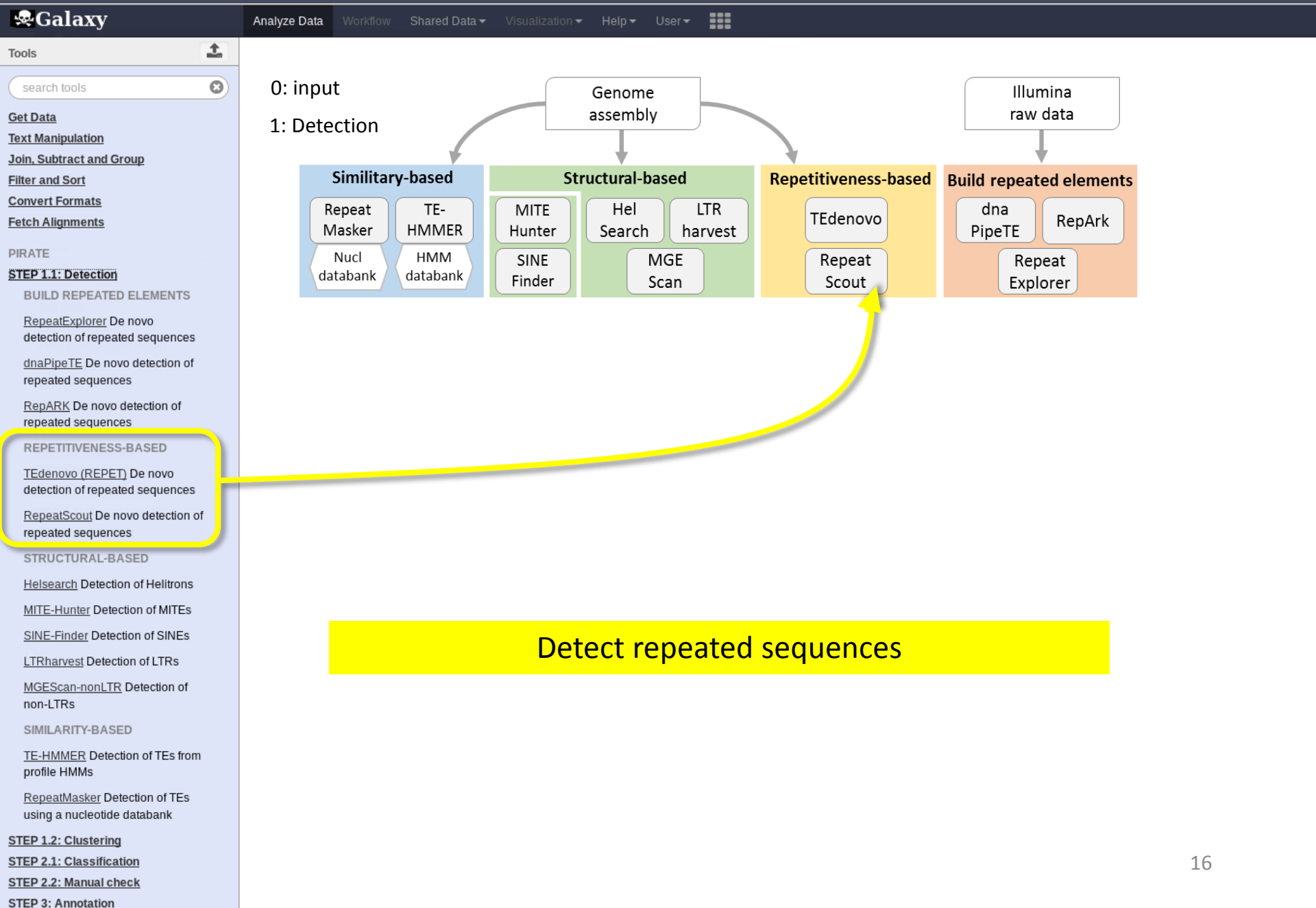
→ To facilitate the classification step

PiRATE-Galaxy: Detection

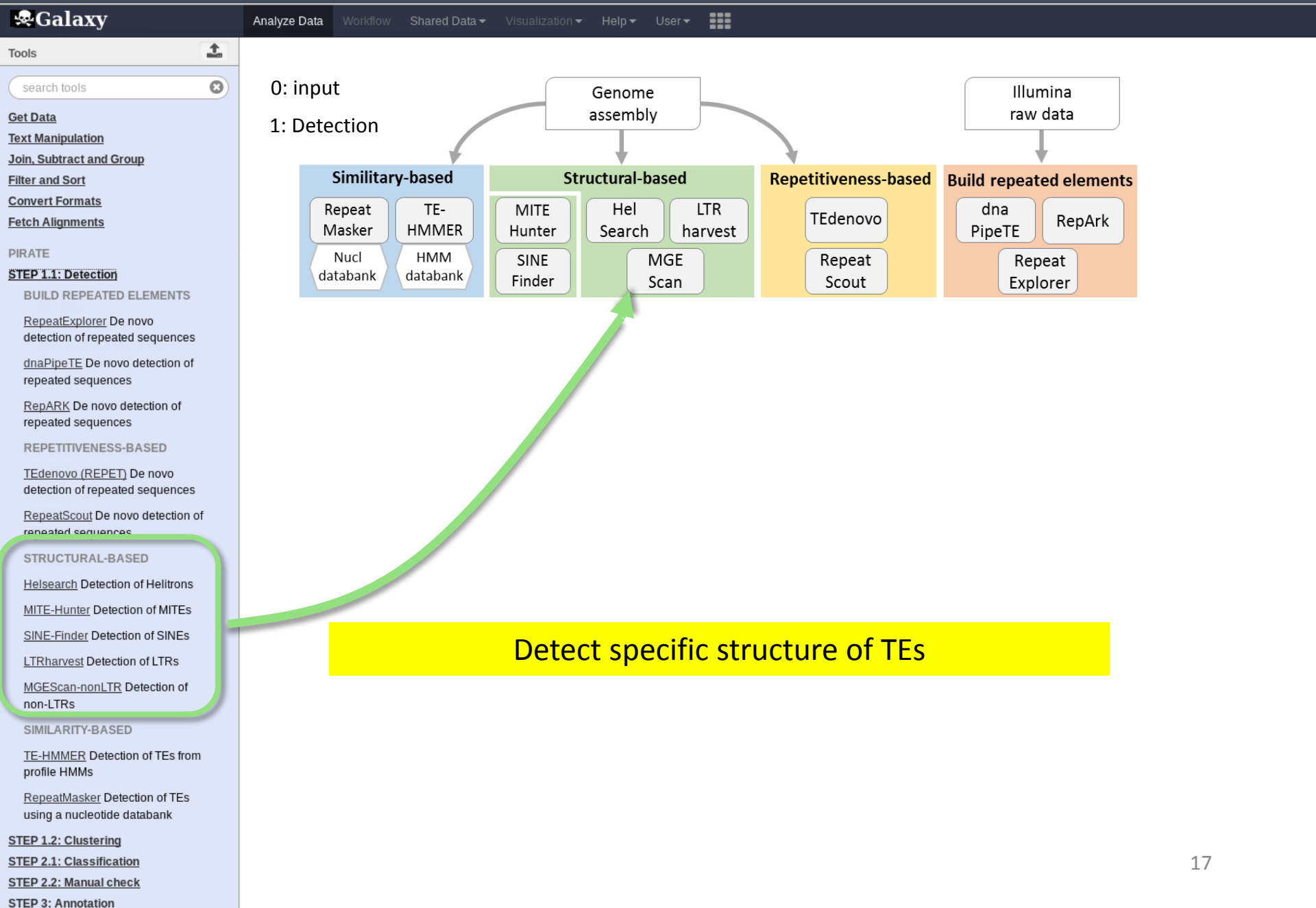


Build repeated sequences

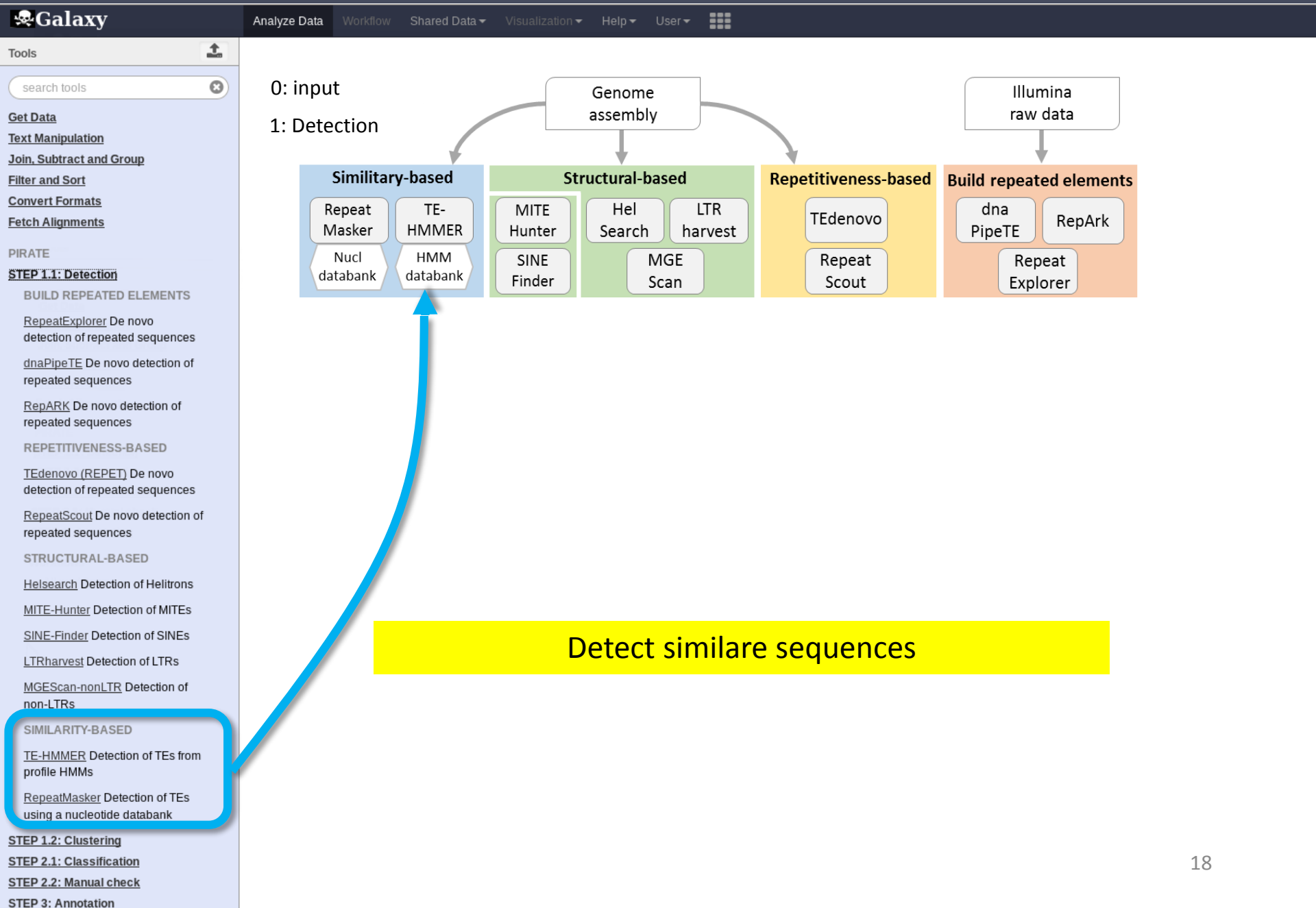
PiRATE-Galaxy: Detection



PiRATE-Galaxy: Detection



PiRATE-Galaxy: Detection



PiRATE-Galaxy: Detection

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

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

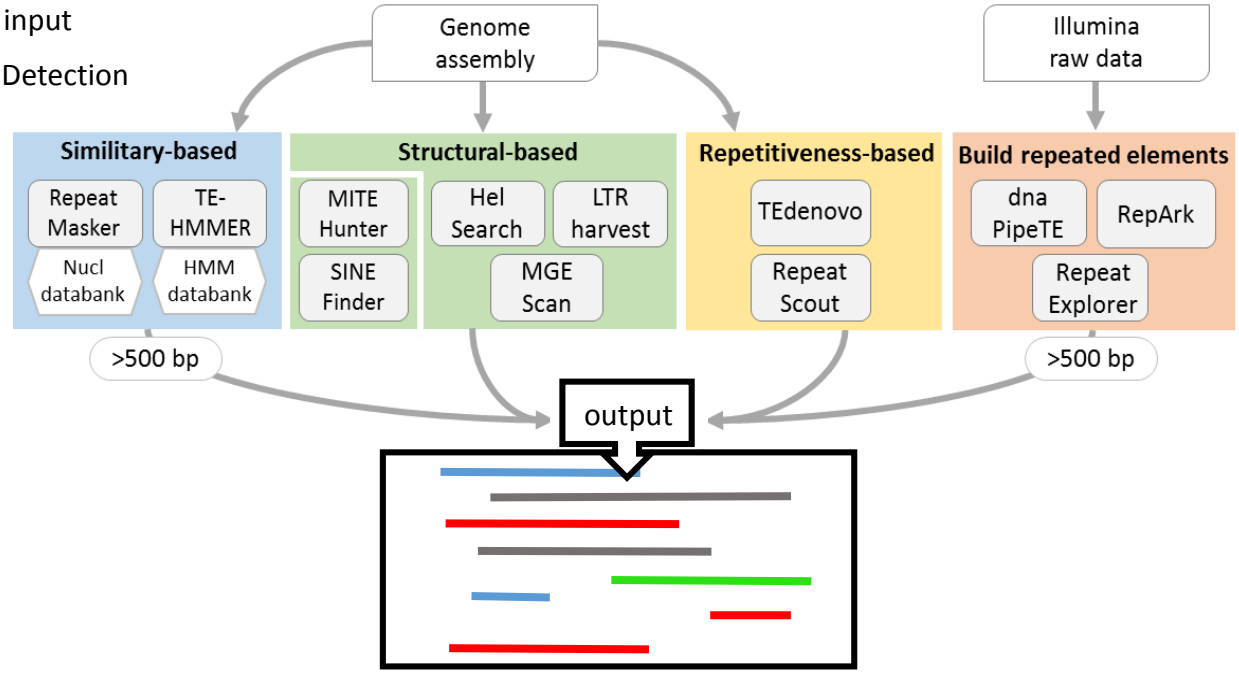
STEP 2.1: Classification

STEP 2.2: Manual check

STEP 3: Annotation

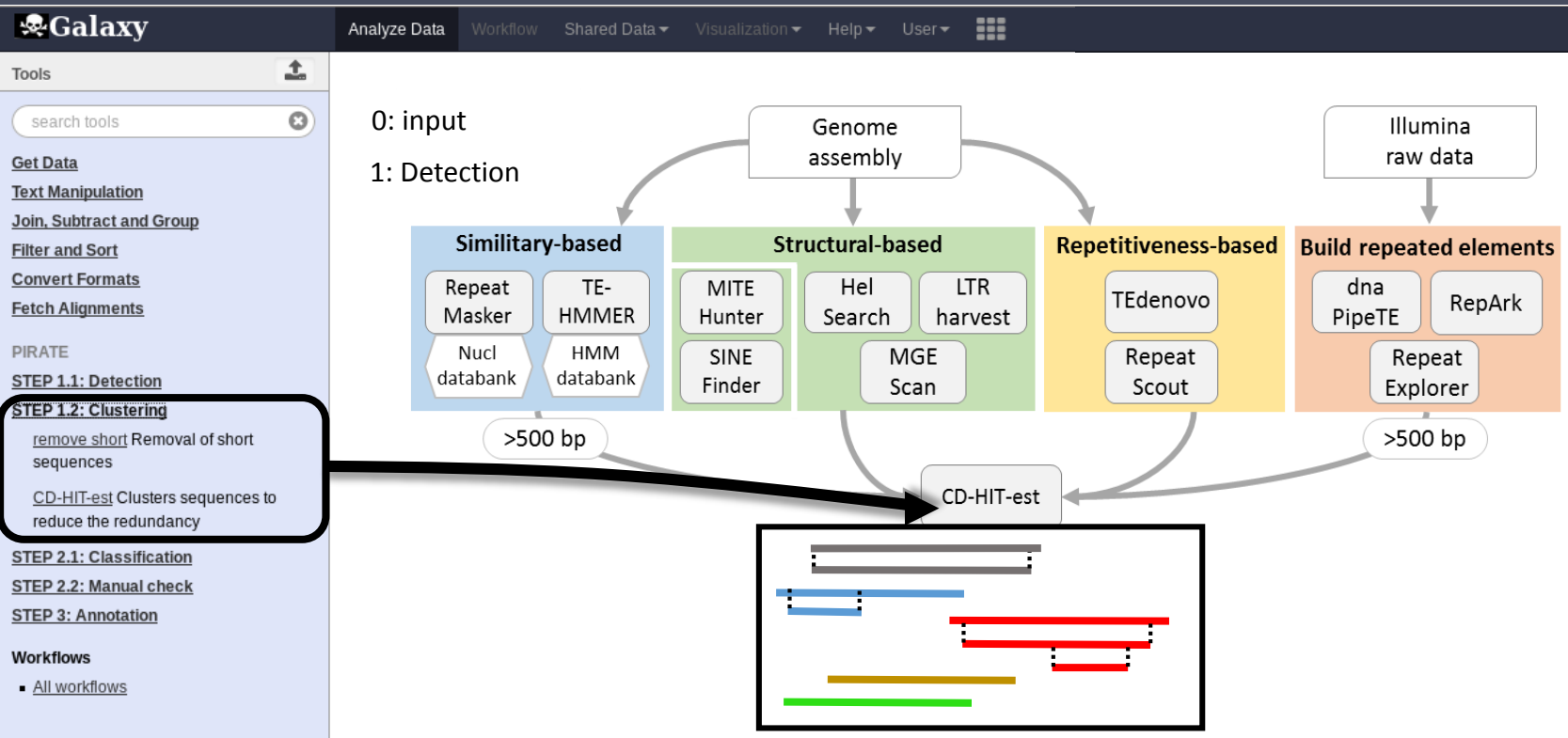
0: input

1: Detection



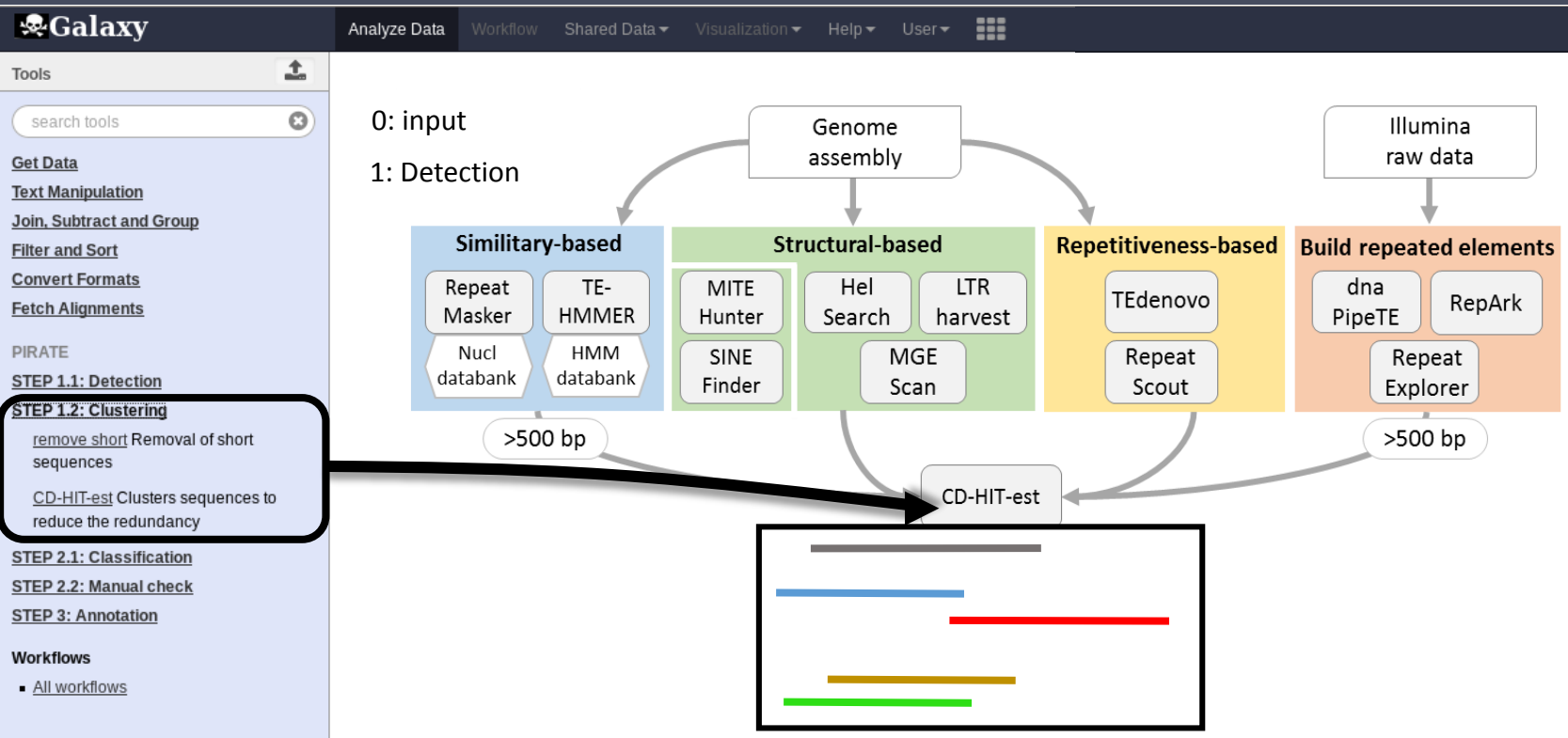
Concatenate the detected sequences

PiRATE-Galaxy: Detection



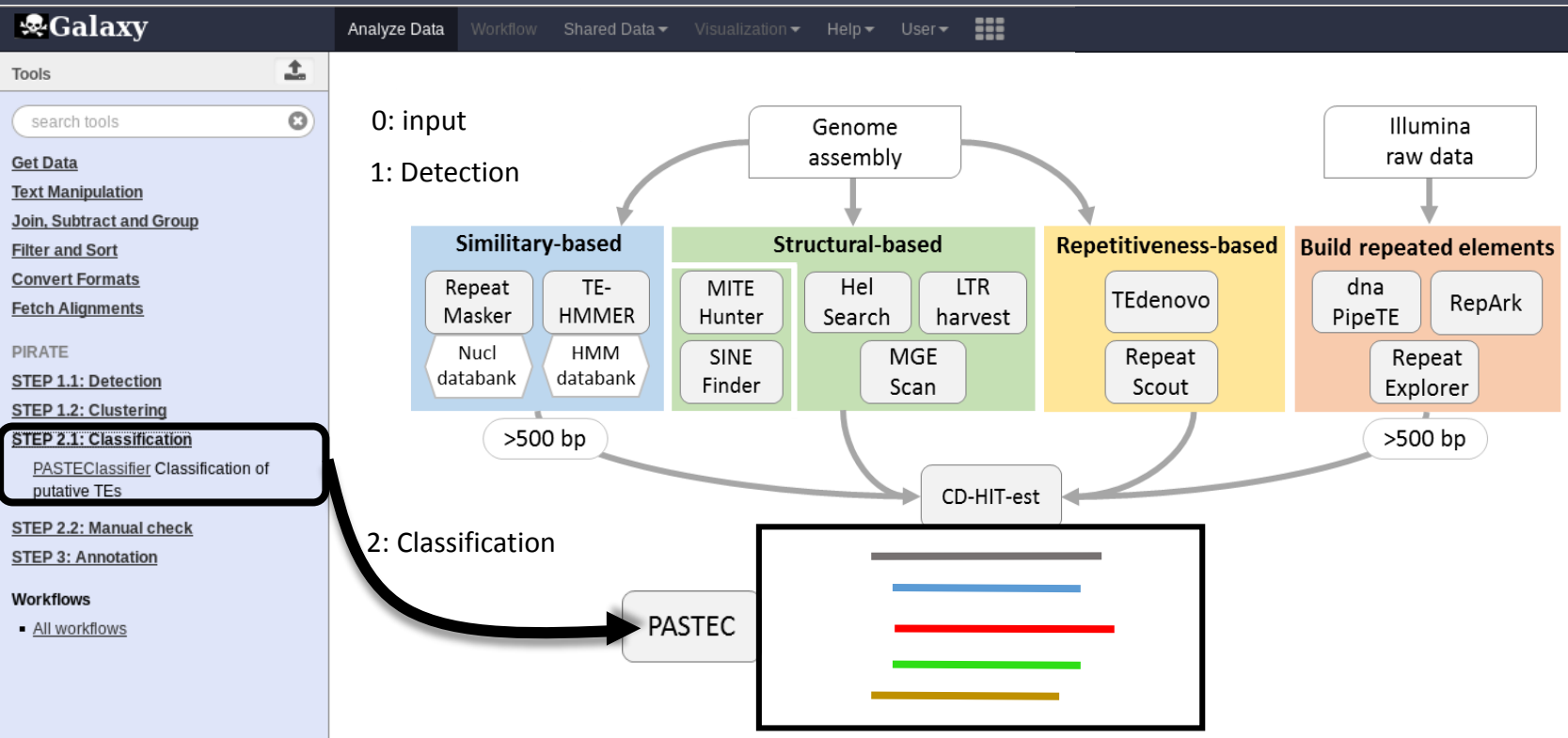
Cluster identical sequences
→ Remove redundancy
→ Only keep the larger putative TEs

PiRATE-Galaxy: Detection



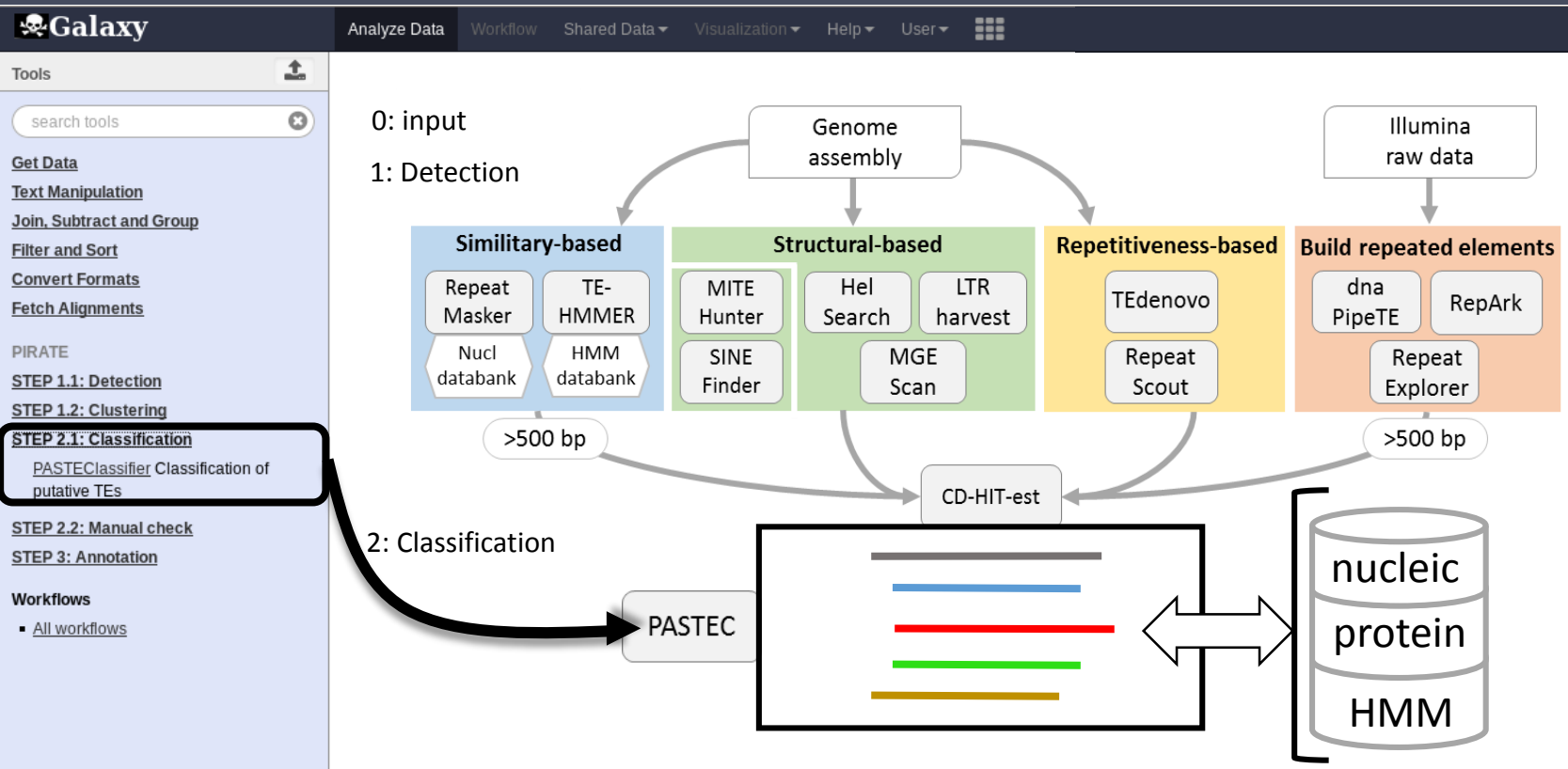
Cluster identical sequences
→ Remove redundancy
→ Only keep the larger putative TEs

PiRATE-Galaxy: Classification



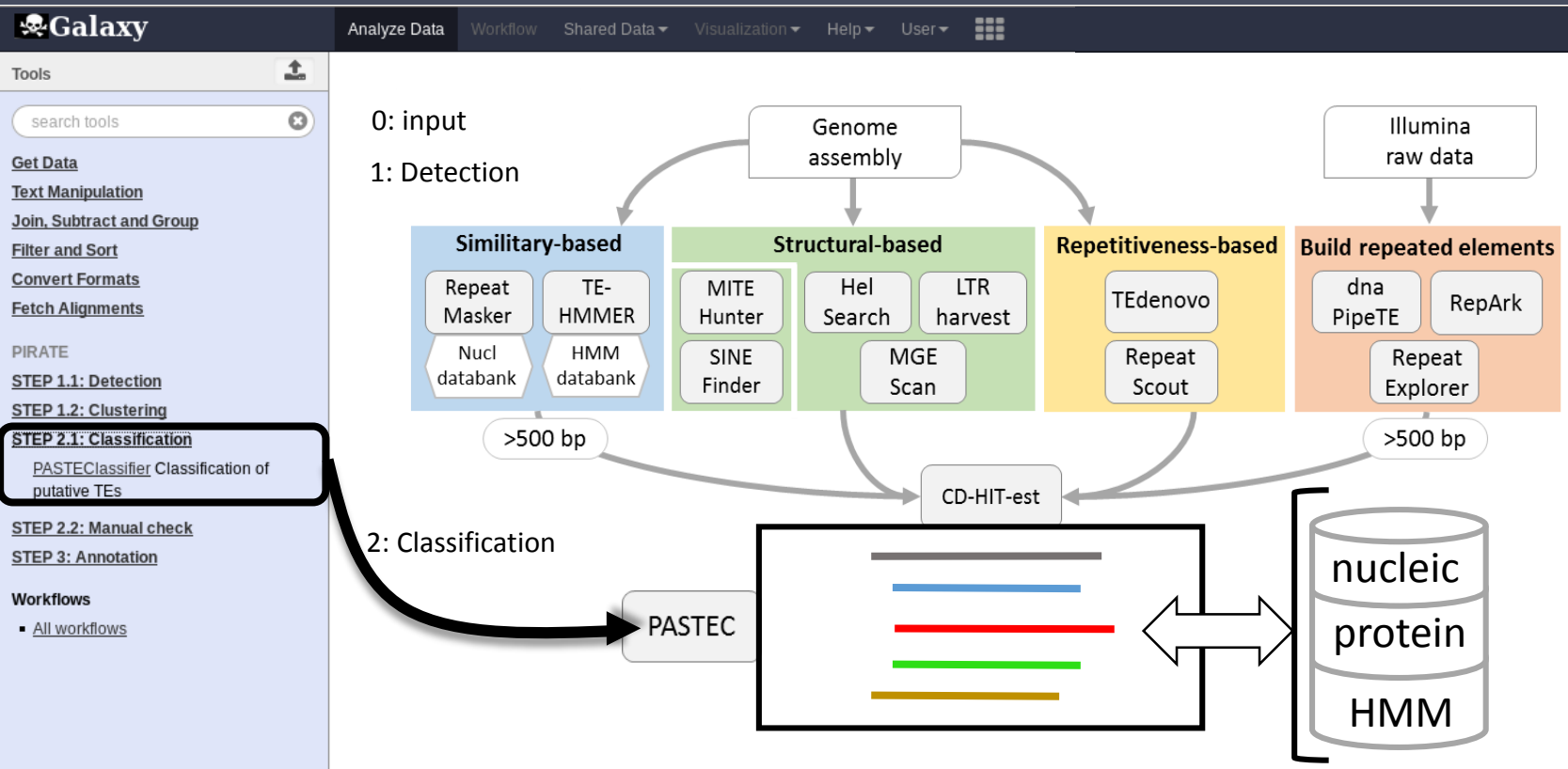
Automated classification

PiRATE-Galaxy: Classification



Automated classification
→ Using databanks

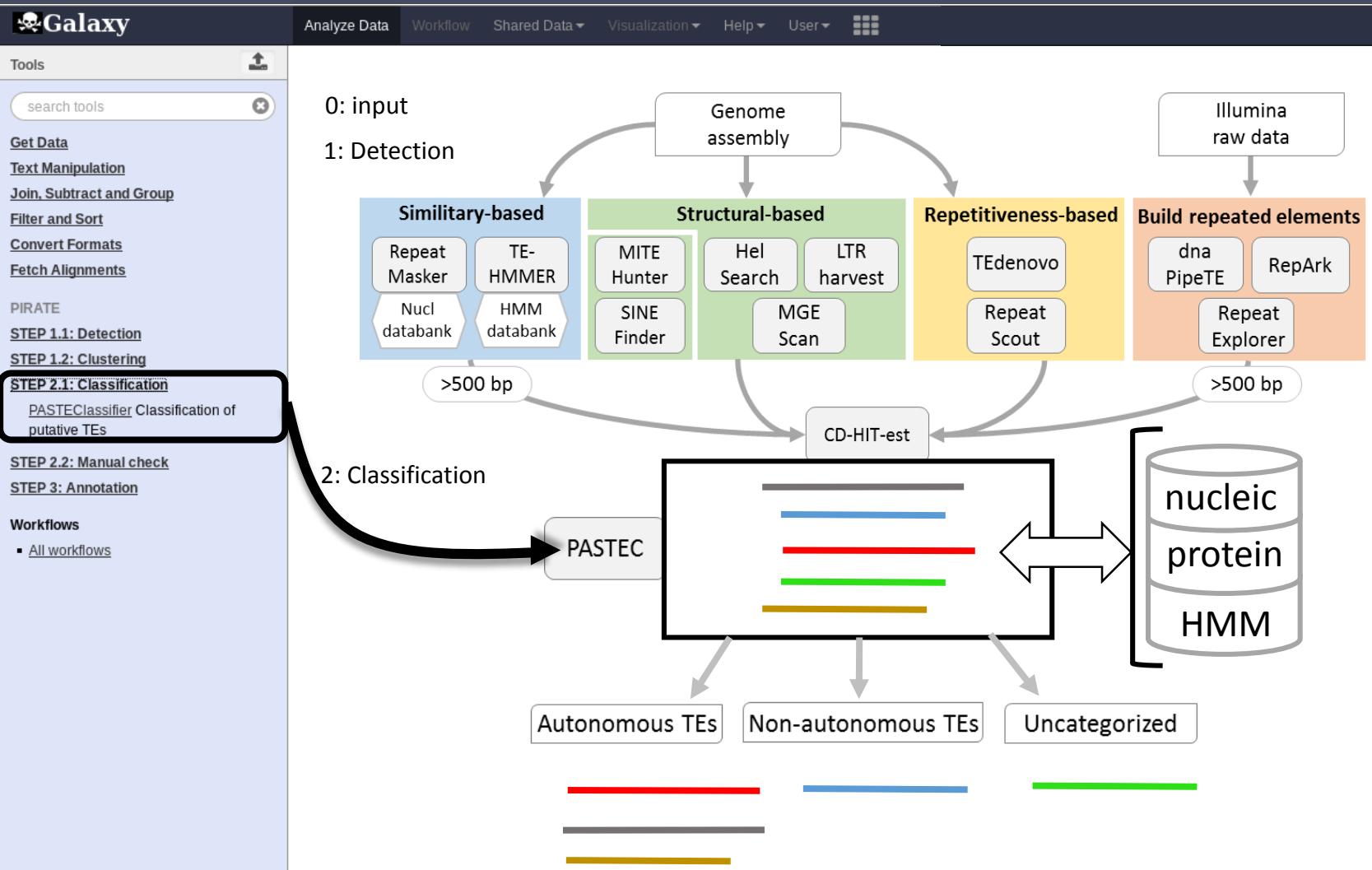
PiRATE-Galaxy: Classification



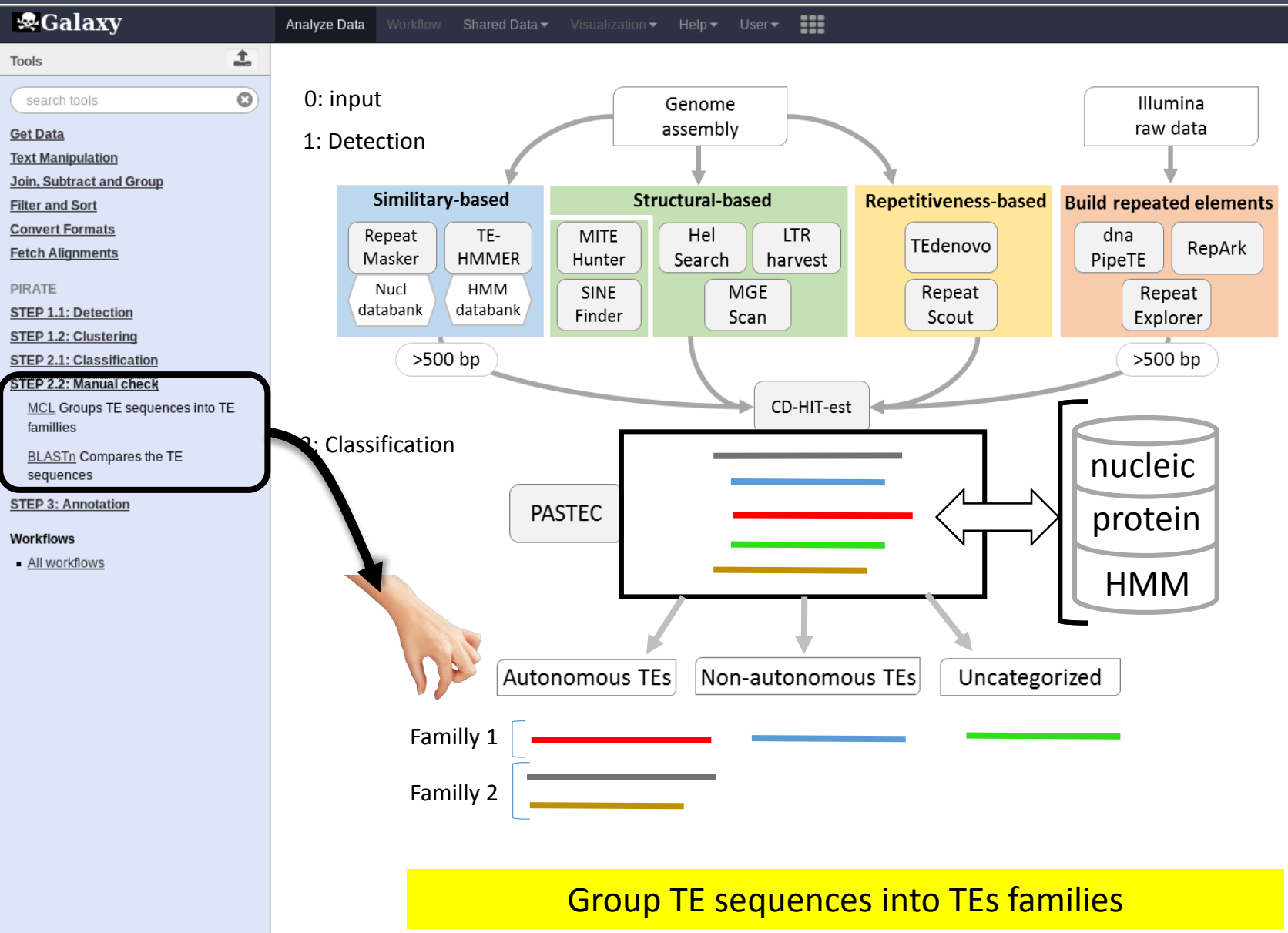
Automated classification
→ Using databanks

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

PiRATE-Galaxy: Classification

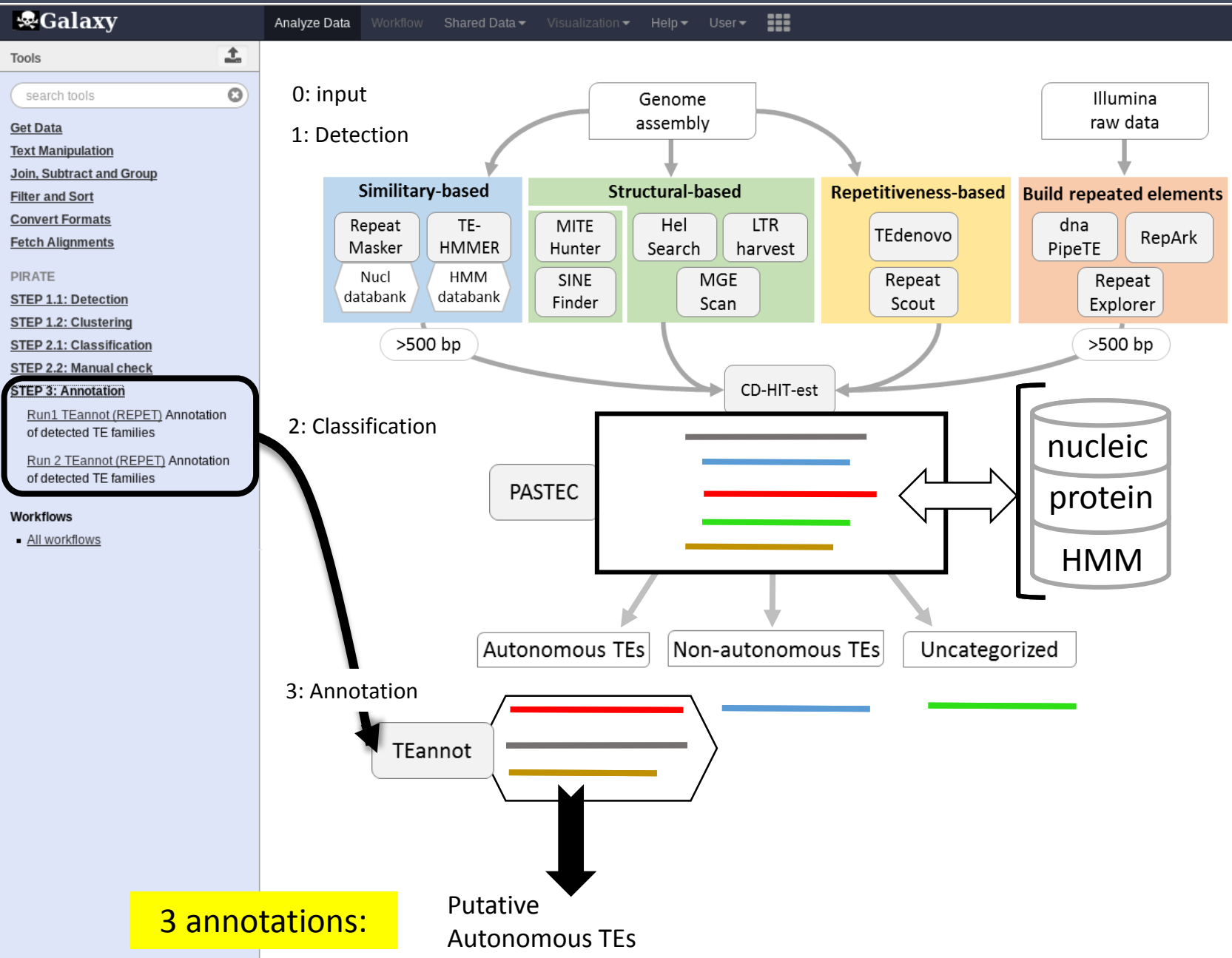


PiRATE-Galaxy: Classification



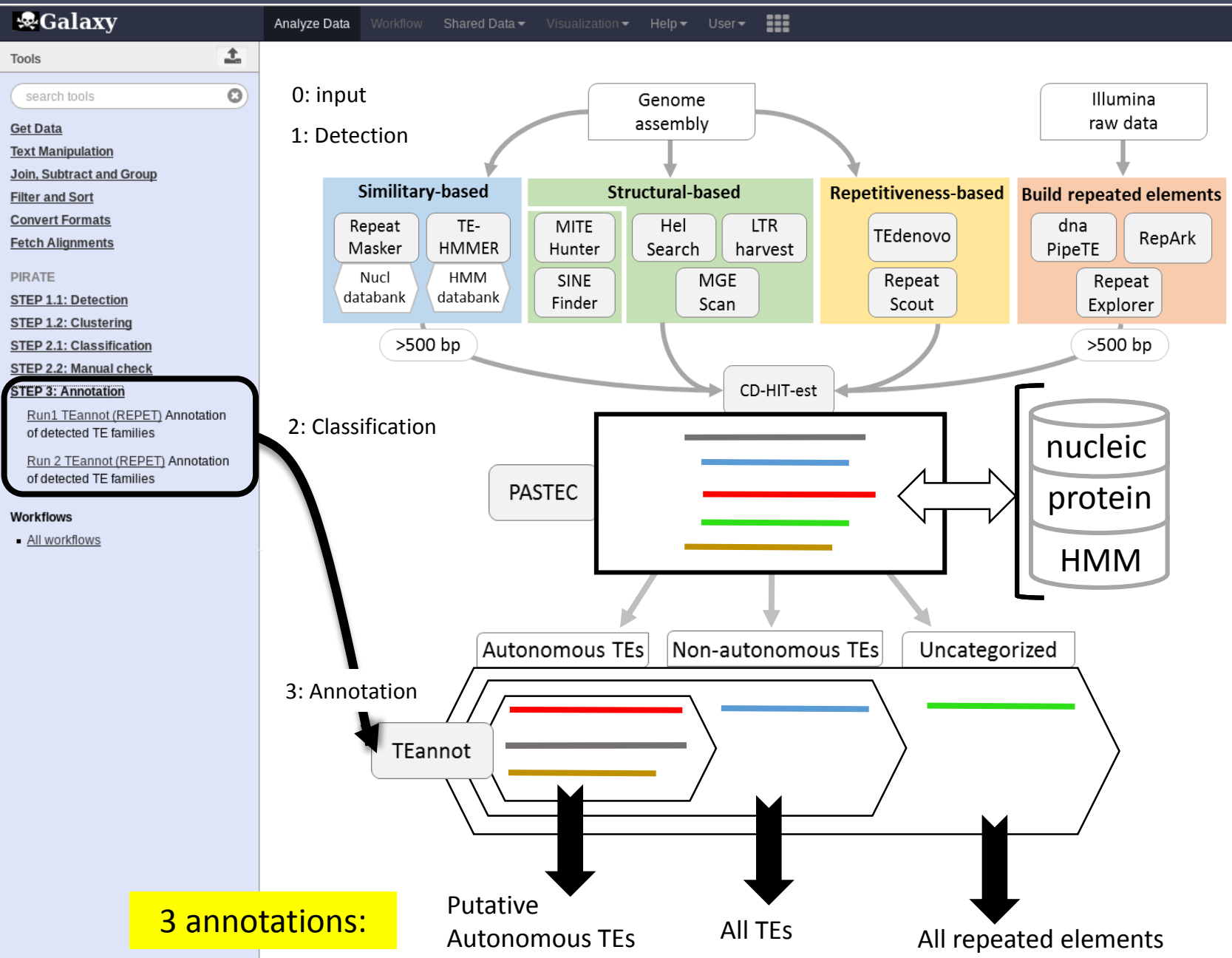
Group TE sequences into TEs families

PiRATE-Galaxy: Annotation



3 annotations:

PiRATE-Galaxy: Annotation



3 annotations:

Control with Arabidopsis thaliana

Arabidopsis thaliana

Genome size: **157 Mb**

259 TE families

- PiRATE detects **81**% of the TE families of *A. thaliana* (difficulty to detect non-autonomous TEs)

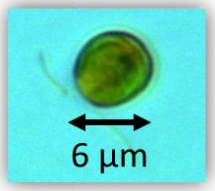
- **75**% of these detectes TEs were correctly classified (difficulty to classify non-autonomous TEs)

Berthelier et al. BMC Genomics ***Under Review***

Use of PiRATE with T. lutea (Haptophyta)

Example with the microalga *Tisochrysis lutea*

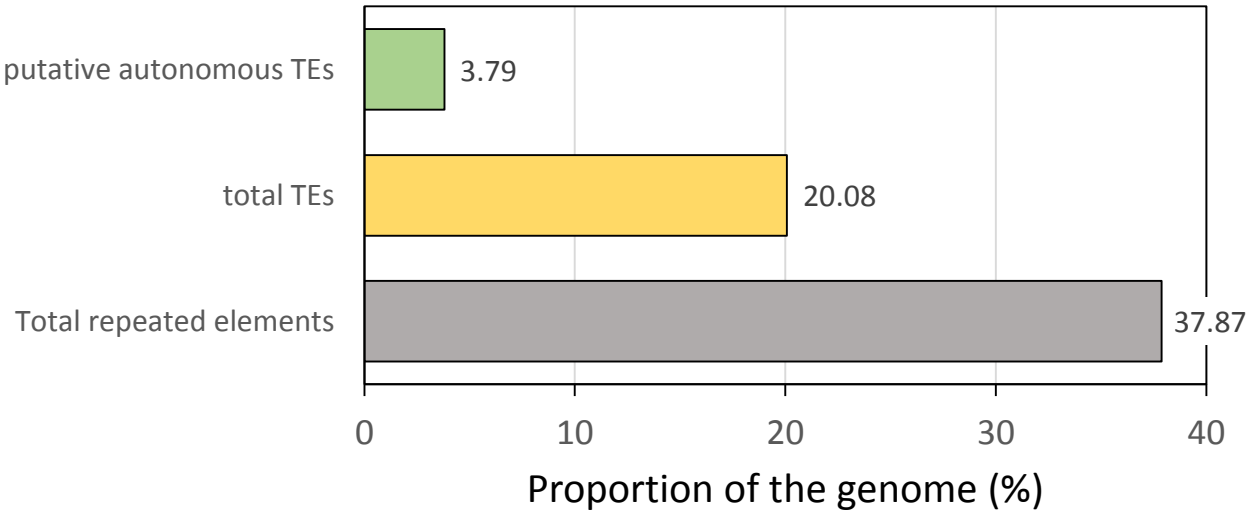
Phylum: Haptophyte



Tisochrysis lutea

Genome size: **82 Mb**
193 contigs

17 TE families
of Haptophytes

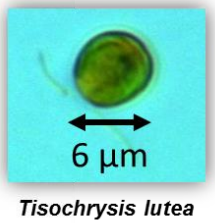


Berthelier et al. BMC Genomics ***Under Review***

Use of PiRATE with T. lutea (Haptophyta)

Example with the microalga *Tisochrysis lutea*

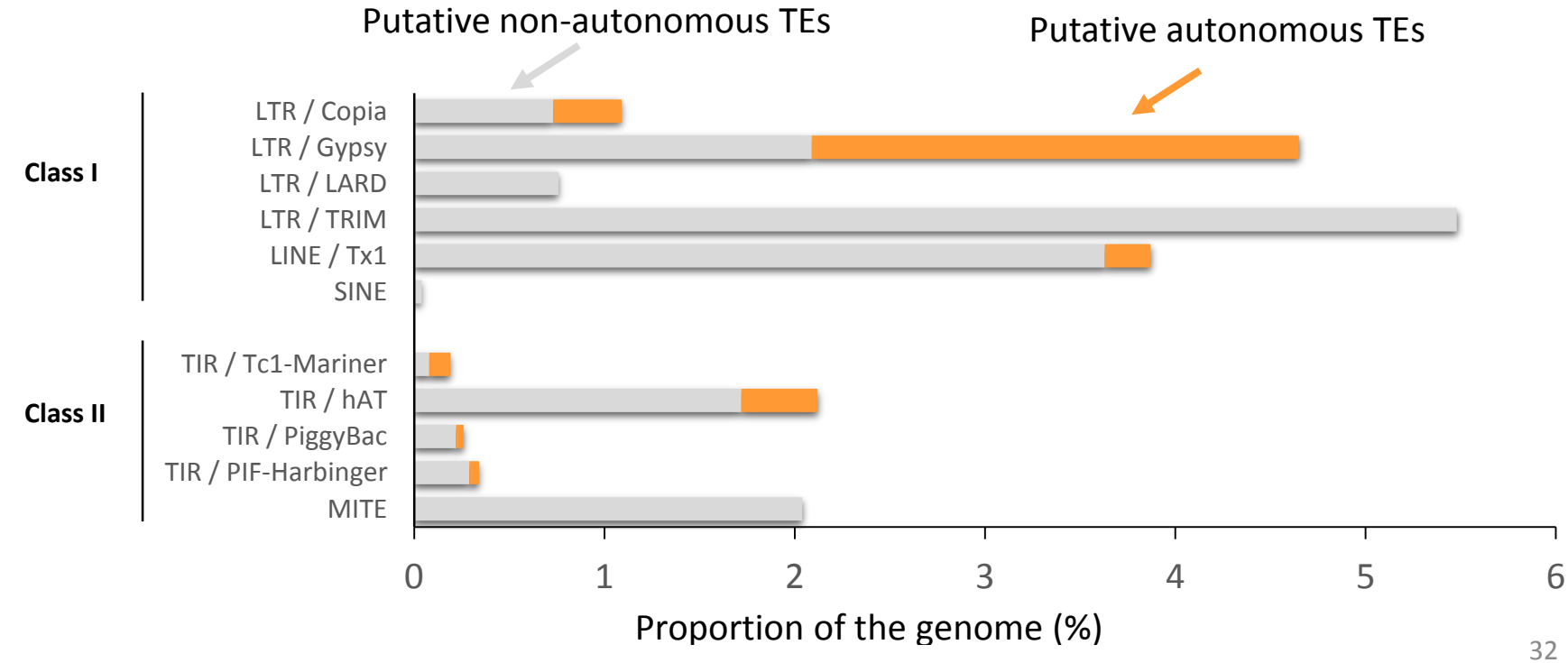
Phylum: Haptophyte



Tisochrysis lutea

Genome size: **82 Mb**
193 contigs

17 TE families
of Haptophytes



Berthelier et al. BMC Genomics ***Under Review***

Thanks for your attention!

Acknowledgements:

Grégory Carrier

Bruno Saint-Jean

Nathalie Casse

- Nicolas Daccord

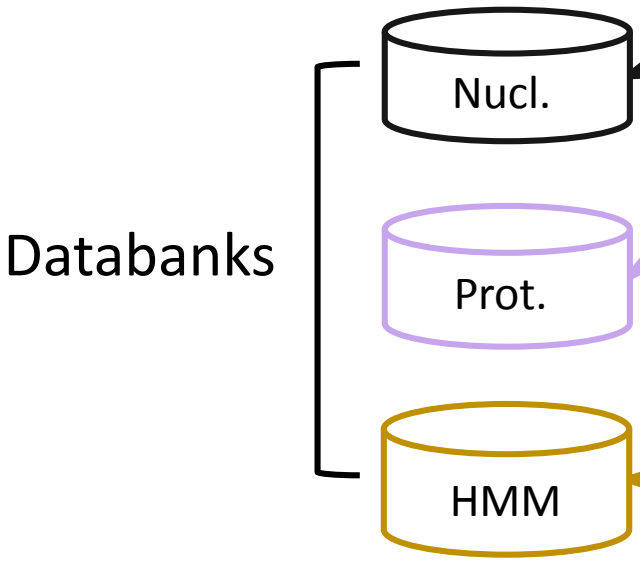
- Véronique Jamilloux

Funding:

How improve TE classification ?

PASTEClassifier

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



Public TE databanks

giri REPBASE

P-MITE

SINE Base

Non-inventoried TEs

NCBI

LITERATURE

Public profil HMM

Pfam

GyDB Gypsy Database 2.0

New profiles HMM

HMMER

done

done for microalgal genomes