Transposable Elements in the Era of Data Science

TE in plant genomes: structure, identification, and classification through bioinformatics and machine learning

Romain Guyot (IRD)

May 13th

Introduction

Importance of transposable elements

Major impact on genome size by contributing extensively to genomic expansion

Restructure genomes by causing duplications, inversions, and insertions that can influence gene regulation by altering expression patterns.

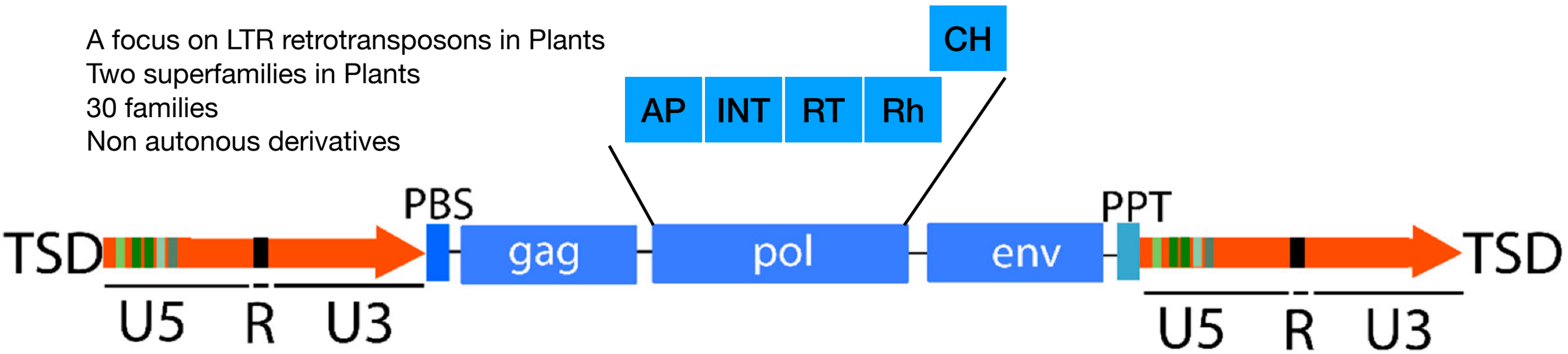
Drive genetic diversity and evolutionary innovation, contributing to phenotypes such as variegation, where mosaic gene activity leads to visible color patterns.

From an agronomic perspective, insertions have been linked to key traits making them crucial targets for crop breeding programs.

Introduction

Structure of LTR retrotransposons

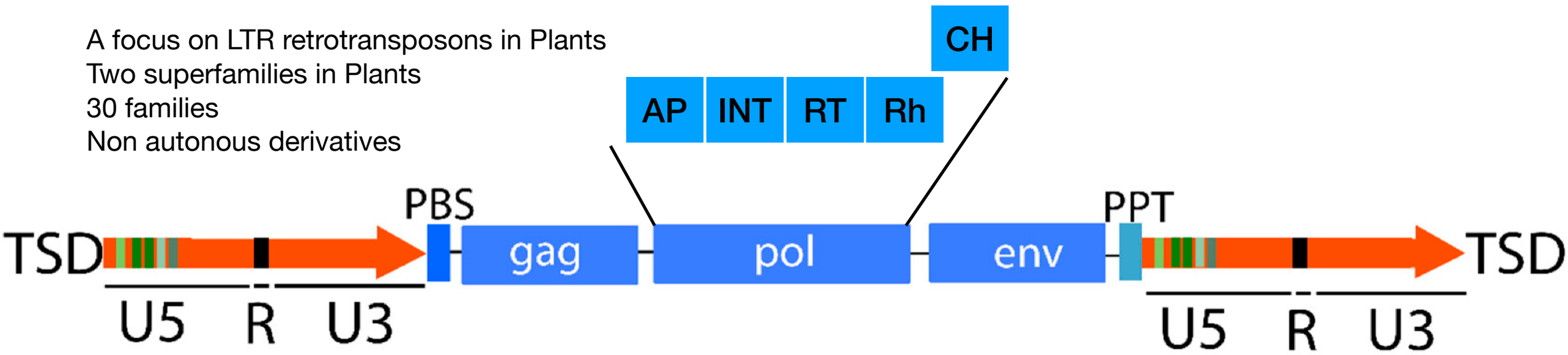
A focus on LTR retrotransposons in Plants
Two superfamilies in Plants
30 families
Non autonous derivatives

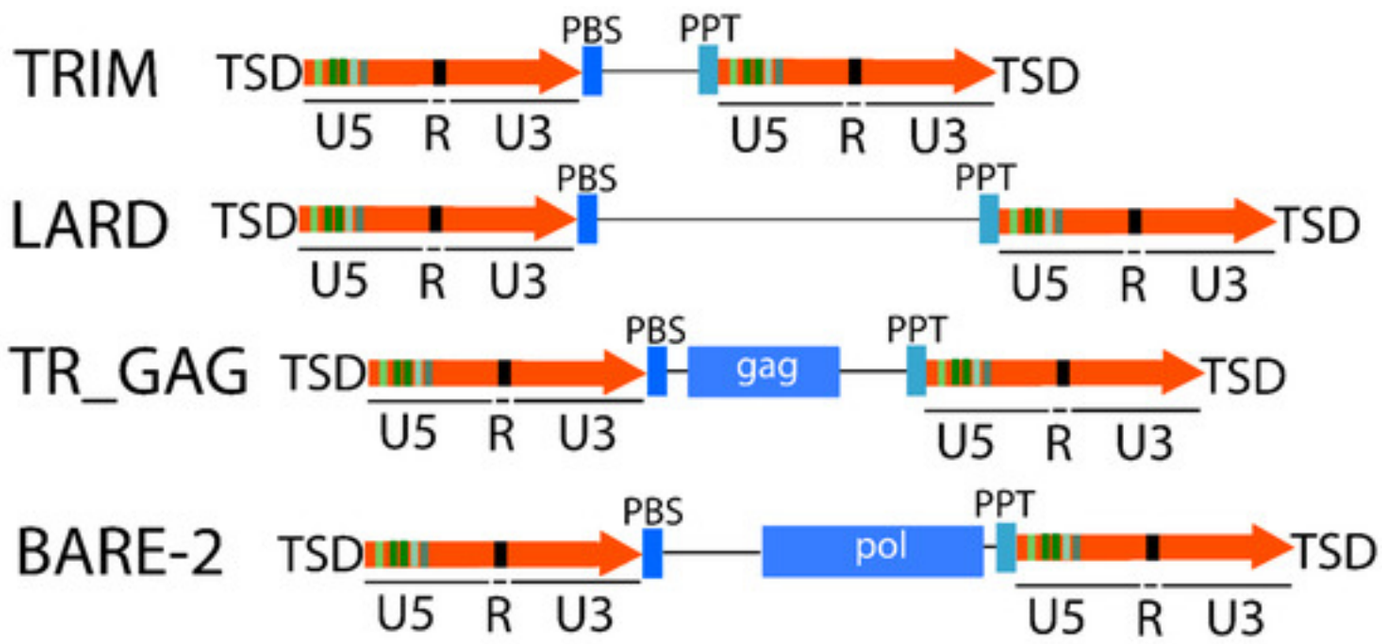


Introduction

Structure of LTR retrotransposons

A focus on LTR retrotransposons in Plants
Two superfamilies in Plants
30 families
Non autonous derivatives





Orozco-Arias el al., 2019

How to Identify and Classify Retrotransposons ?

Current Strategies and Methodologies

Different methods to identify and classify :

1_ Structure-Based Methods *LTRStruc/LTRharvest/LTRfinder — PASTEC*

2_ Homology-Based Methods RepeatMasker/Censor — *PASTEC TEsorter*

3_ *De novo RepeatModeler, REPET*

4_Comparative genomics

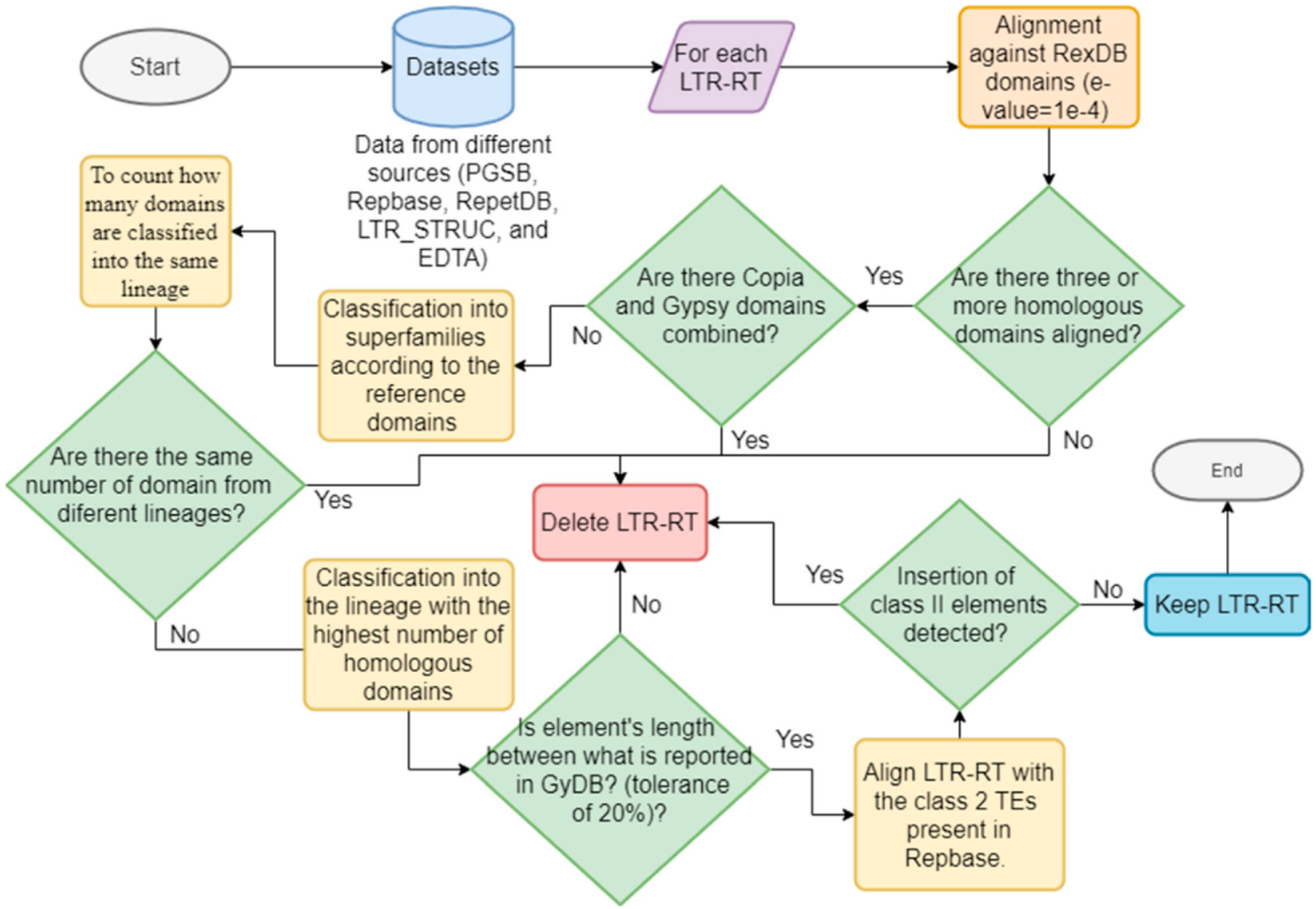
Challenges:

-Difficulties in constructing a representative and comprehensive library of TE sequences, since it depends on the sensibility and specificity of the bioinformatics programs used
-Nested elements, old and fragmented elements
-False identification of TEs (for example, large gene families).
-Difficulties in classifying non-autonomous elements
-Computational bottleneck due to the necessity to align sequences

How to Identify and Classify Retrotransposons ?

Kmer-based ML method to classify

Step_1 Obtain an expert-reviewed and comprehensive database of elements : InpactorDB https://inpactordb.github.io/



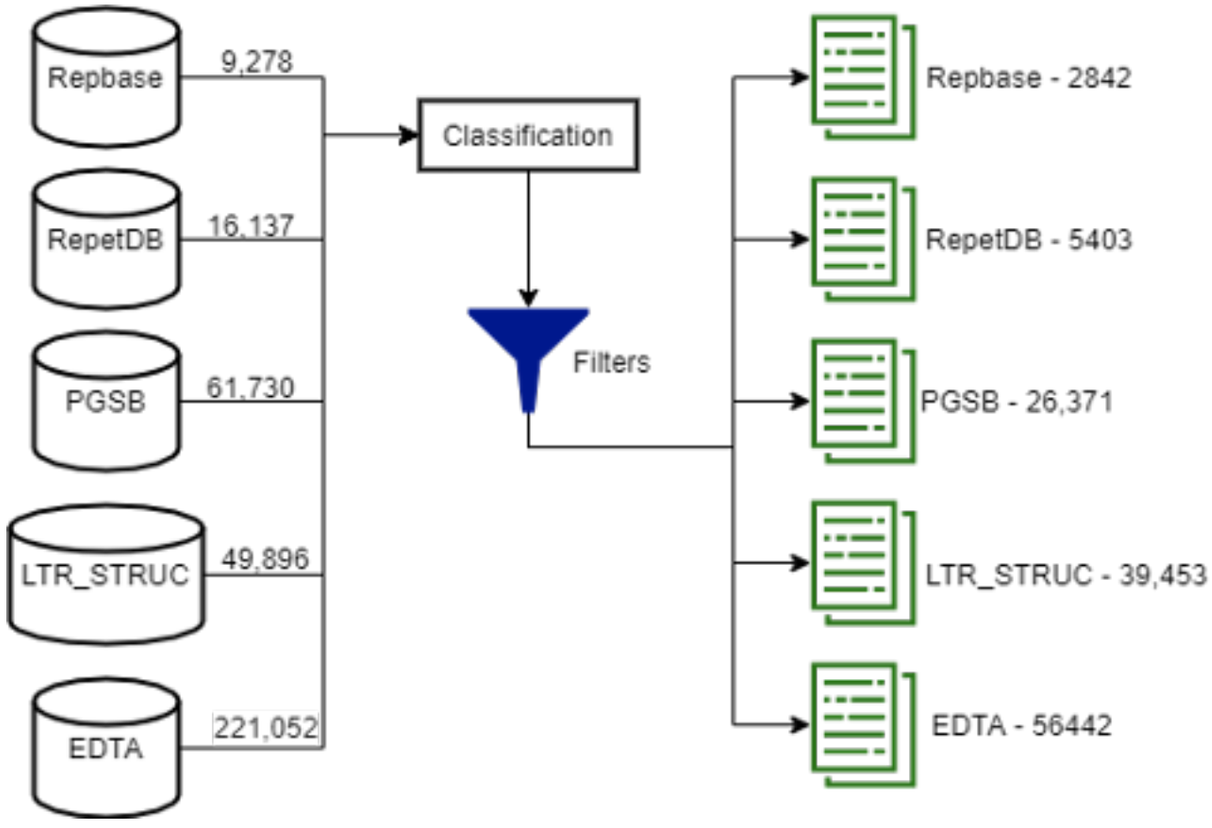
130,439 elements from195 plant genomes from NCBI classified into Superfamily and family

Orozco-Arias et al., 2021

How to Identify and Classify Retrotransposons ?

Kmer-based ML method to classify

Step_1 Obtain an expert-reviewed and comprehensive database of elements : InpactorDB https://inpactordb.github.io/



130,439 elements from195 plant genomes from NCBI classified into Superfamily and family

Orozco-Arias et al., 2021

How to Identify and Classify Retrotransposons ?

Kmer-based ML method to classify

Step_2 Select features explaining variations between families: **k-mer frequency** with k=1 to 6. Pre-processed data by scalling for methods based on distances (organized data) and by dimension reduction (simplified complex data) by PCA.

Step_3 Machine learning algorithmes used:

Binary classification (is a sequence a Retrotransposons?): Linear Support Vector Classifier (SVC), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Naive Bayesian Classifier (NB), Multi-Layer Perceptron (MLP), Decision Trees (DT), and Random Forest (RF)

Family classification : Linear Support Vector Classifier (SVC), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN)

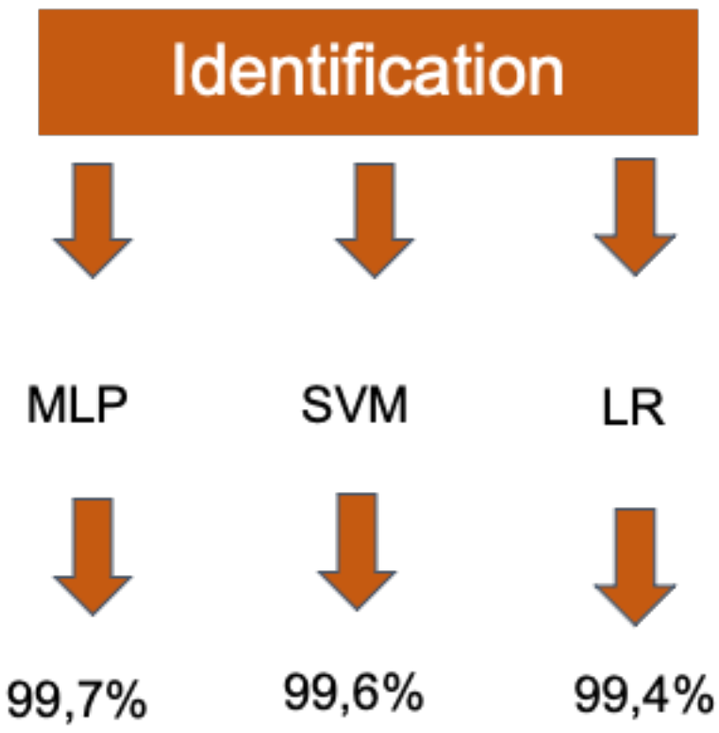
Orozco-Arias et al., 2021

How to Identify and Classify Retrotransposons ?

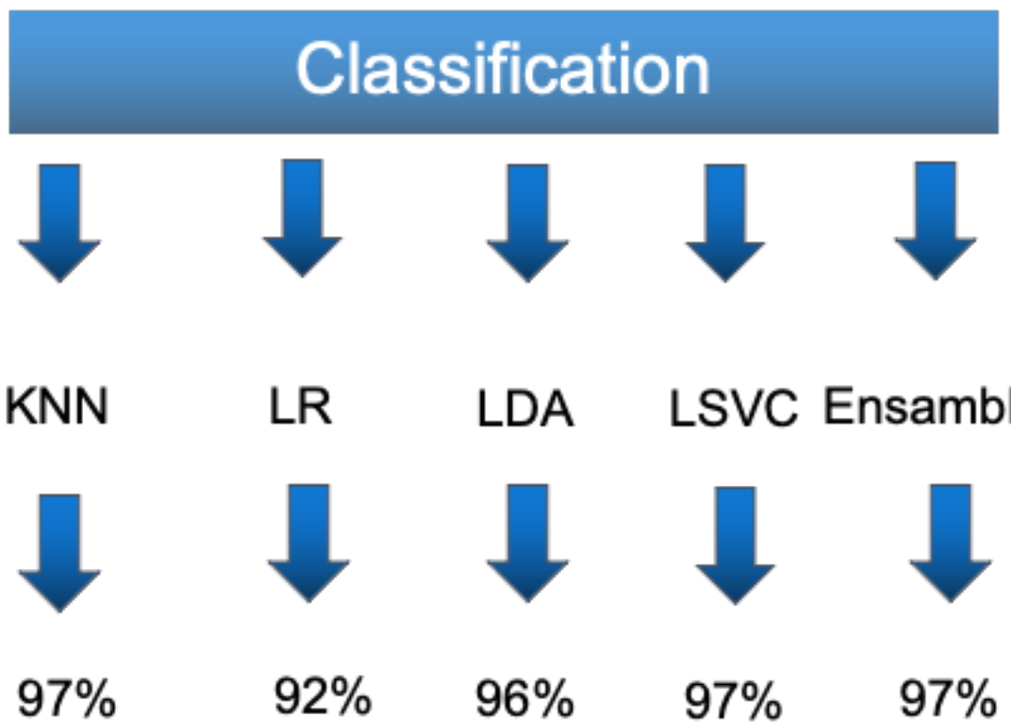
Kmer-based ML method to classify

Step_4 Performance Test:

Binary classification



Multi-class classification



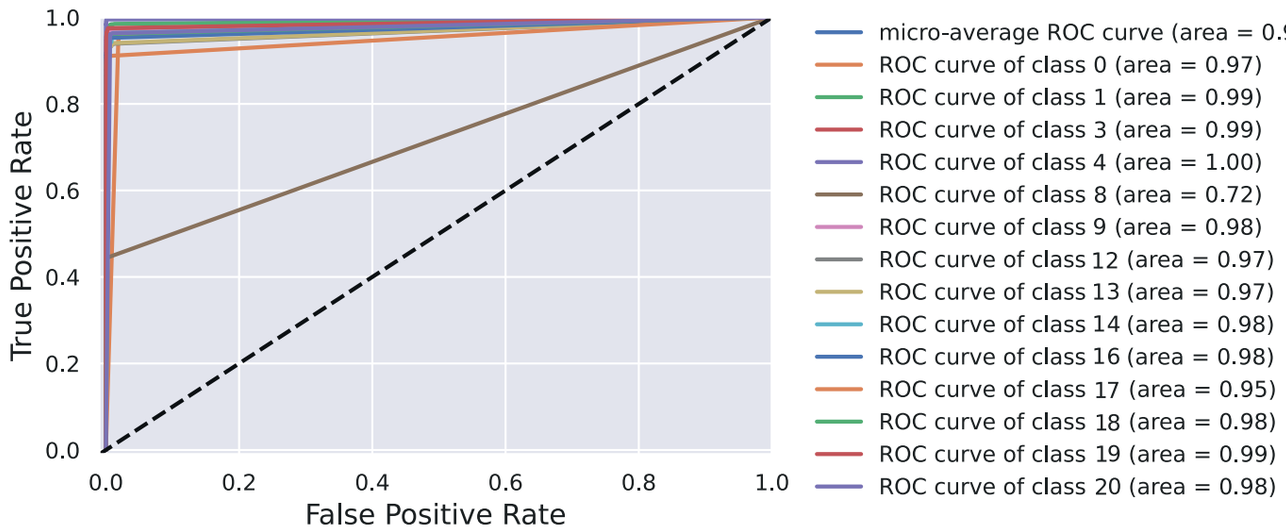
Orozco-Arias et al., 2021 Orozco-Arias. PhD thesis, 2022

How to Identify and Classify Retrotransposons ?

Kmer-based ML method to classify

Step_4 Performance Test: ROC curve

Class8 (Ikeros, Copia) and 16 (Galadriel, Gypsy) showed low performance.



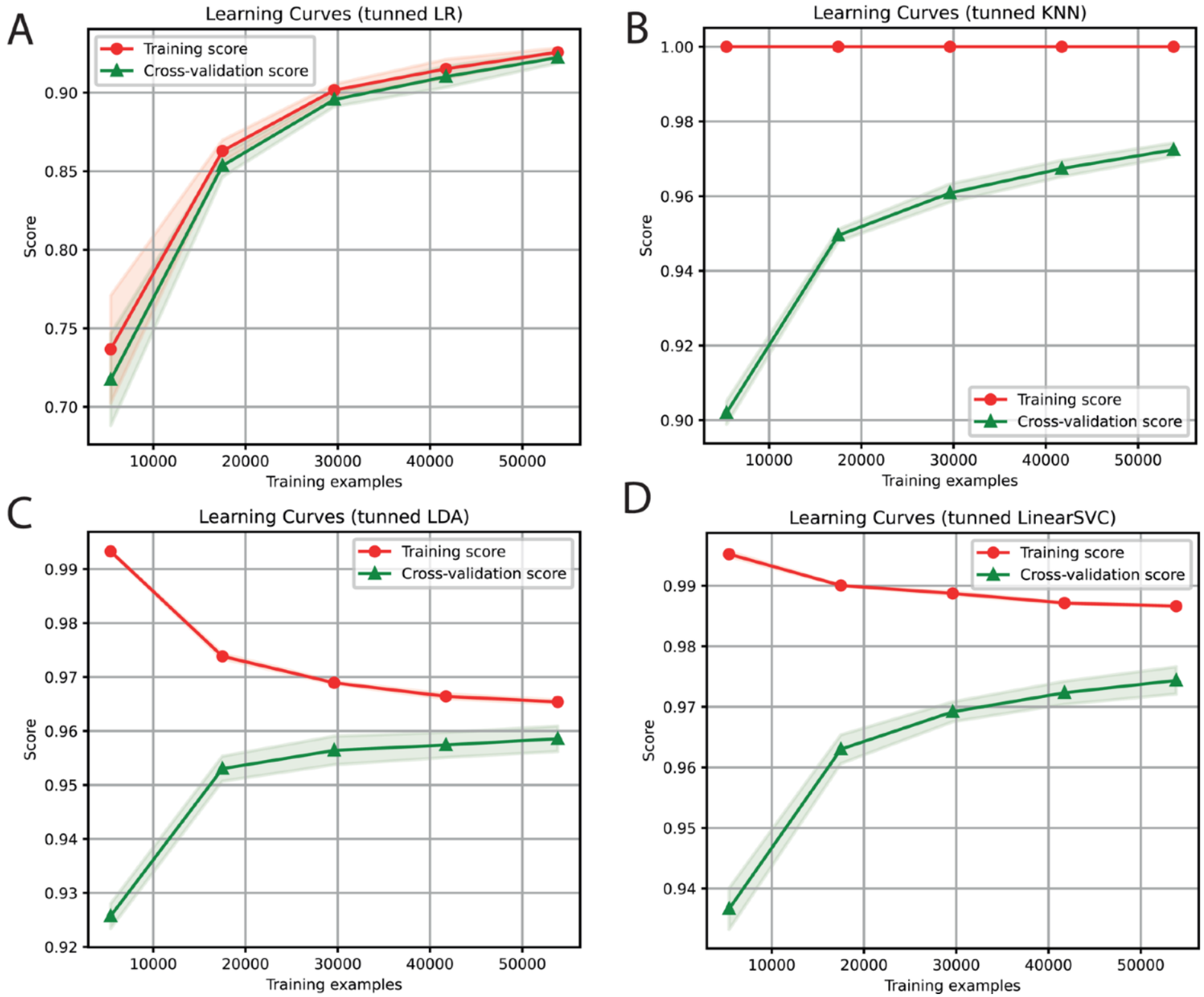
Class	Superfamily	Name	Number of classes
0	–	Other genomic features	34,823
1	Copia	ALE/Retrofit	12,031
3	Copia	Angela	1,458
4	Copia	Bianca	1,827
8	Copia	Ikeros	84
9	Copia	Ivana/Oryco	3,556
12	Copia	Tork/Tar	6,180
13	Copia	SIRE	3,130
		Total Copia	**28,266**
14	Gypsy	CRM	2,136
16	Gypsy	Galadriel	549
17	Gypsy	Reina	4,532
18	Gypsy	Tekay/DEL	10,396
19	Gypsy	Athila	3,499
20	Gypsy	TAT	17,927
		Total Gypsy	**39,039**

Orozco-Arias et al., 2021 Orozco-Arias. PhD thesis, 2022

How to Identify and Classify Retrotransposons ?

Kmer-based ML method to classify

Step_4 Performance Test

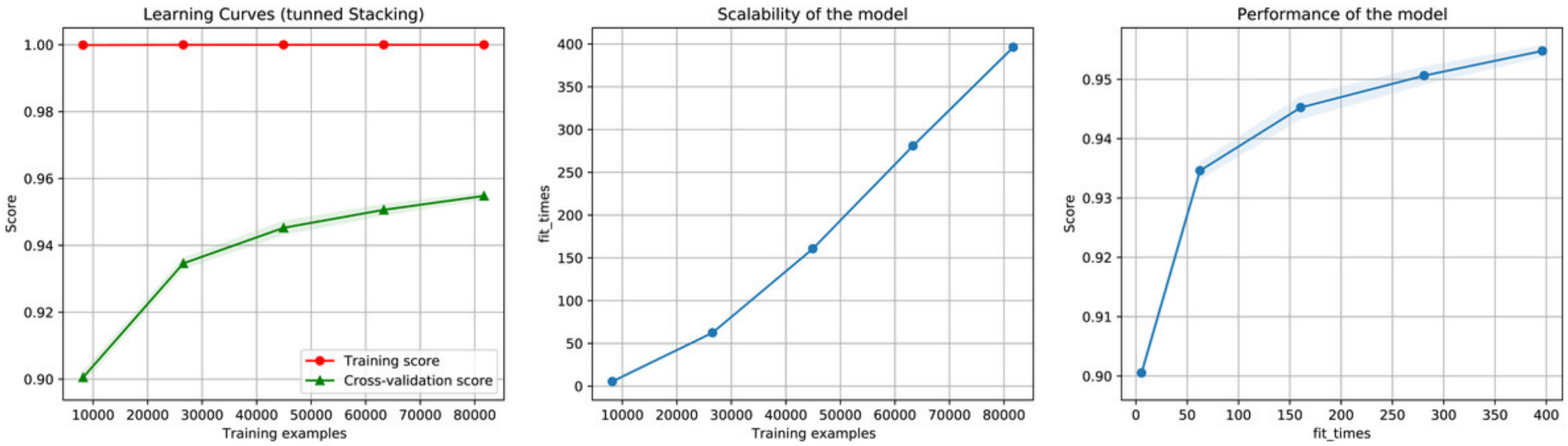


Orozco-Arias et al., 2021 Orozco-Arias. PhD thesis, 2022

How to Identify and Classify Retrotransposons ?

Kmer-based ML method to classify

Step_5 Determining relevant features (K-mer):
289 features out of 5,460 (5.29%) are relevant. The 10 most important features are: A, T, AAAAAA, ATAT, AGGGGG, CCCCC, TTTTTT, AGCT, GATC, GATGA but generally, increasing the length of k decreases the percentage of top selected features with greater importance



96% F1-Score). This task can be performed with only 289 *k*-mer frequencies, allowing low computational resources and time

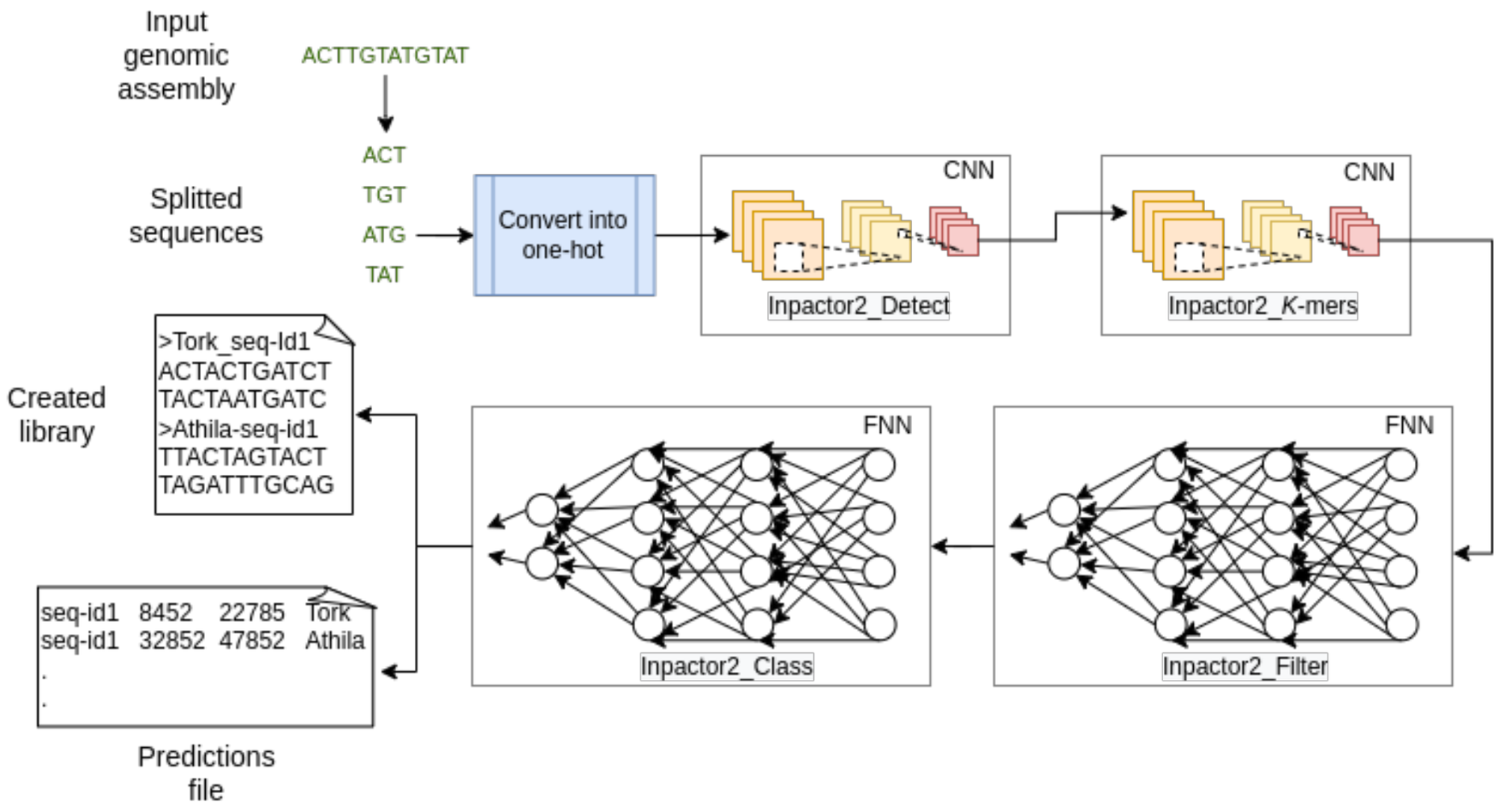
-> Developpemnt of an Automatic and free alignment tool: Inpactor2

Orozco-Arias et al., 2021 Orozco-Arias. PhD thesis, 2022

How to Identify and Classify Retrotransposons ?

Inpactor2

https://github.com/simonorozcoarias/Inpactor2

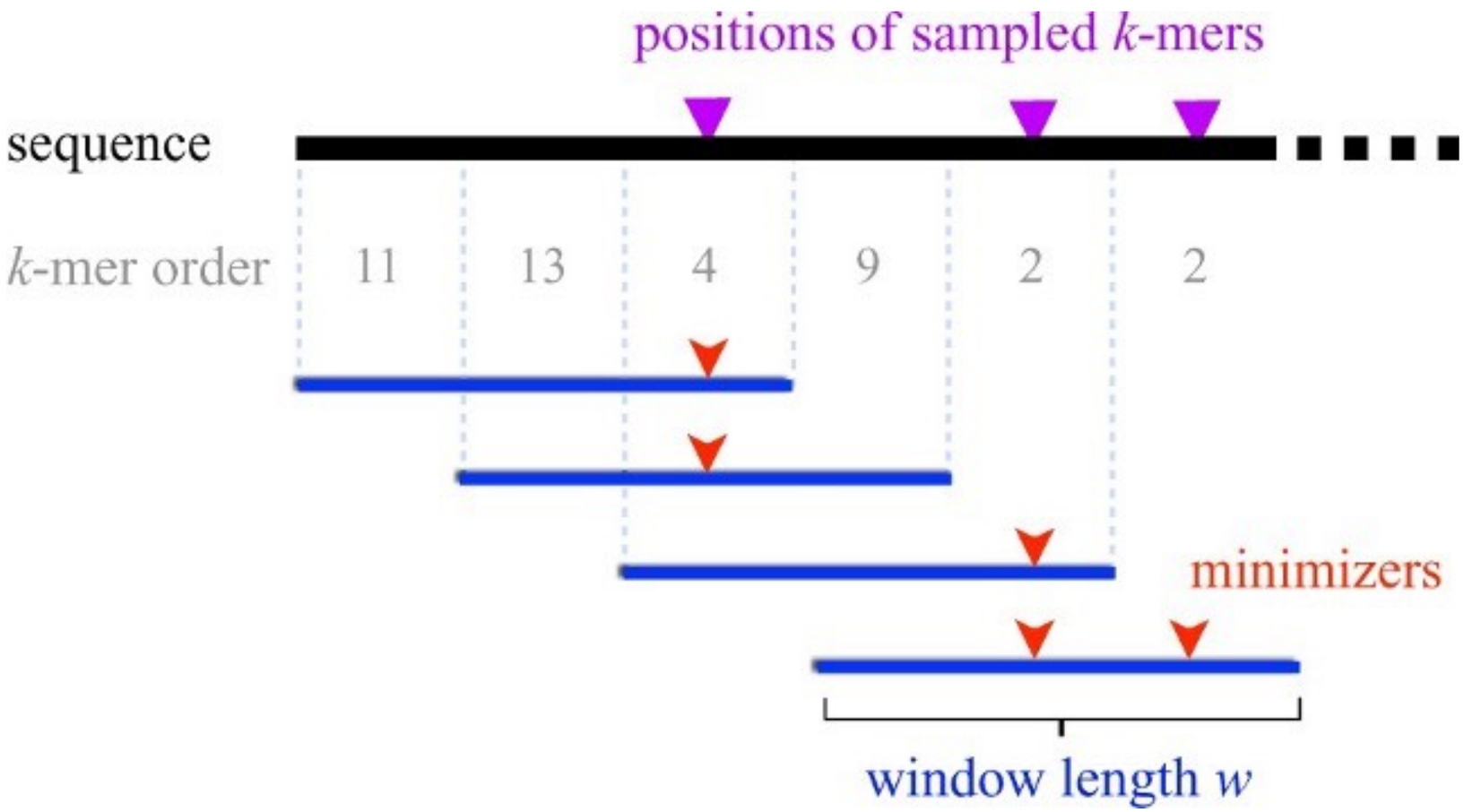


Orozco-Arias et al., 2023

How to Identify and Classify Retrotransposons ?

Alignment-free method to replace RM

Replace the alignment step of RepeatMasker by the search of minimizers



Jain et al., 2020

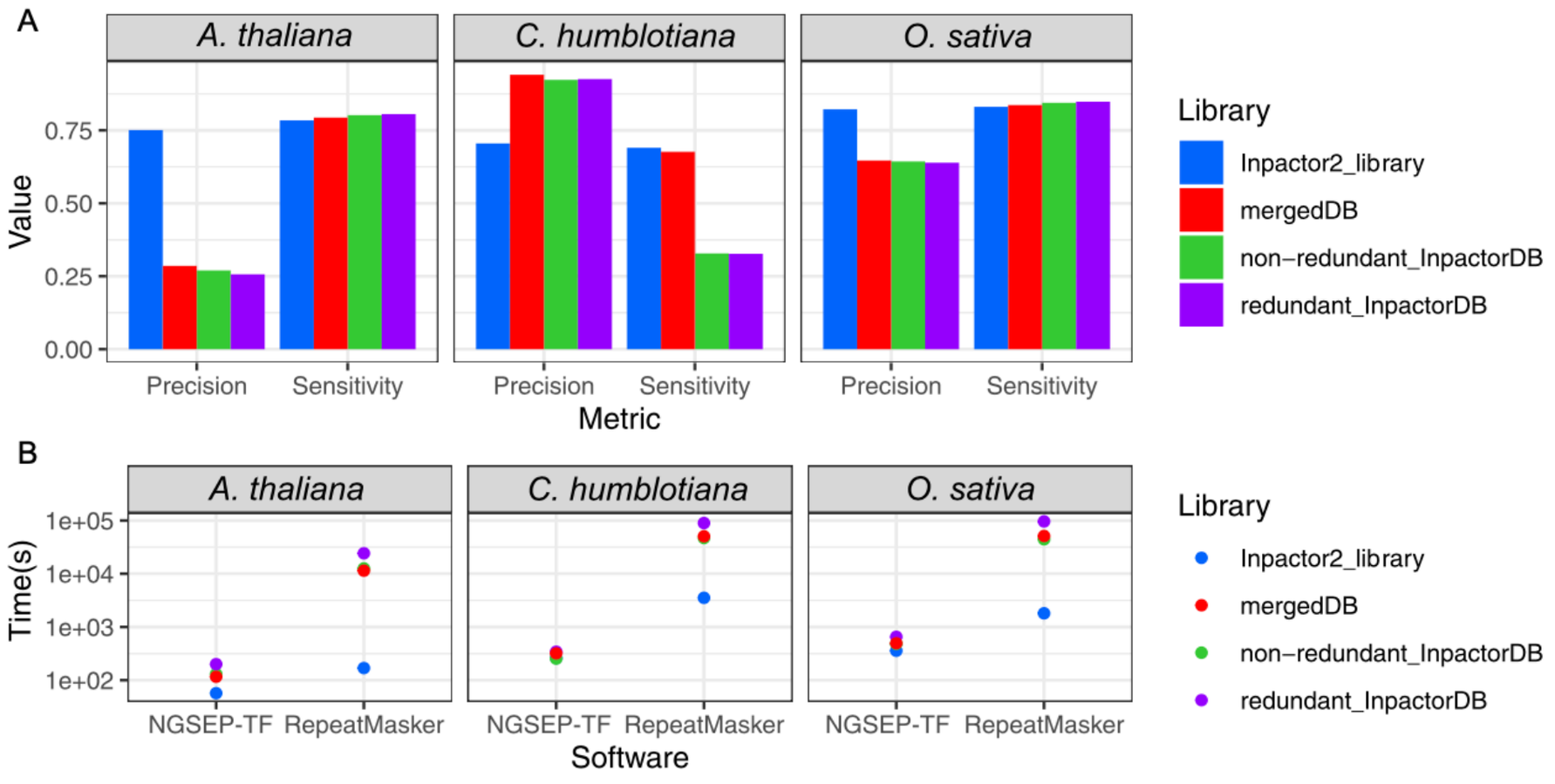
A minimizer is the smallest *k*-mer selected within a sliding window to represent a part of a sequence compactly.

How to Identify and Classify Retrotransposons ?

Alignment-free method to replace RM

Replace the alignment step of RepeatMasker by the search of minimizers : Transposon Finder (NGSEP)

https://github.com/NGSEP/NGSEPcore



2 to 20 times faster than RepeatMasker

Up to 75% precision and 70% sensitivity

Jain et al., 2020

Wheat genome annotation:
Less than 2 hours per chromosome using less than 250 GB RAM

Gonzalez-Garcia et al., 2023

Conclusion

Alignment free approaches and ML:
_ improves annotation and classification of TE
_ improves speed of computation
_ low accuracy with fragmented and degenerated TEs

Future needs :
_ high quality training dataset