Transposable Elements in the Era of Data Science

TE classification systems

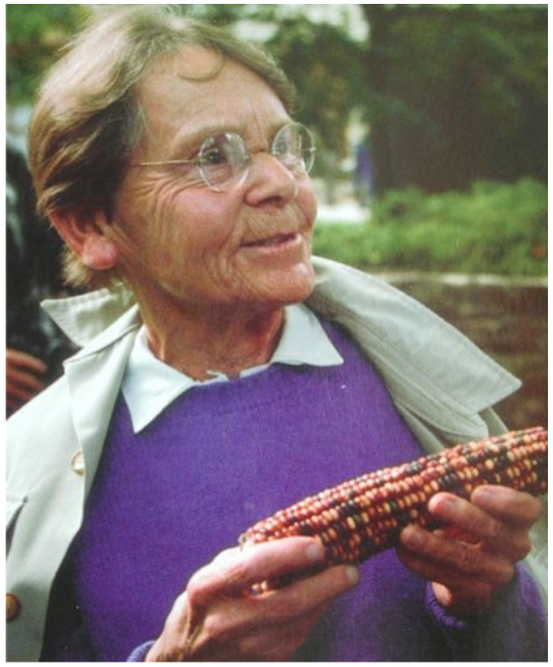
Anna-Sophie Fiston-Lavier (ISEM-U. Montpellier) & Romain Guyot (IRD)

May 12th 2025

Introduction

what are transposable elements?

Barbara McClintock (1940s–50s), when working on Maize transposable elements were called « Controlling elements ». According to Barbara McClintock, the elements had a regulatory effect on gene expression and acted as dynamic regulators of genetic information.



"Controlling elements are segments of DNA that can move (transpose) from one location to another within the genome, and in doing so, they can alter the expression of genes. »

Barbara McClintock, Nobel Lecture, 1983

"The genome is a highly sensitive organ of the cell that monitors genomic activities and corrects common errors, senses unusual and unexpected events, and responds to them, often by restructuring itself."

Modern definition

Transposable elements are DNA sequences that can change their position within a genome, sometimes creating or reversing mutations and altering the genome's size. »

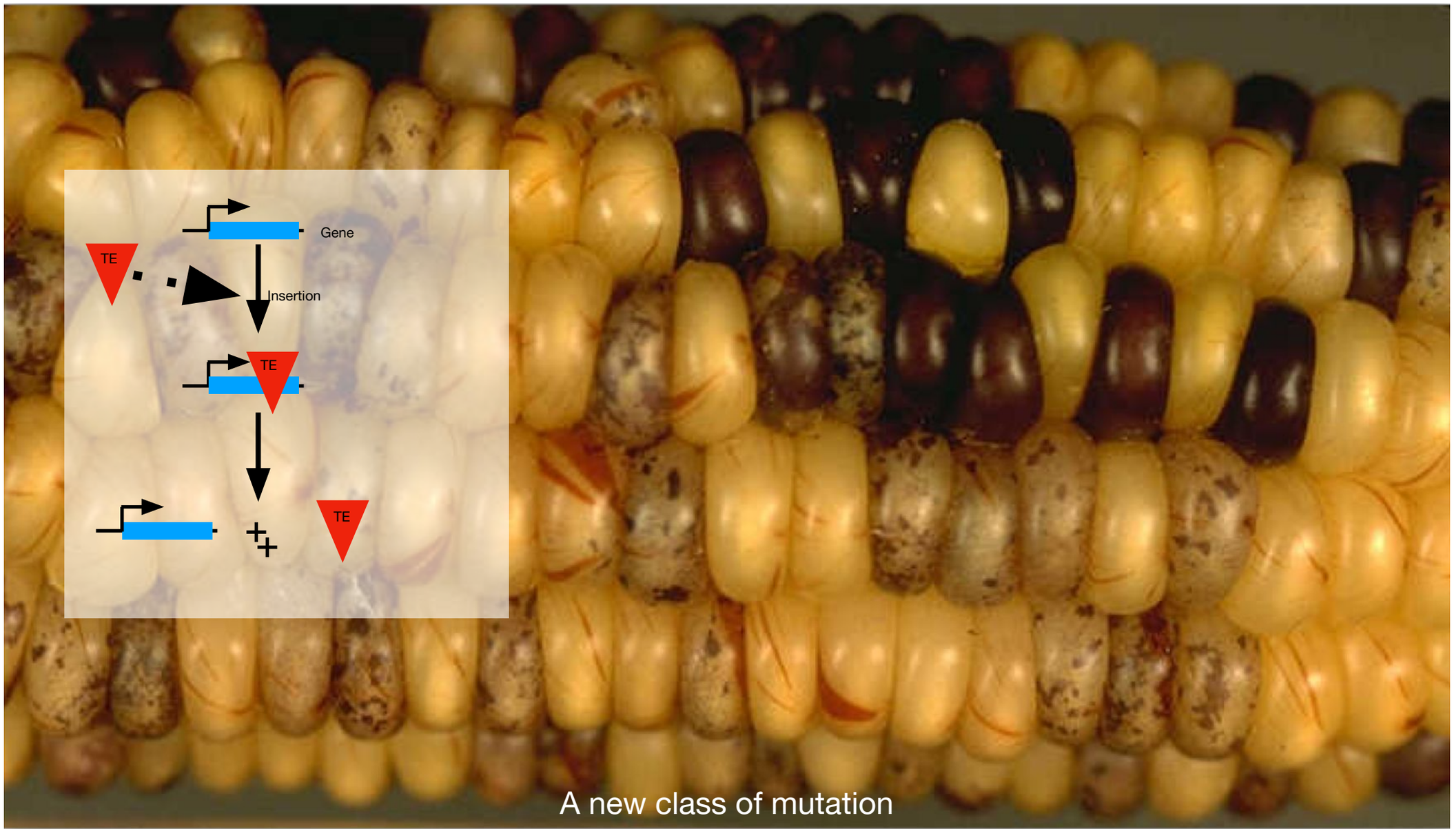
Introduction

brief historical background

In 1940s, Barbara McClintock showed that certain DNA fragments, termed transposons, can be activated to “jump” from one position on a chromosome to another. She hypothesized that transposition provides a means to rapidly organize genes in response to environmental stress.

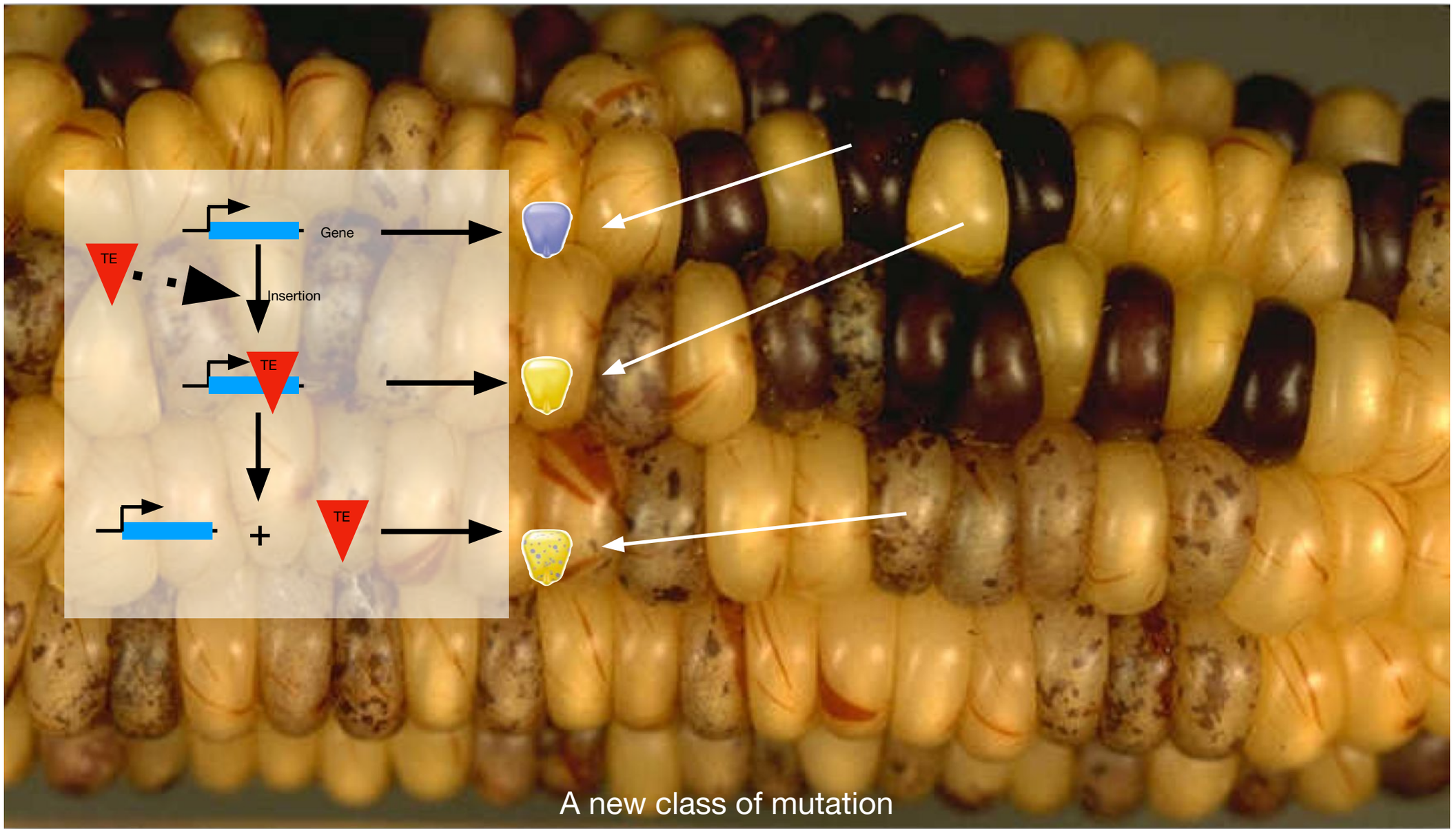
Introduction

brief historical background



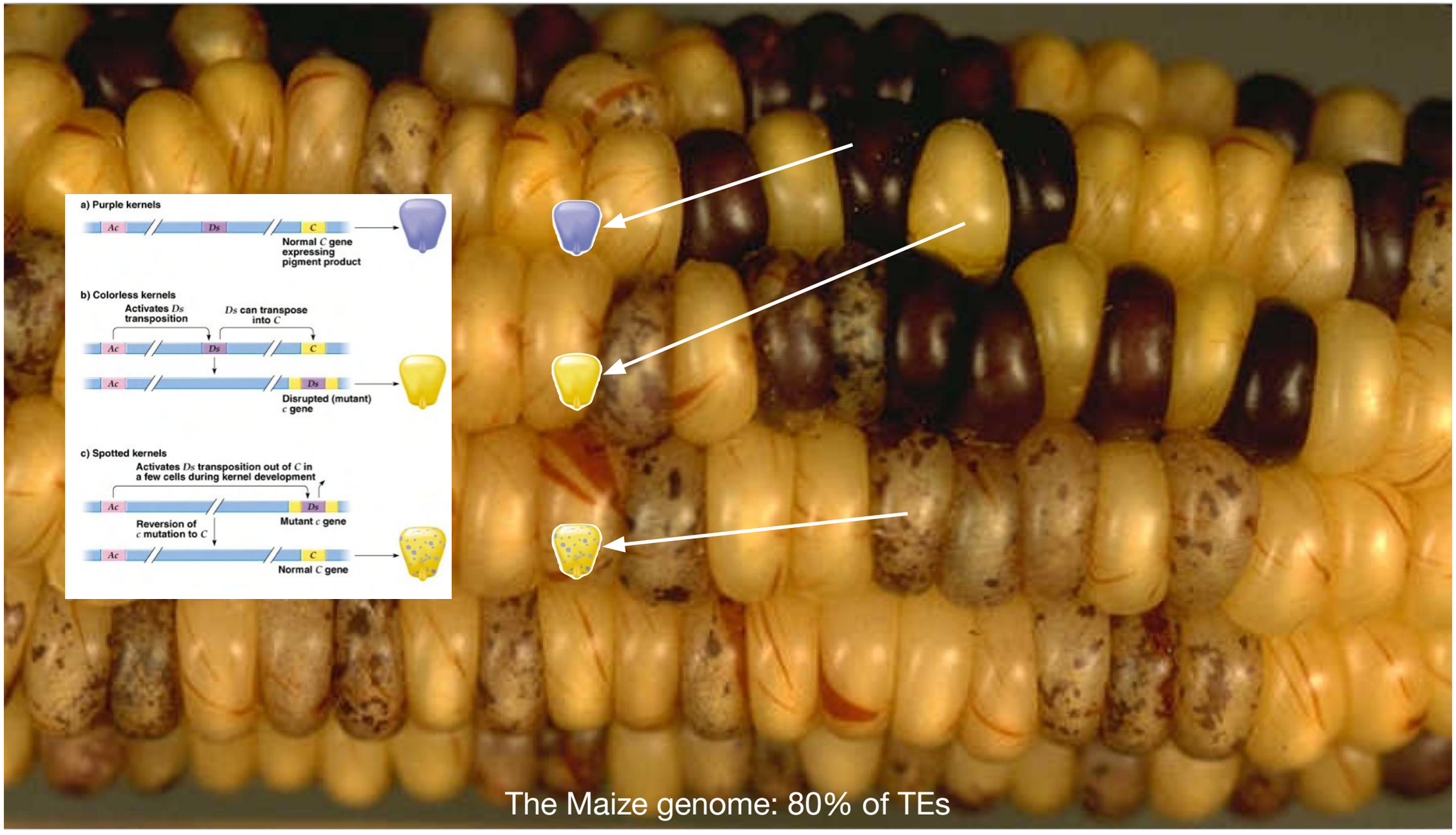
Introduction

brief historical background



Introduction

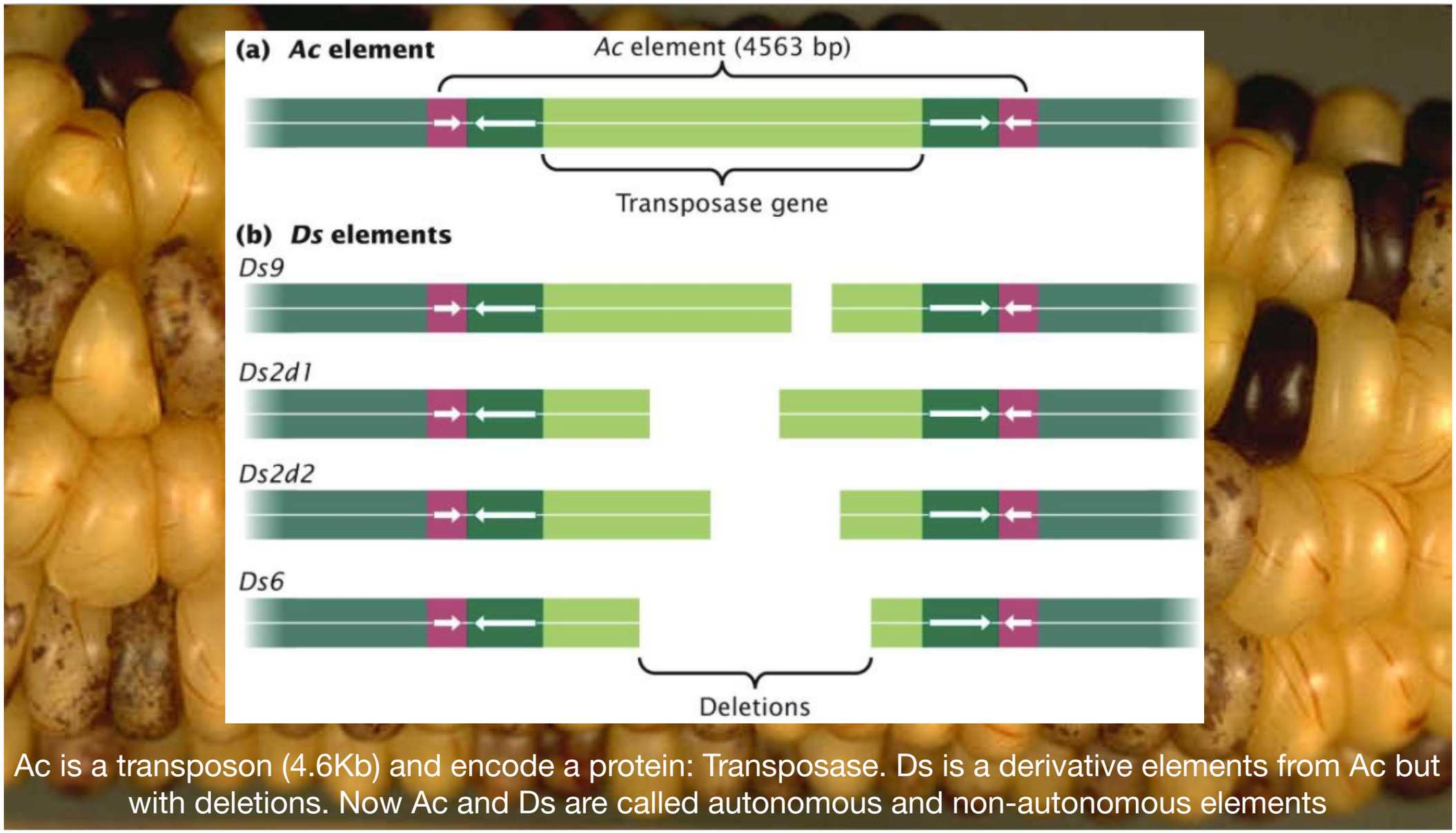
brief historical background



Adapted from "Molecular Biology of the Gene" by Watson et al. 6th edition

Introduction

brief historical background



Ac is a transposon (4.6Kb) and encode a protein: Transposase. Ds is a derivative elements from Ac but with deletions. Now Ac and Ds are called autonomous and non-autonomous elements

Adapted from "Molecular Biology of the Gene" by Watson et al. 6th edition

Introduction

why classify TEs ?

Understand genome structure and dynamics at different levels

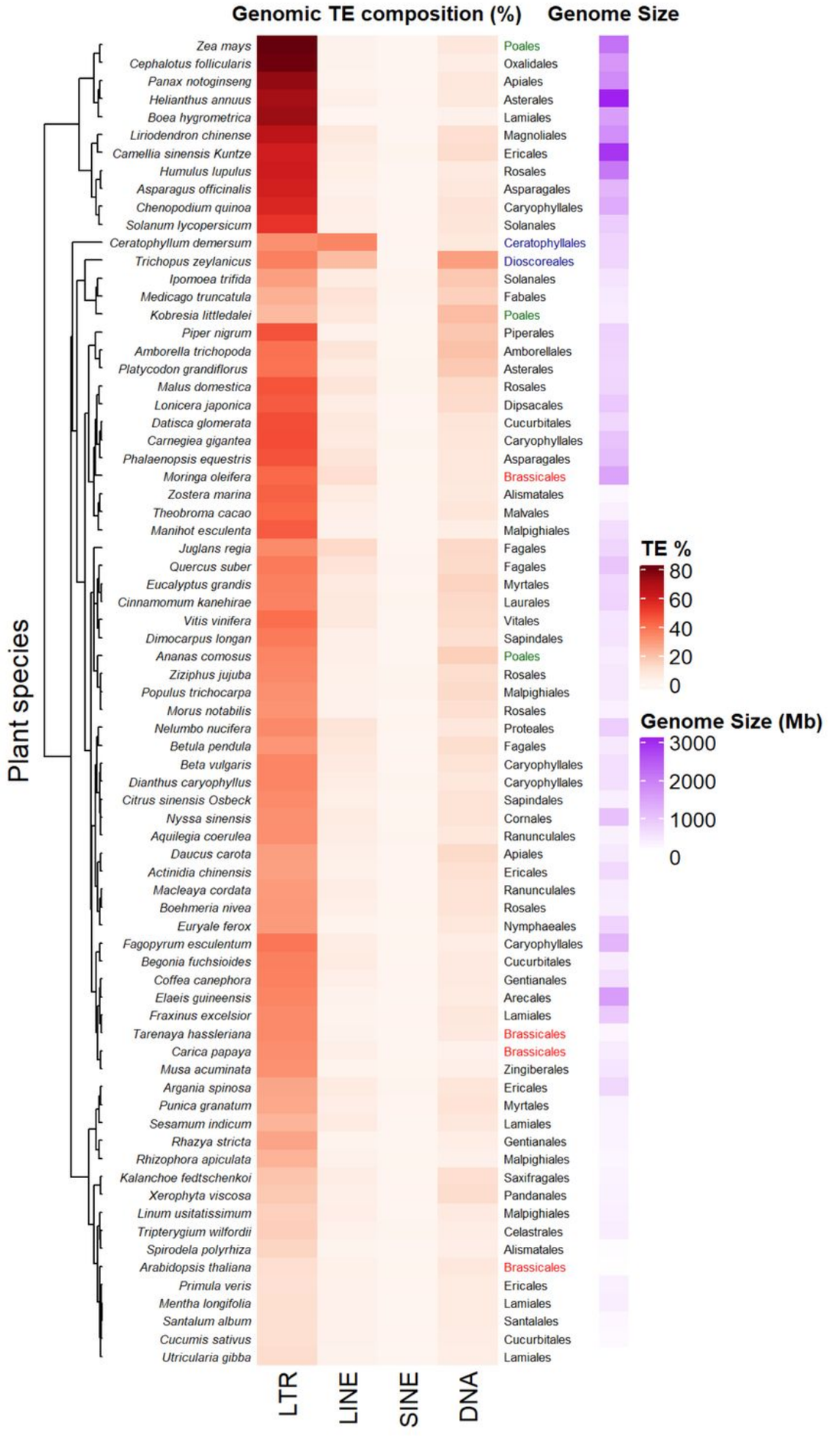
Trace genome evolution

Differentiate TE functions and impacts

Improve genome annotation

Facilitate comparative genomics

Enable biotechnological applications

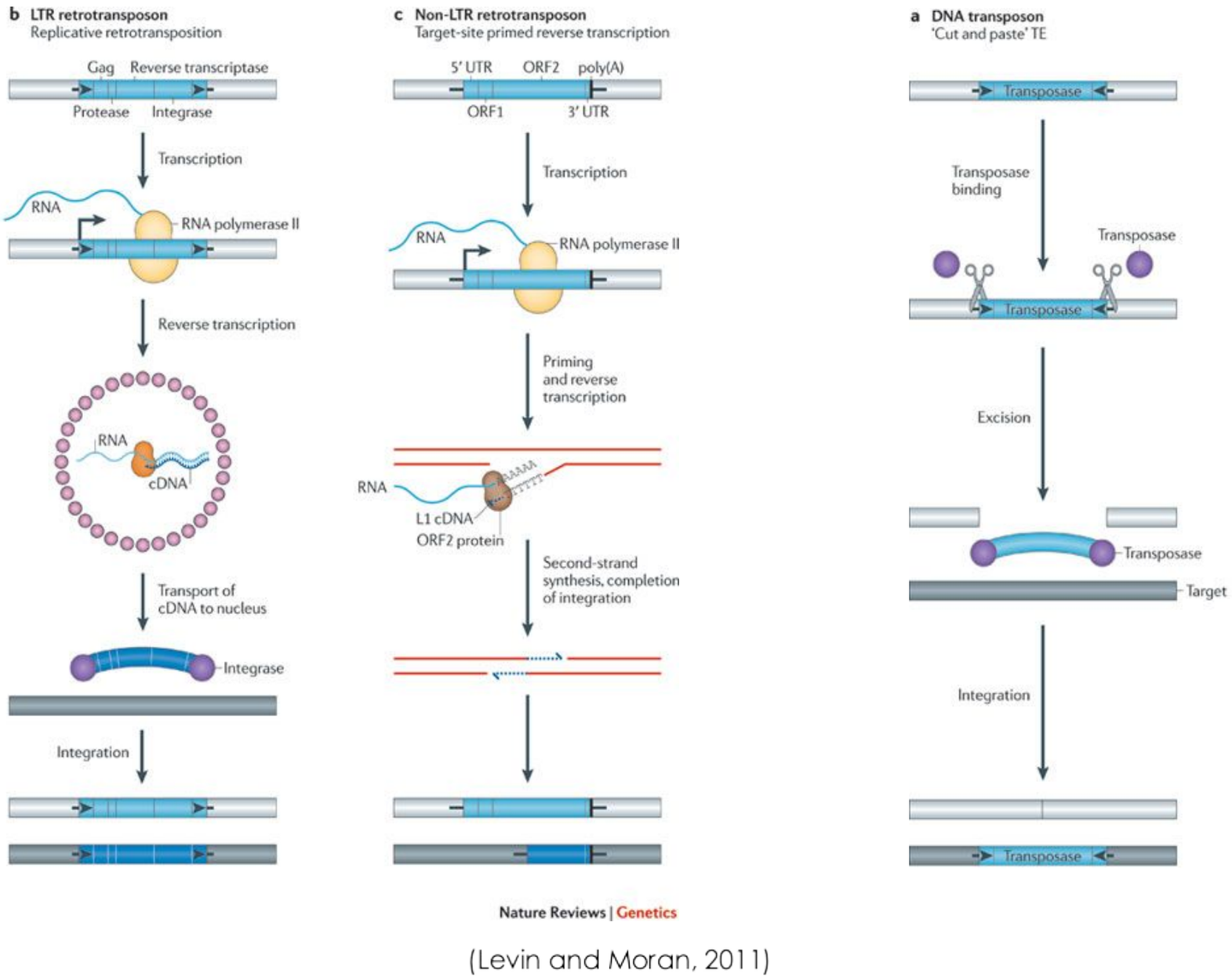


Mhiri et al. 2022

Major classes of Transposable elements

2 major classes of TEs

- Class I: Transposition via an RNA intermediate (“copy and paste”)
  → *retrotransposons*.
- Class II: Direct DNA transposition (“cut and paste”)
  → *DNA transposons*.



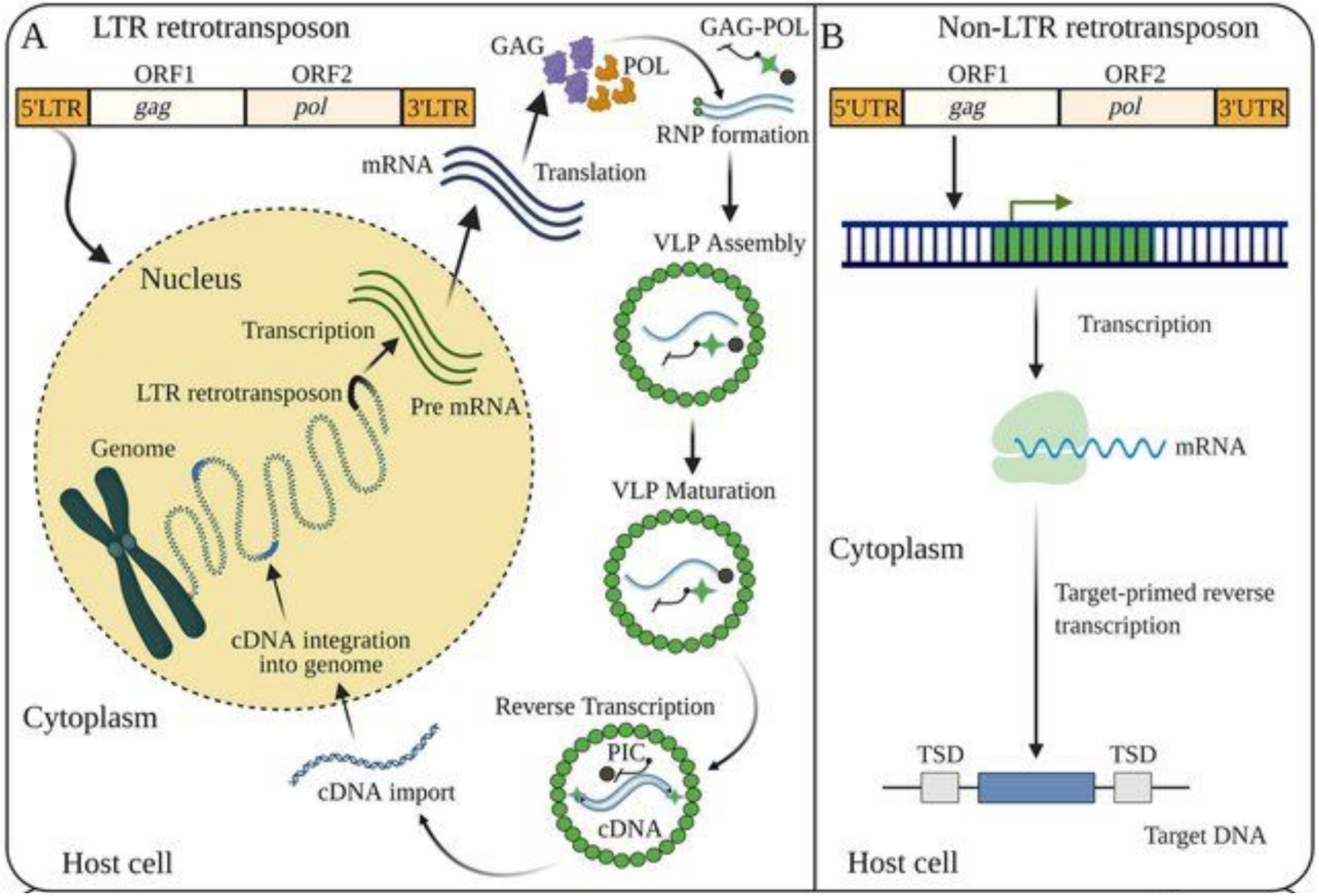
(Levin and Moran, 2011)

Major classes of Transposable elements

2 major classes of TEs

Two main mechanisms:

- Class I: Transposition via an RNA intermediate (“copy and paste”) → *retrotransposons*.



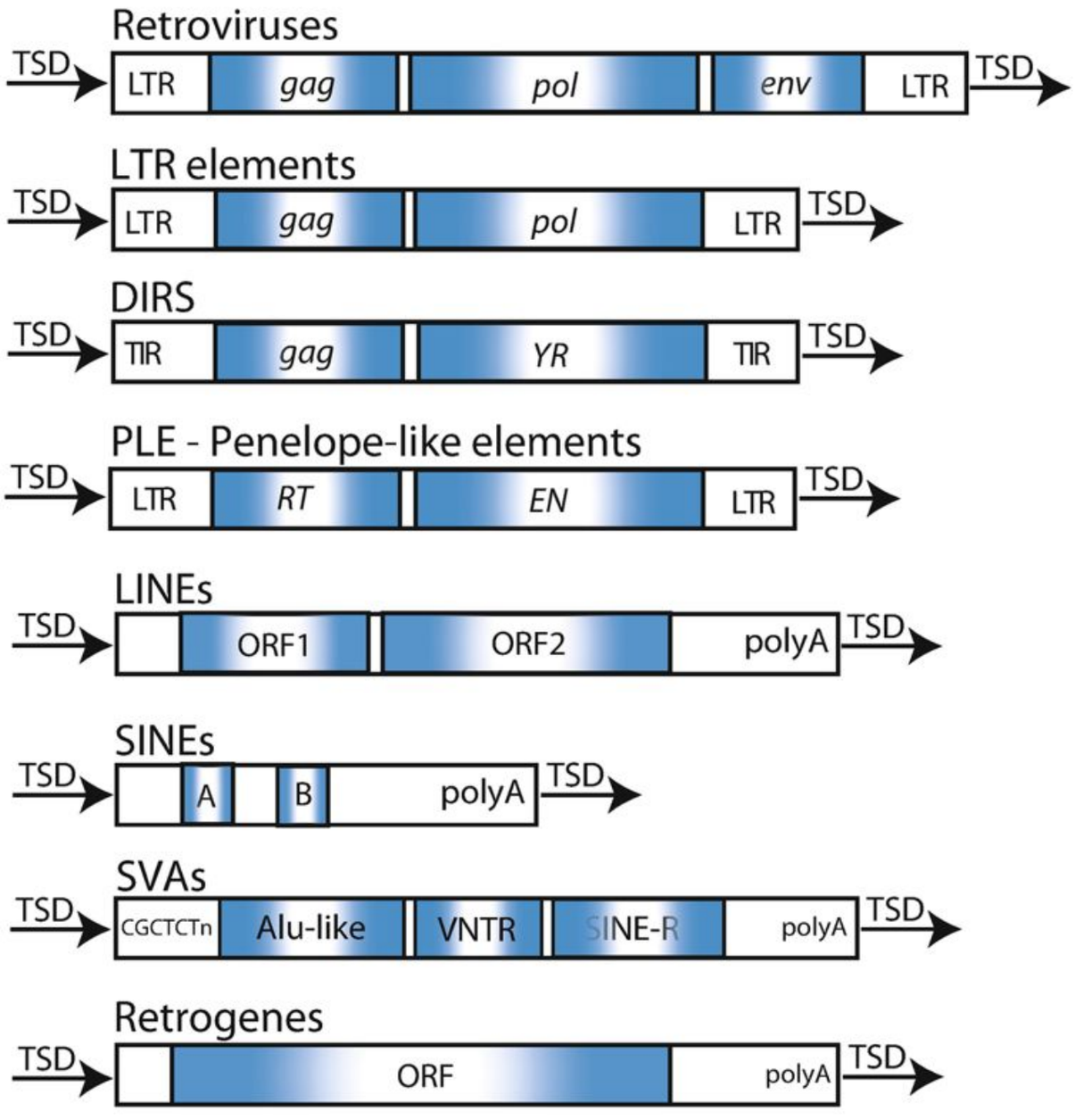
Ramakrisshnan et al. 2022

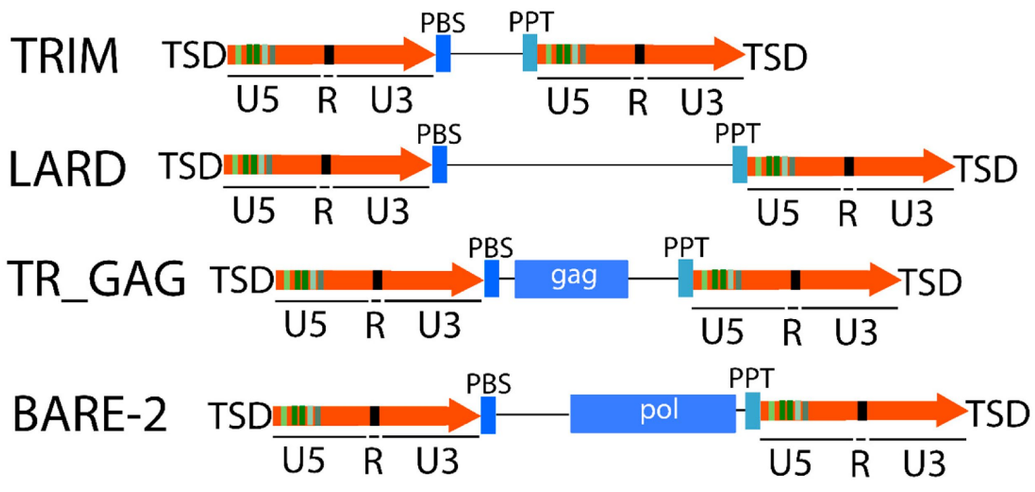
Major classes of Transposable elements

2 major classes of TEs

Two main mechanisms:

- Class I: Transposition via an RNA intermediate ("copy and paste") → *retrotransposons*.





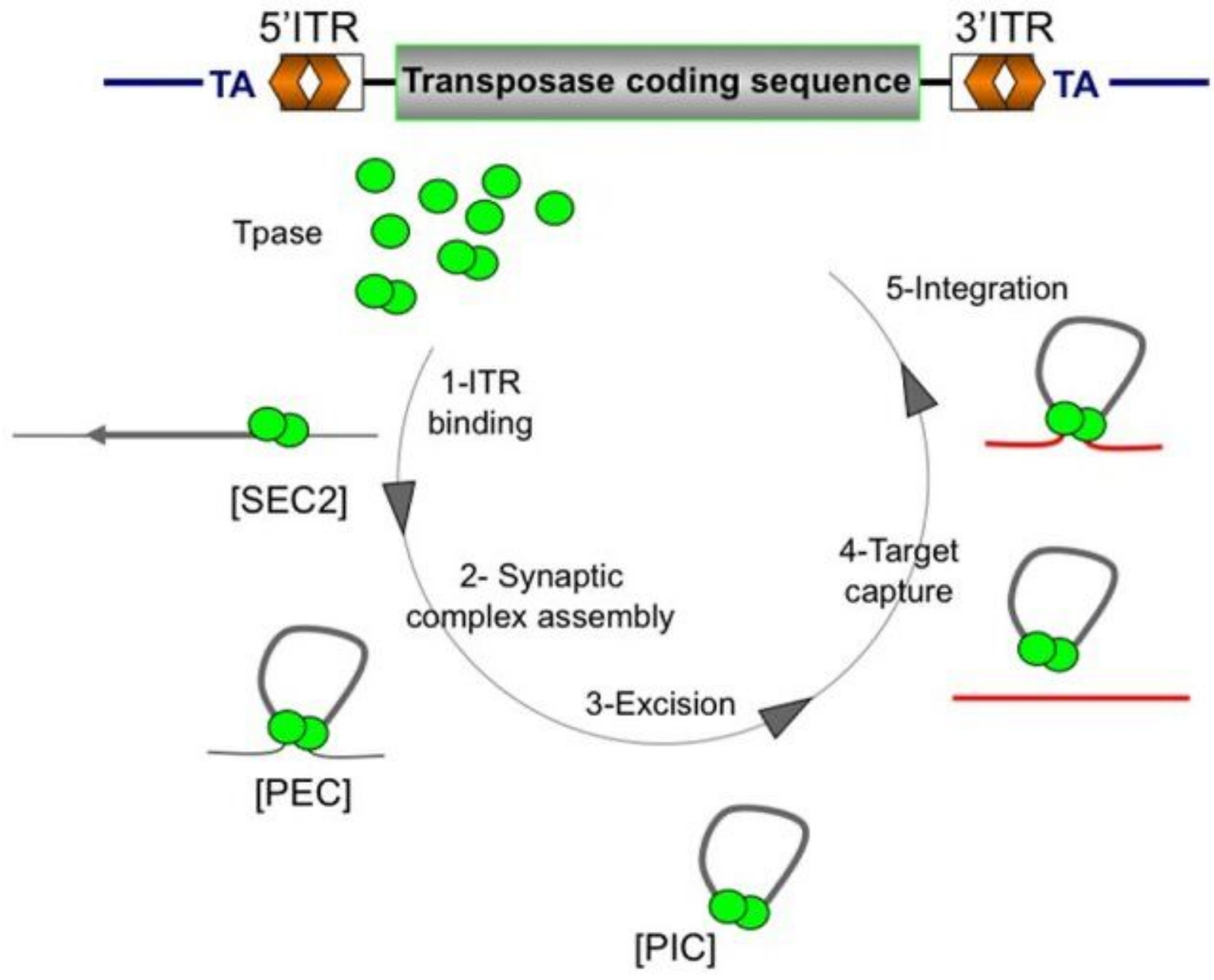
Makalowski el al., 2019

Major classes of Transposable elements

2 major classes of TEs

Two main mechanisms:

- Class II: Direct DNA transposition (“cut and paste”) → *DNA transposons*.



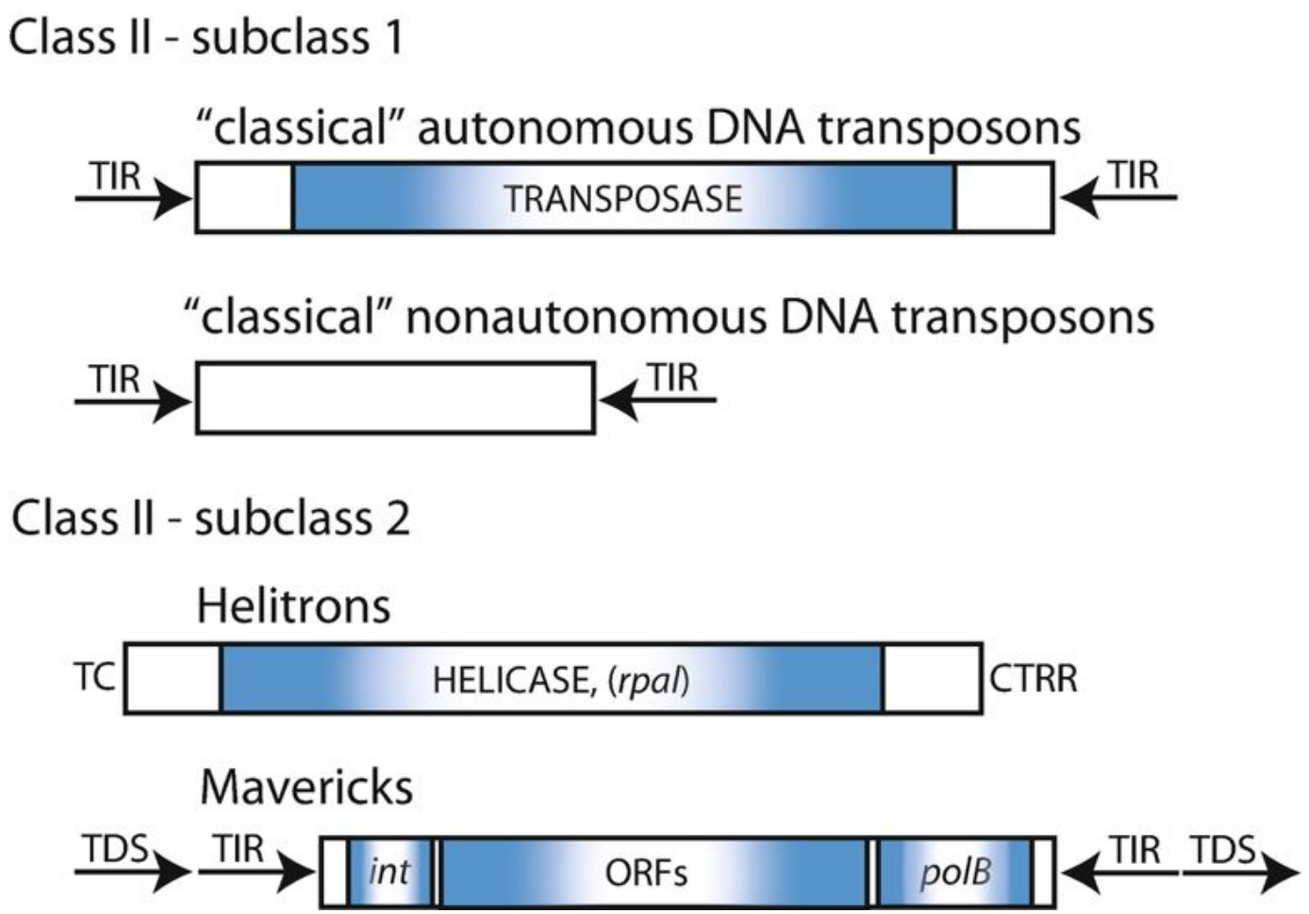
Caillet el al., 2012

Major classes of Transposable elements

2 major classes of TEs

Two main mechanisms:

- Class II: Direct DNA transposition (“cut and paste”) → *DNA transposons*.



Makalowski el al., 2019

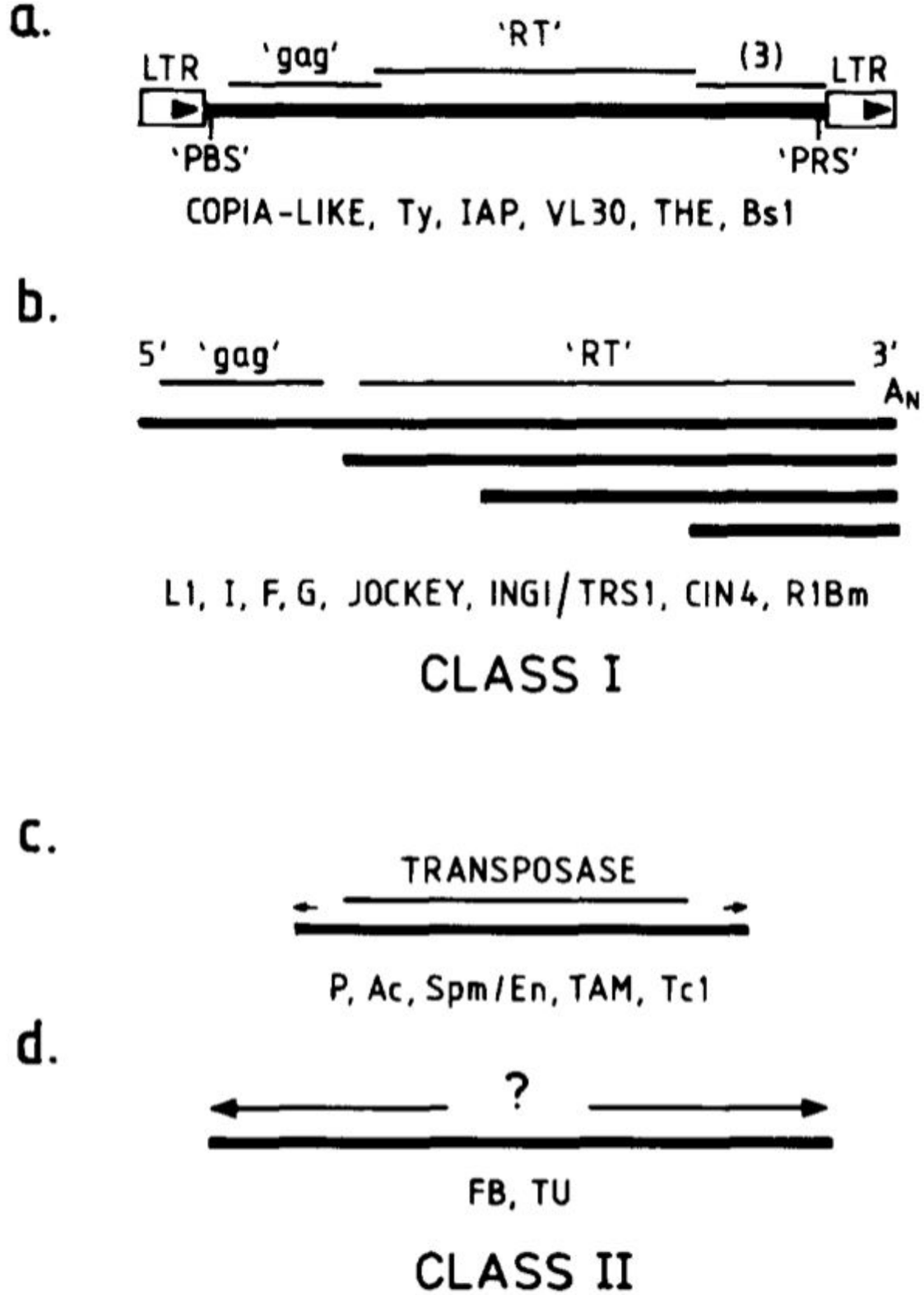
Detailed classification system

Most widely used systems

Most of the systems are based on : Mechanisms, structural features and sequence similarity

1_ First tentative of classification in 1989 by **Finnegan** : First division between Class1/Class2

2_ Classification in 1995 with **RepeatMasker**. Division between Class1/Class2 into 3 levels of sub classifications. Introduction of « Non - autonomous » elements: elements lacking coding properties.



Finnegan, D. 1989

Detailed classification system

Most widely used systems

3_ Tentative of classification in 2007 by **Wicker** and plant genomics Colleagues : A hierarchical system divided into 2 main classes (Class1/Class2) further subdivided by structural properties into Order/Superfamily and possibly into family/subfamily.

Introduction of a 3-letter code to describe the group (ie Class/Order/superfamily; RLC/RLG/ RXX)

Definition of a family with the 80/80/80 rules

Sub classification into family and classification of non autonomous elements using structural features

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF– Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

Protein coding domains

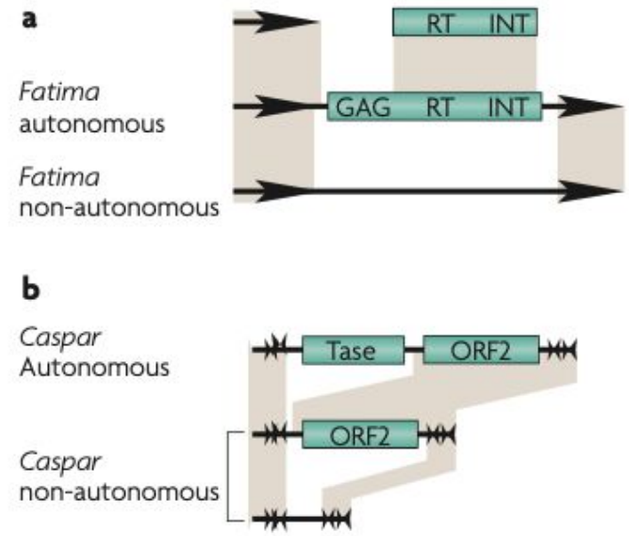
AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

Species groups

P, Plants; M, Metazoans; F, Fungi; O, Others

Wicker et al., 2007

Classification		Structure	TSD	Code	Occurrence
Order	**Superfamily**				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF– Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O



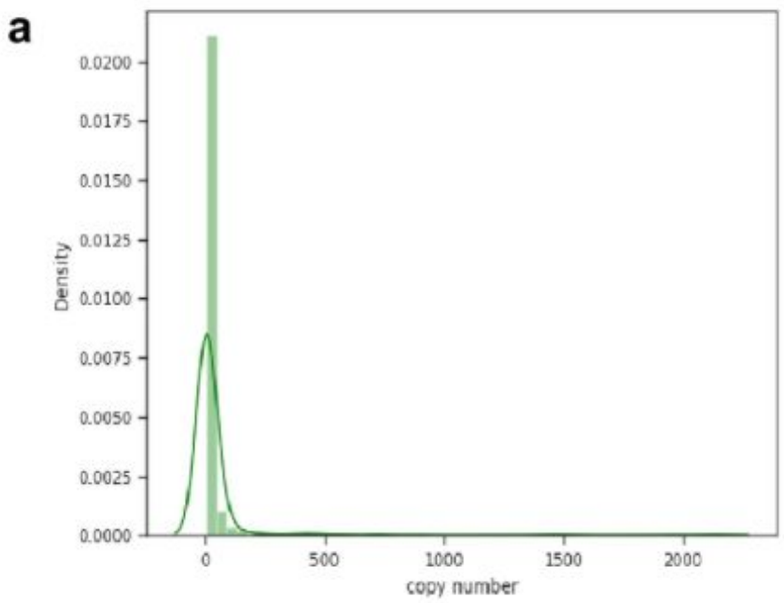
Structural features

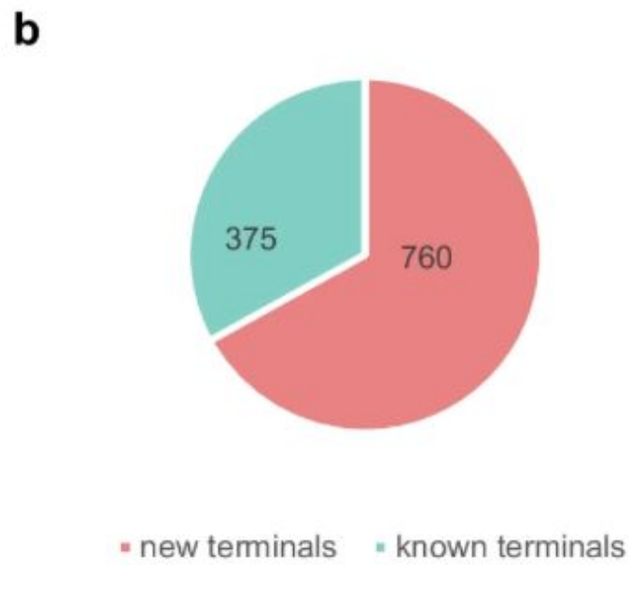
Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

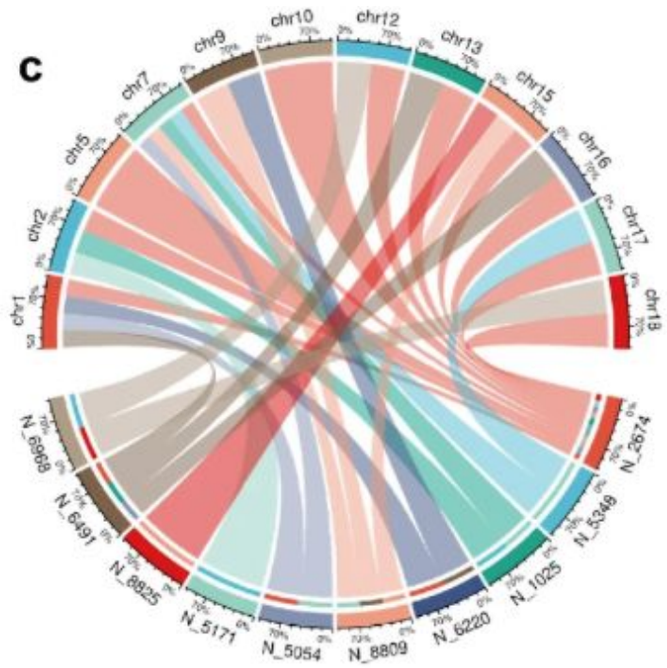
Protein coding domains

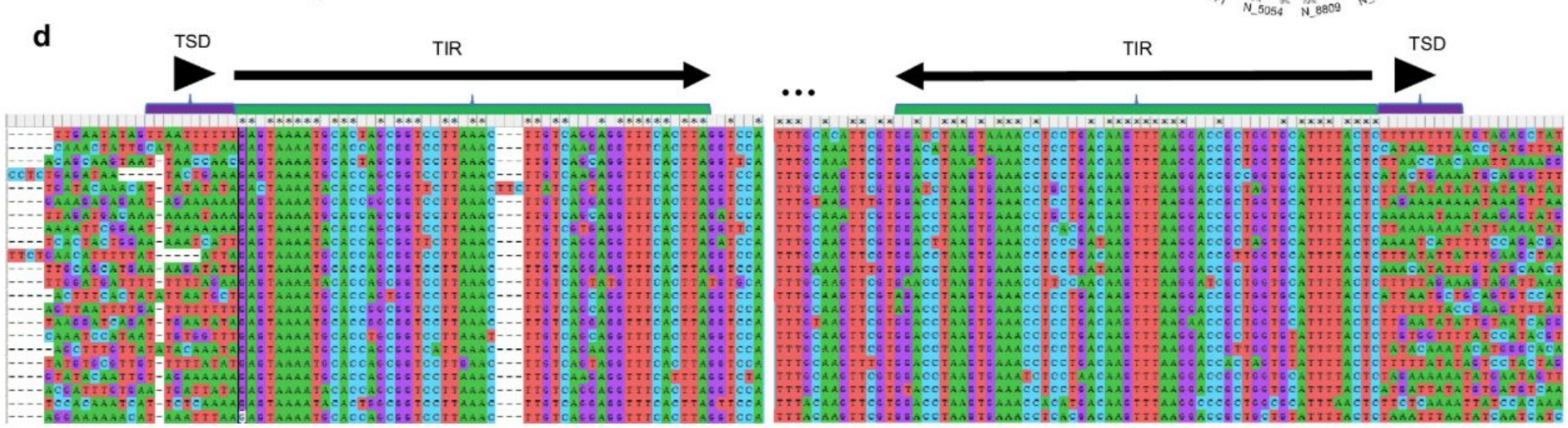
AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function

Detailed classification system





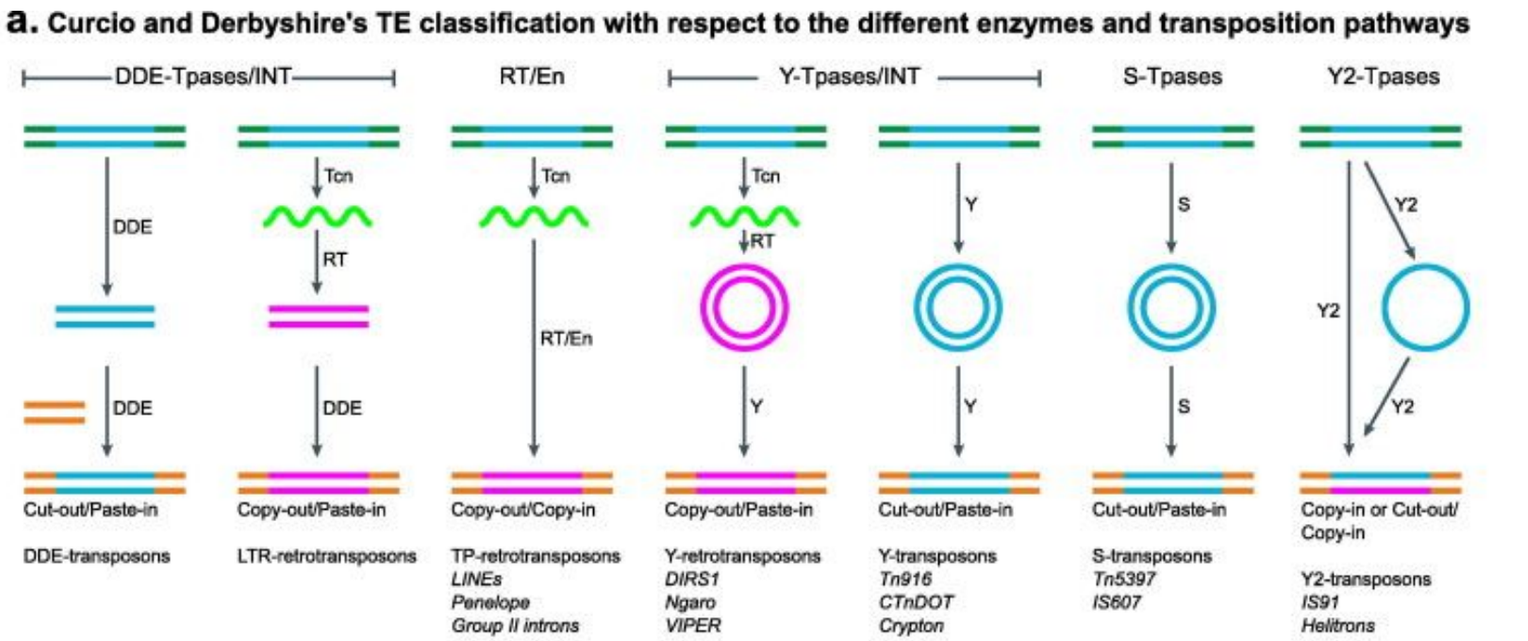




Characteristic of novel TIR elements identified by HiTE-TIR based on *O. sativa.*

Hu et al., 2024

More detailed classification system



(Piégu et al., 2015)

Detailed classification system

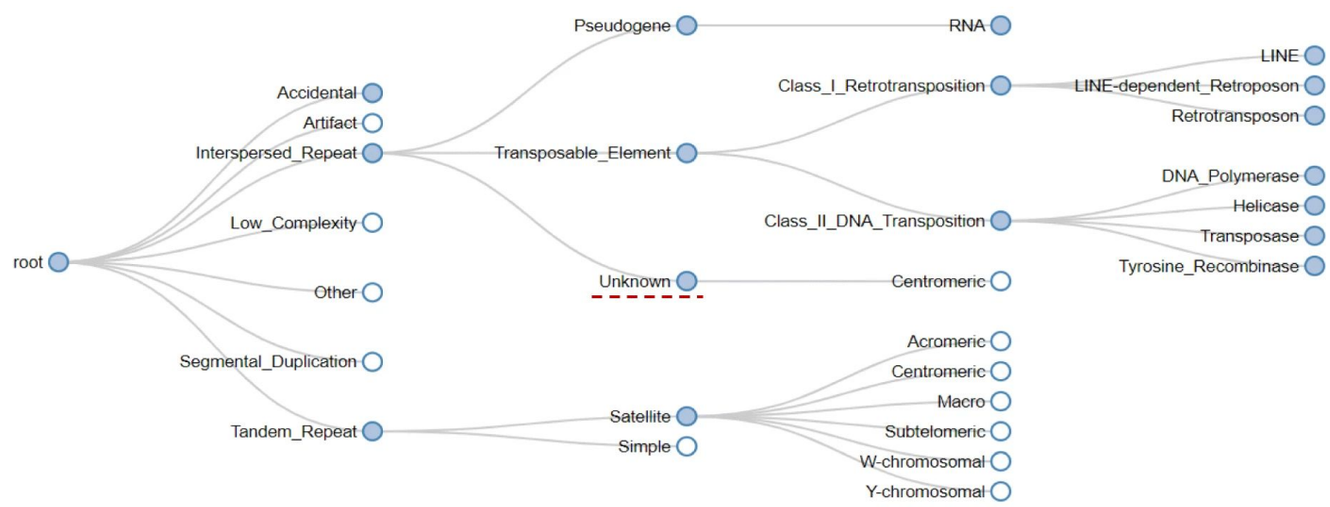
Most widely used systems

4_ Tentative of classification in 2021 by **Storer** into Dfam project (https://dfam.org): A hierarchical system divided into 2 main classes (Class1/Class2) further subdivided by structural, mechanistic and cladistics properties.

Dfam both allows the identification and classification of TE using a collection of HMM and consensus sequences.

Storer et al., 2021

A.



B.

Dfam Classification: Interspersed_Repeat;Transposable_Element;Class_I_Retrotransposition;LINE;Group-II;Group-1;L1-like;L1-group;L1

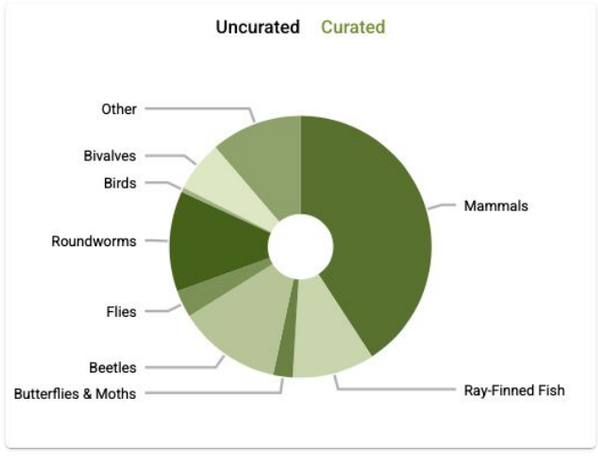
RepeatMasker Type/Subtype: LINE/L1

RepBase: Non-LTR/L1

Wicker: R/I (LINE)/L (L1)

Curcio/Derbyshire: TP-retrotransposons

Dfam release 3.9 (March 2025)



4,121,397	Total Families
26,279	Curated Families
2,784	Species

Detailed classification system

Key classification criteria

Class

Transposition mechanism: RNA vs DNA)

Order to superfamily

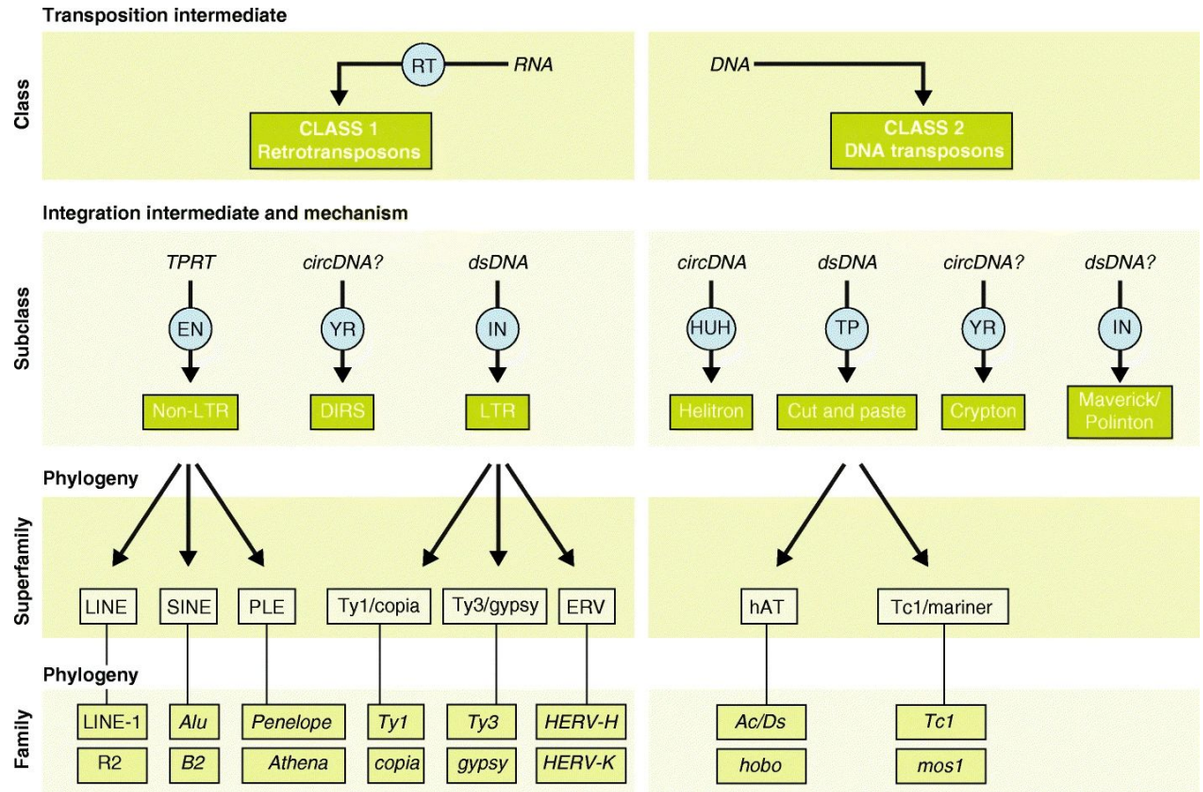
Structural features: presence of LTRs, TIRs, enzymatic domains like reverse transcriptase or transposase

Internal sequence structure: size, terminal repeats, coding capacity (ie. autonomous vs non autonomous)

Family to subfamily

Sequence similarity (80/80/80 rules)

Phylogeny

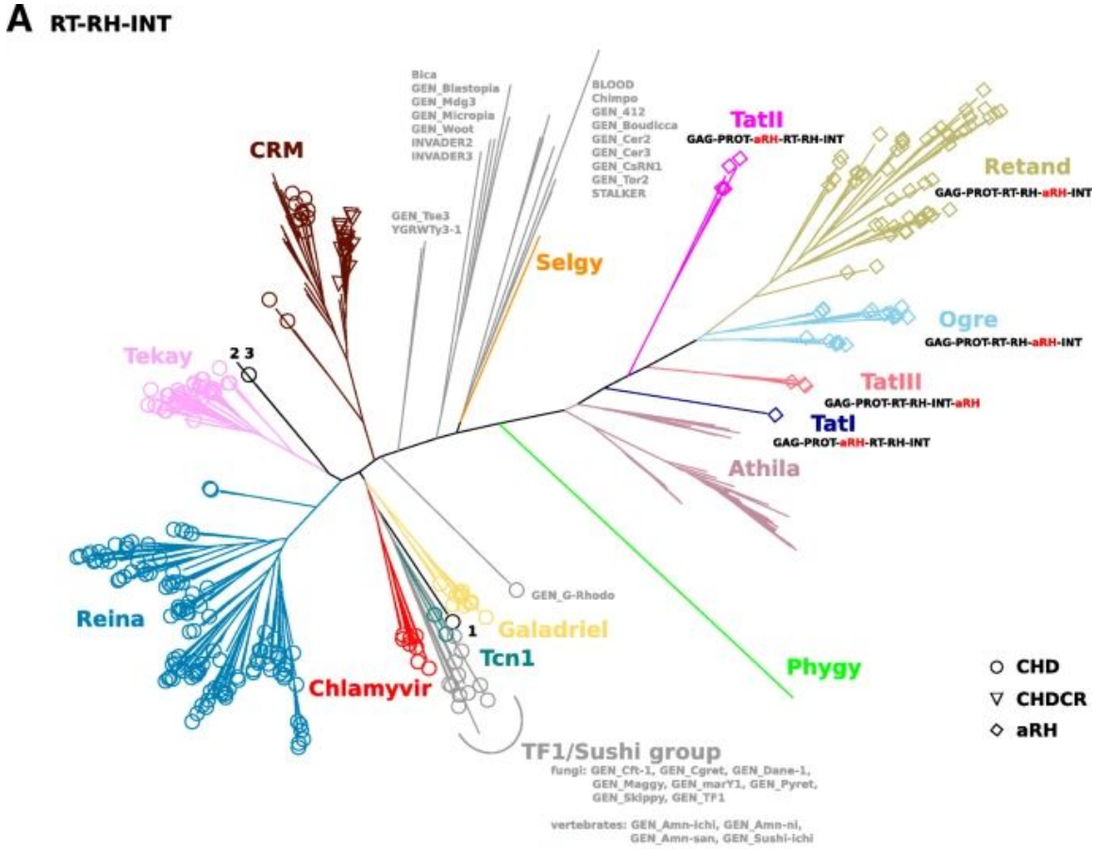


Storer et al., 2021

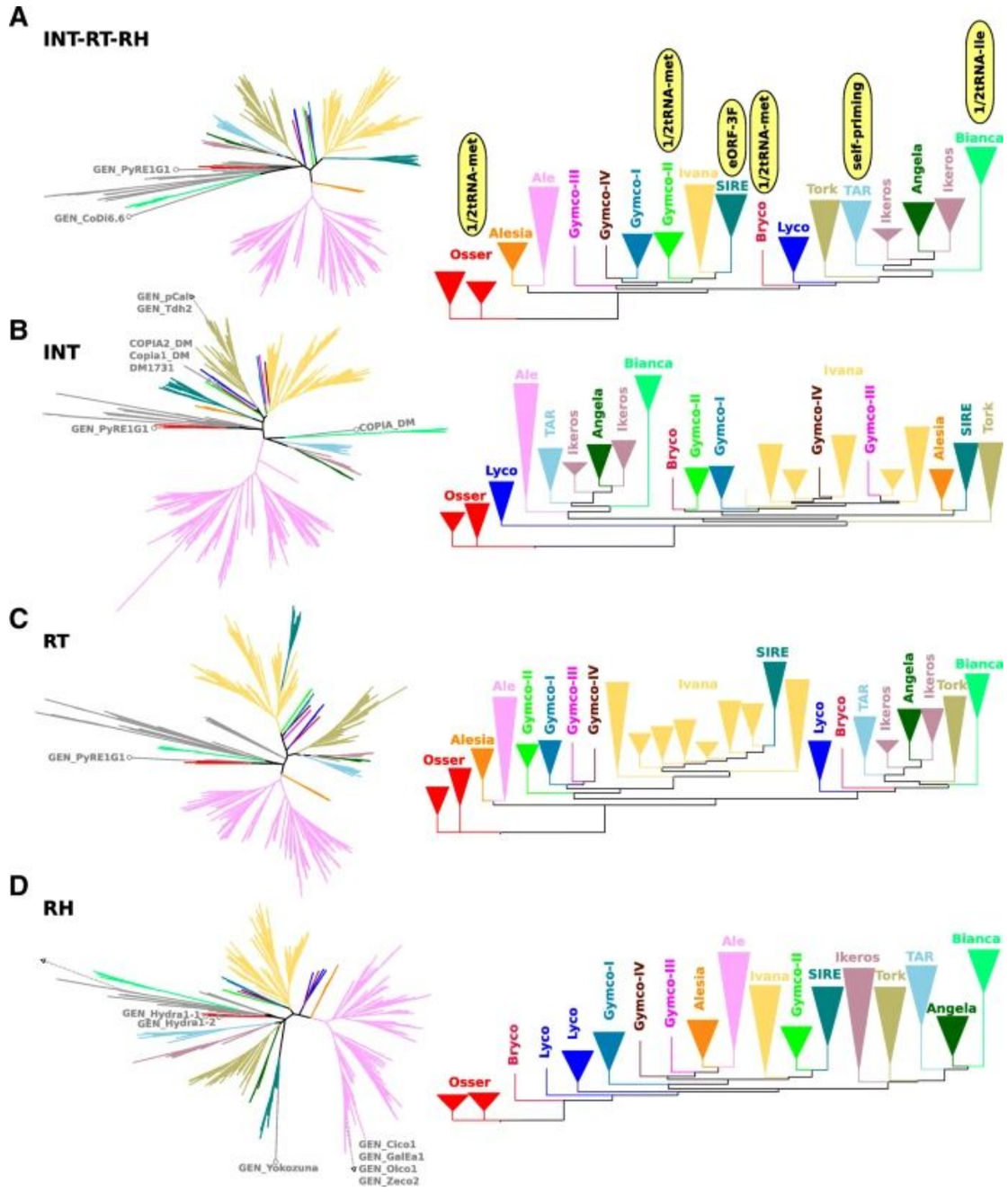
Example of Major Superfamilies

Class I

Class I / LTR retrotransposons / Gypsy



Class I / LTR retrotransposons / Copia

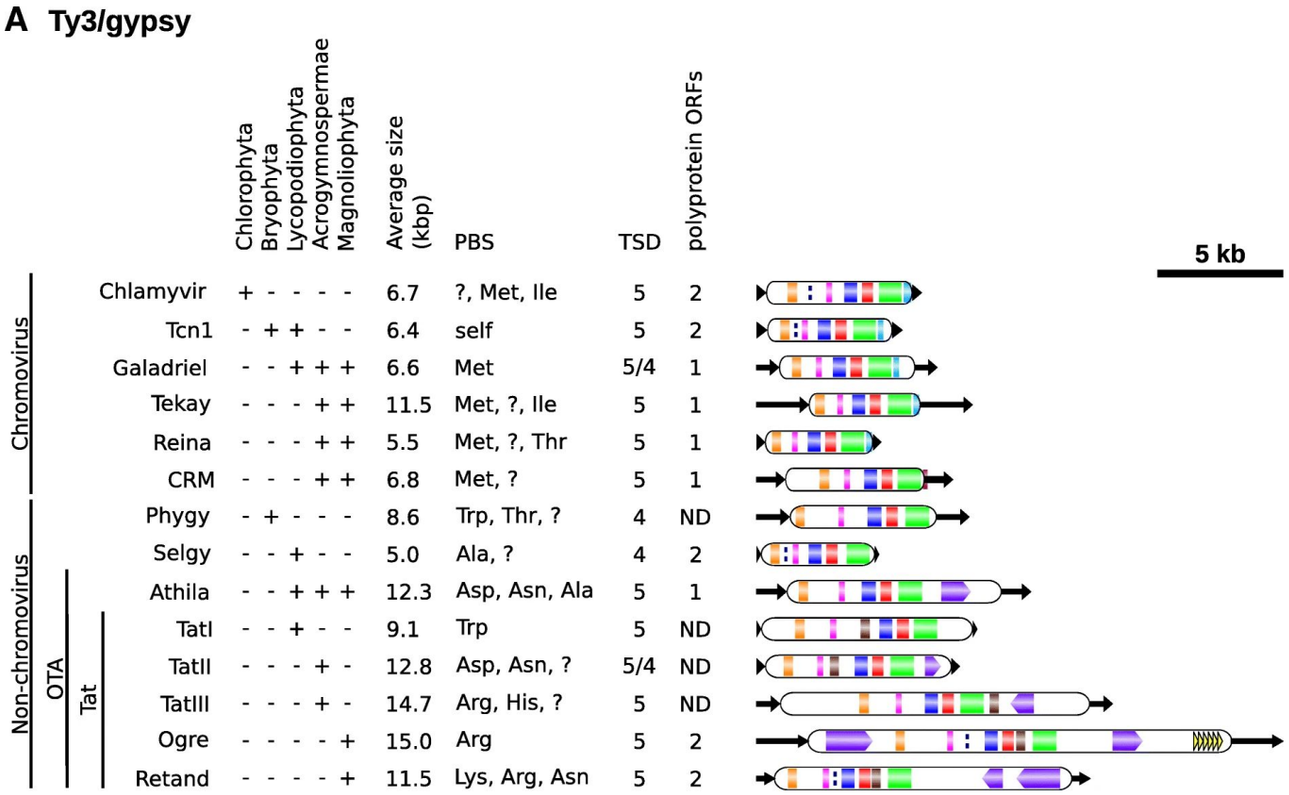


Neumann et al., 2019

Example of Major Superfamilies

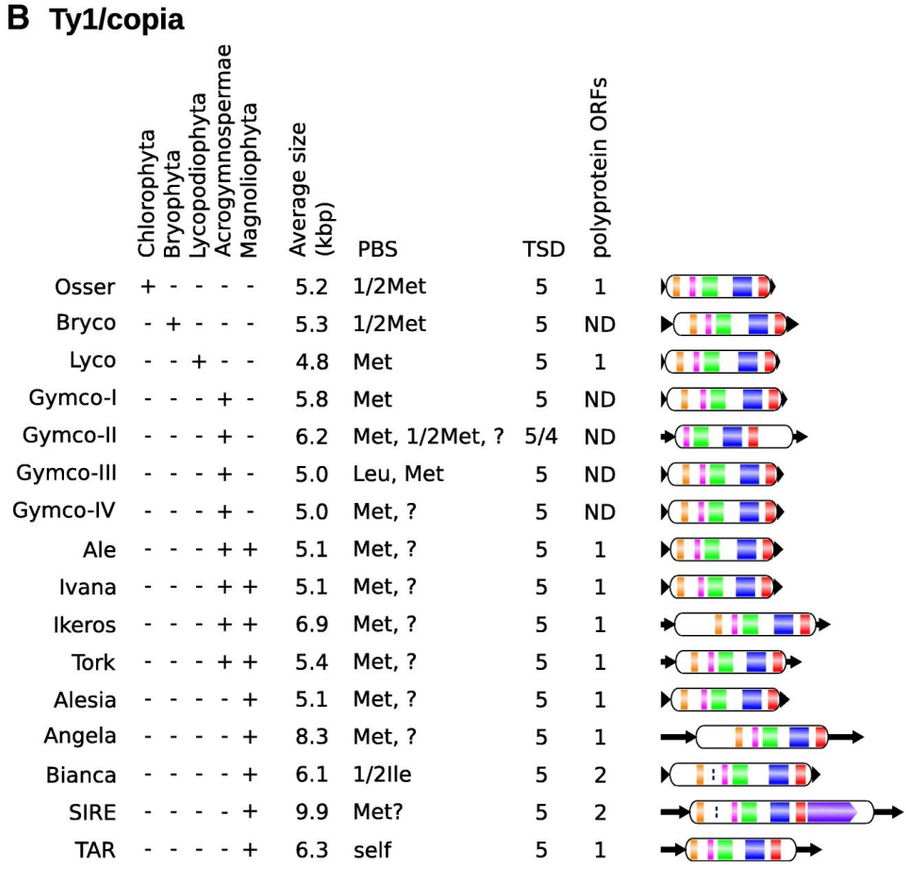
Class I

Class I / LTR retrotransposons / Gypsy



14 families

Class I / LTR retrotransposons / Copia



16 families

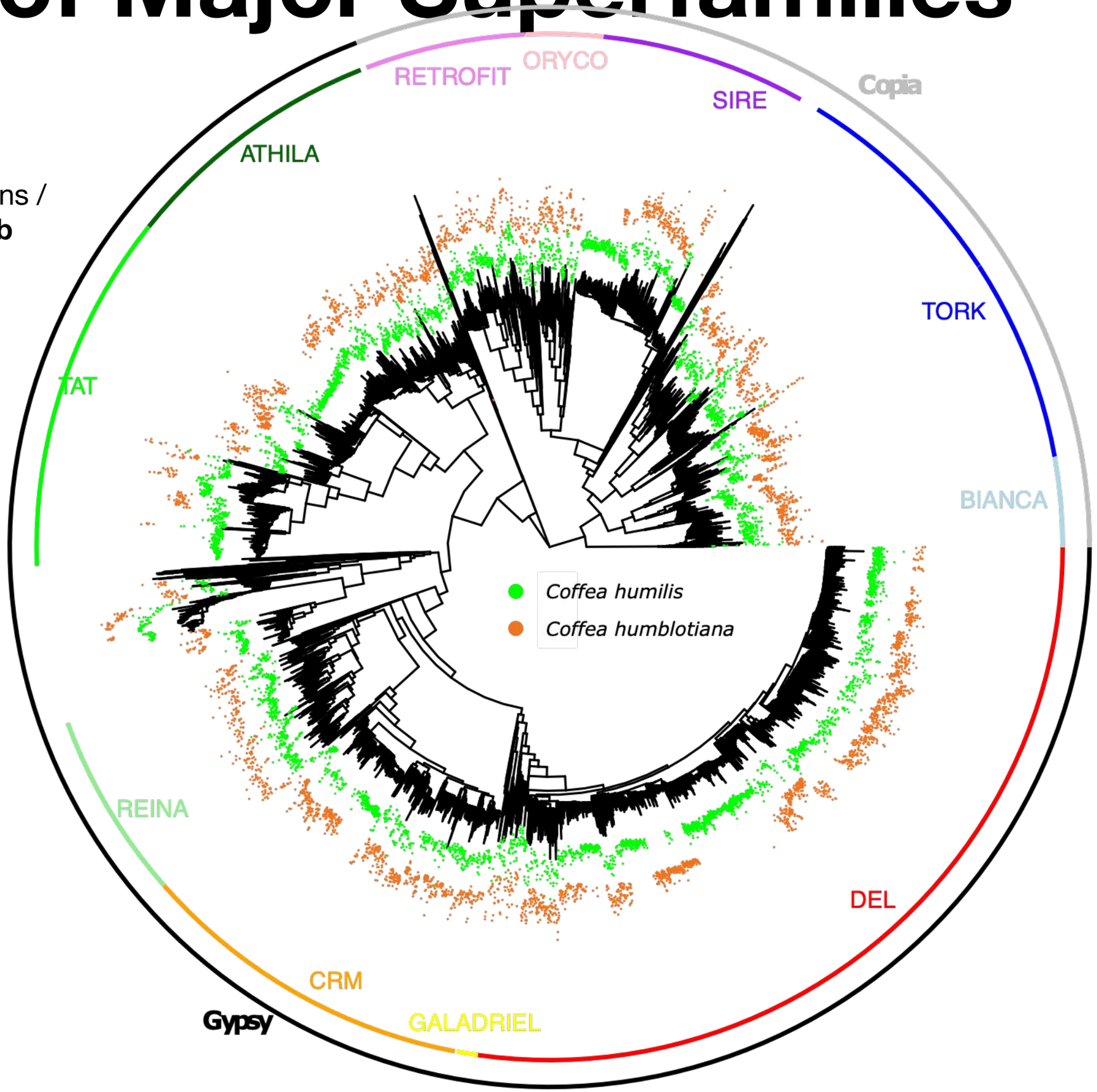
Neumann et al., 2019

Example of Major Superfamilies

Class I

Class I / LTR retrotransposons / Gypsy-Copia/ **Families/ Sub Families**

Reverse Transcriptase based phylogeny of two *Coffea* species



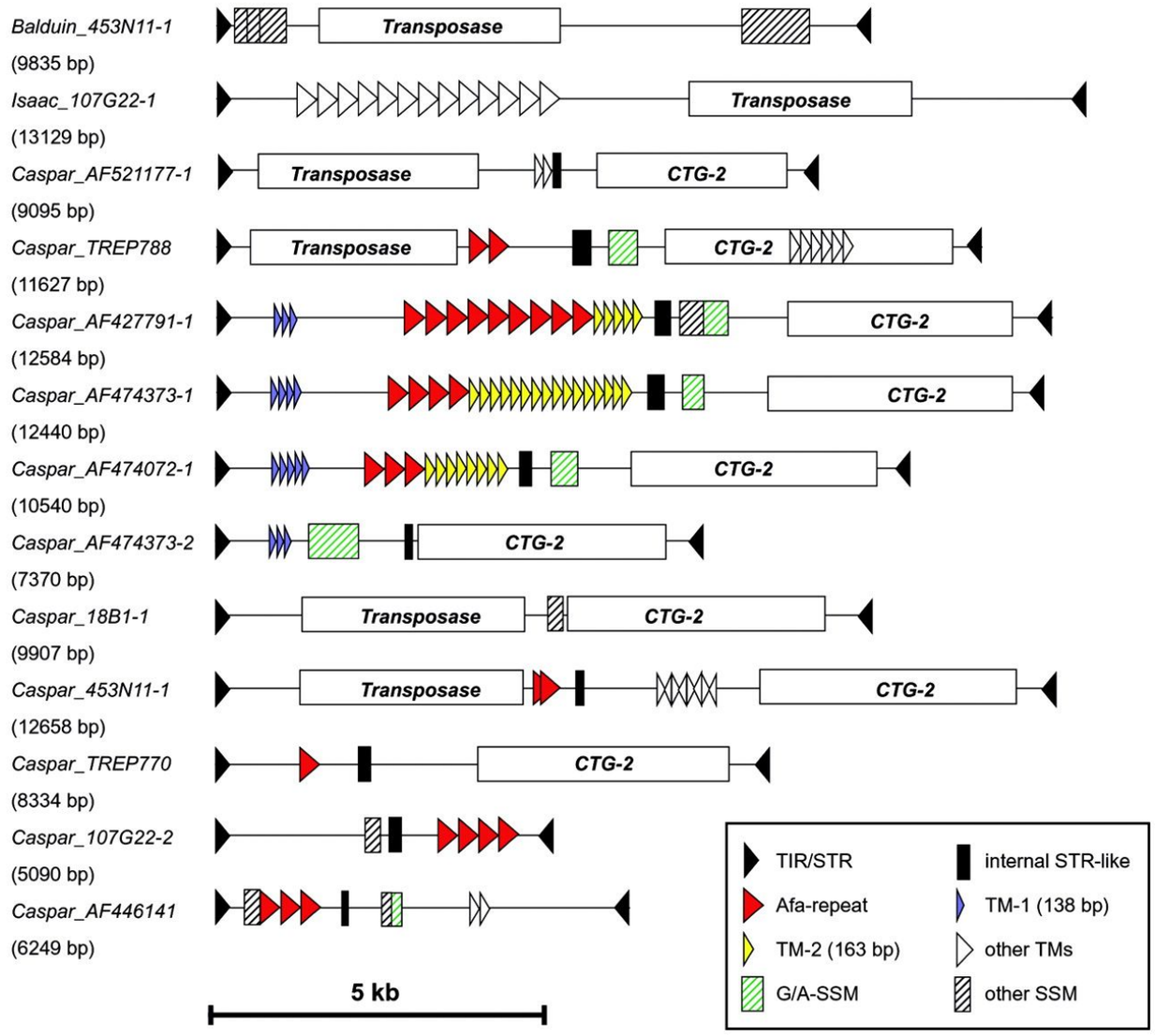
Guyot et al.

Example of Major Superfamilies

Class II

Class II / DNA transposons / En/Spm (CACTA)

Sub classification into family and classification of non autonomous elements using structural features



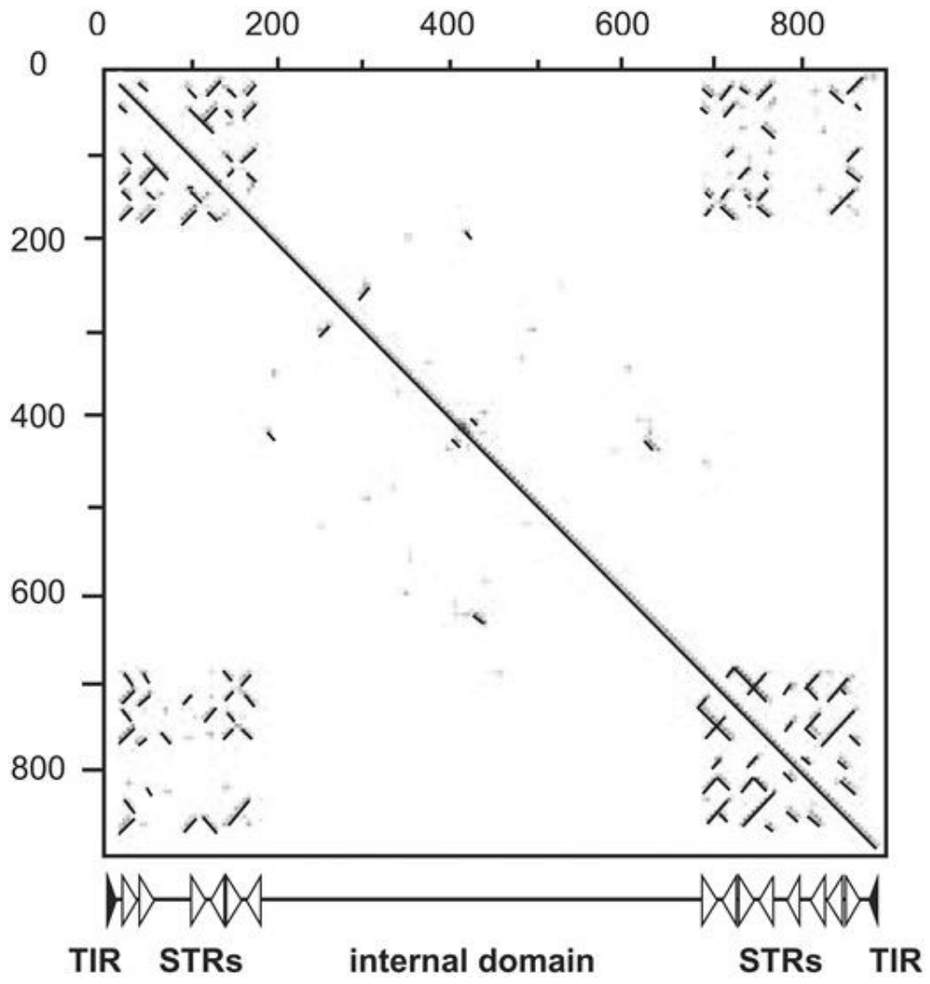
Wicker et al., 2007

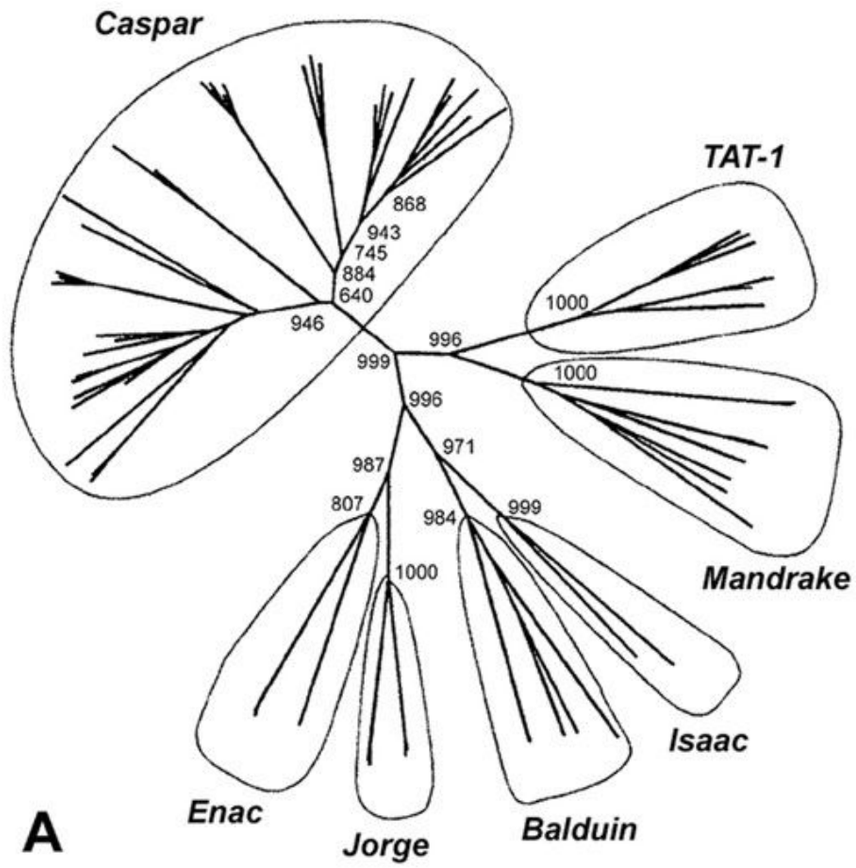
Example of Major Su

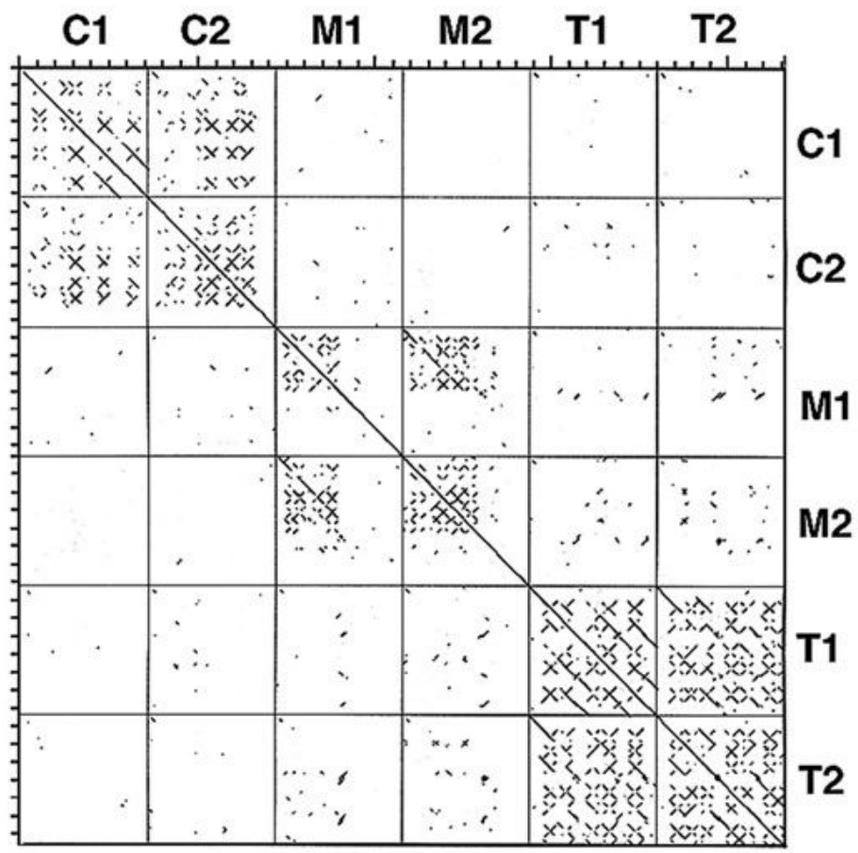
Class II

Class II / DNA transposons / En/Spm (CACTA)

Sub classification into family and classification of non autonomous elements using structural features



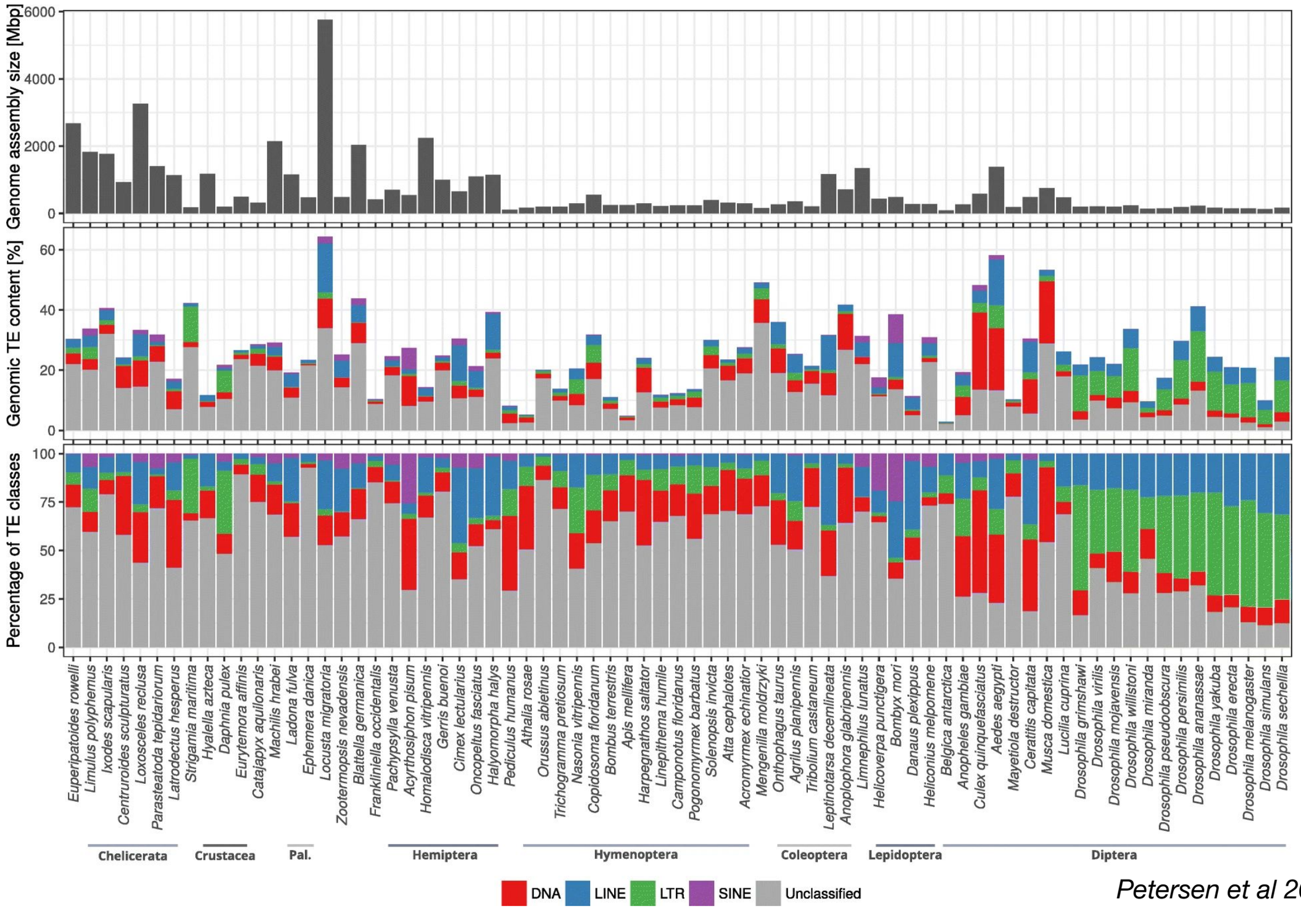




Wicker et al., 2007

TE diversity means different TE dynamics

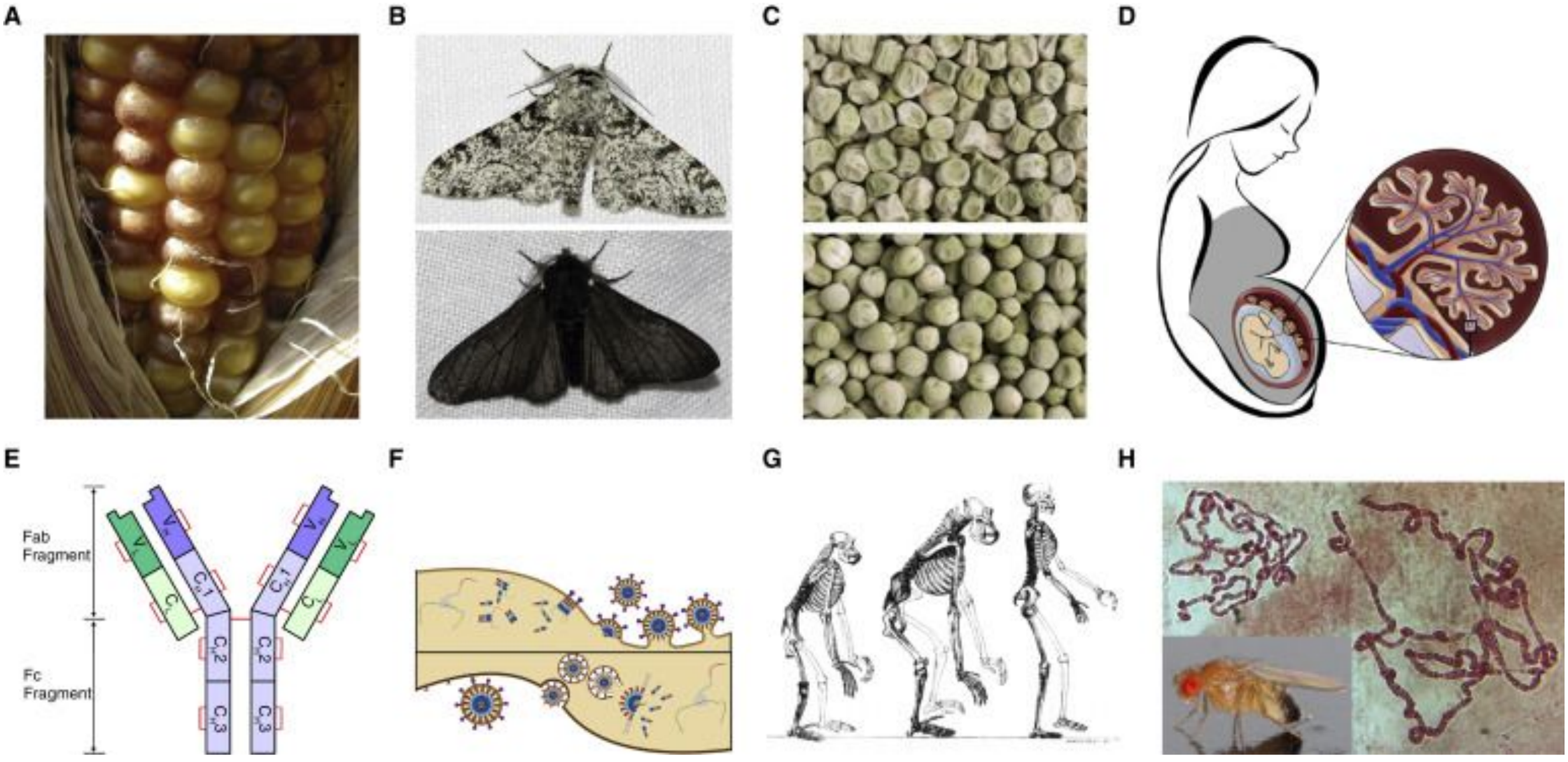
in Arthopoda genomes



Petersen et al 2019

Applications and importance

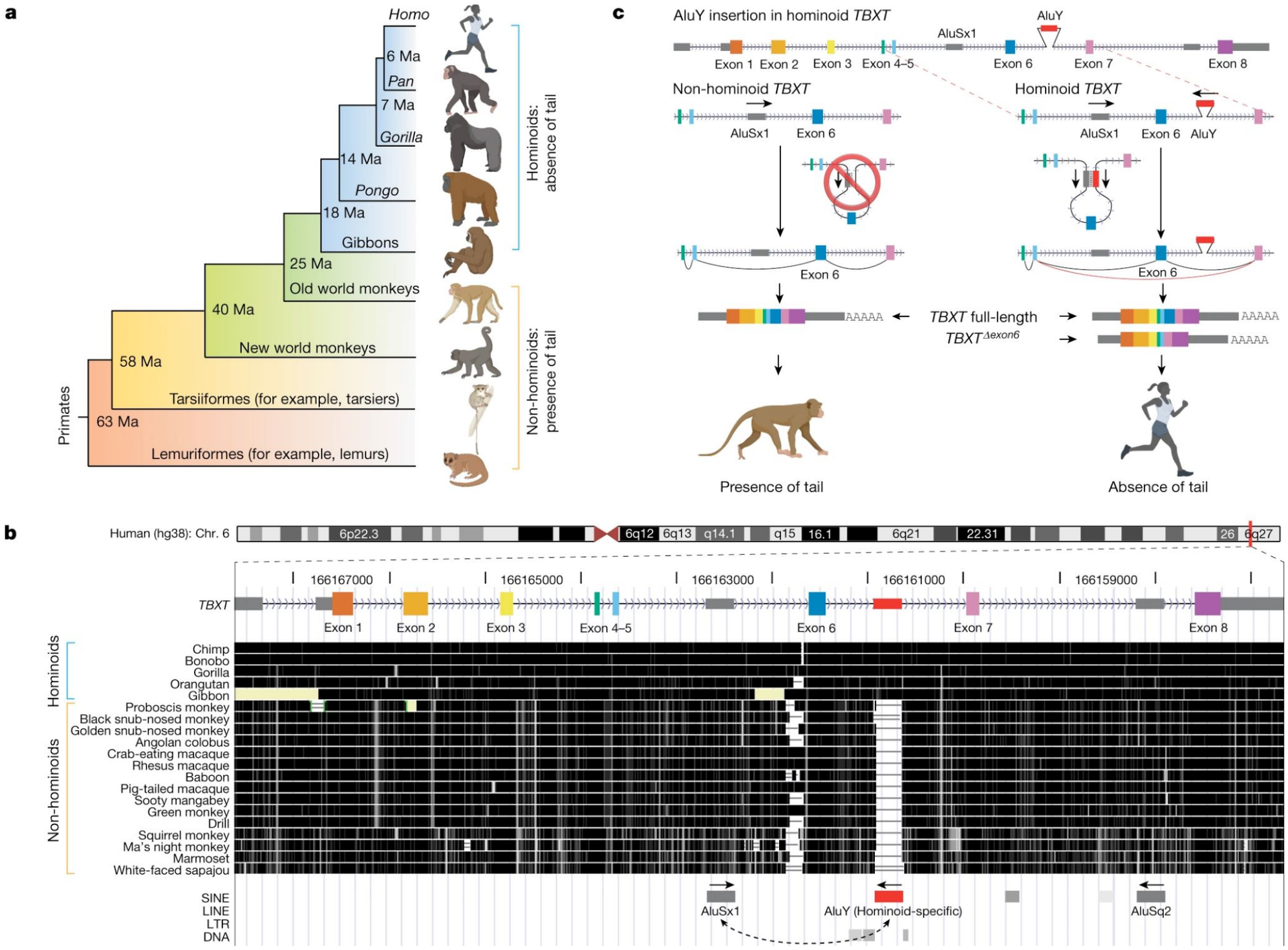
Role in genome evolution and diversity



Hayward and Gilbert 2022

Applications and importance

Involvement in human evolution



Xia et al 2024

Conclusions

Complexity of classification
What level of classification is necessary
Importance of classification to understand the impact of TE

Element	Organism	Disease	Gene	Mechanism
L1	Human	Haemophilia A	Factor VIII	Insertion
L1	Human	DMD	Dystrophin	Insertion
L1	Human	Colon carcinoma	APC	Insertion
Alu	Human	Breast cancer	BRCA2	Insertion
Alu	Human	Haemophilia B	Factor IX	Insertion
Alu	Human	Dent's disease	CLCN5	Insertion
SVA	Human	Various	Various	Insertion
L1	Mouse	Myoclonus	Glrb	Insertion
IAP	Mouse	Various	Various	Insertion
L1	Human	Glycogen storage dis.	PHKB	Ectopic recombination (del)
Alu	Human	Hunter disease	IDS	Ectopic recombination (del)
Alu	Human	Wiskott-Aldrich sdr	WASP	Ectopic recombination (del)
Alu	Human	Lesch-Nyhan sdr	HPRT1	Ectopic recombination (del)
Alu	Human	Lesch-Nyhan sdr	HPRT	Ectopic recombination (dupl)
L1	Mouse	Beige-like coat color	Lystbg	Ectopic recombination (del)
Alu	Human	Ornithine AT def.	OAT	Exonisation
Alu	Human	Sly syndrome	GUSB	Exonisation
Alu	Human	Alport syndrome	COL4A3	Exonisation