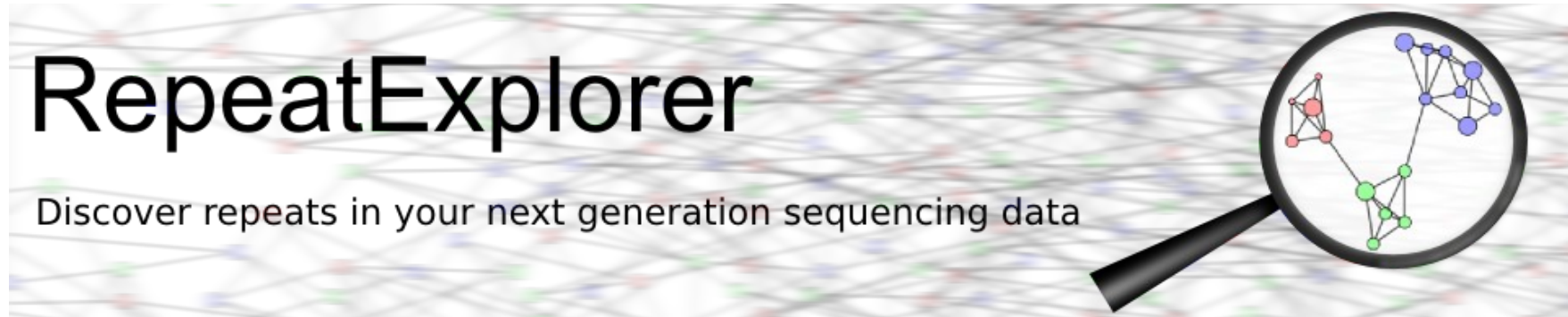


REXdb database and transposon classification using conserved protein domains



Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification

[Pavel Neumann](#) ✉, [Petr Novák](#), [Nina Hošťáková](#) & [Jiří Macas](#)

[*Mobile DNA*](#) **10**, Article number: 1 (2019) | [Cite this article](#)

Classification of transposable elements

A unified classification system for eukaryotic transposable elements

Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel & Alan H. Schulman ✉

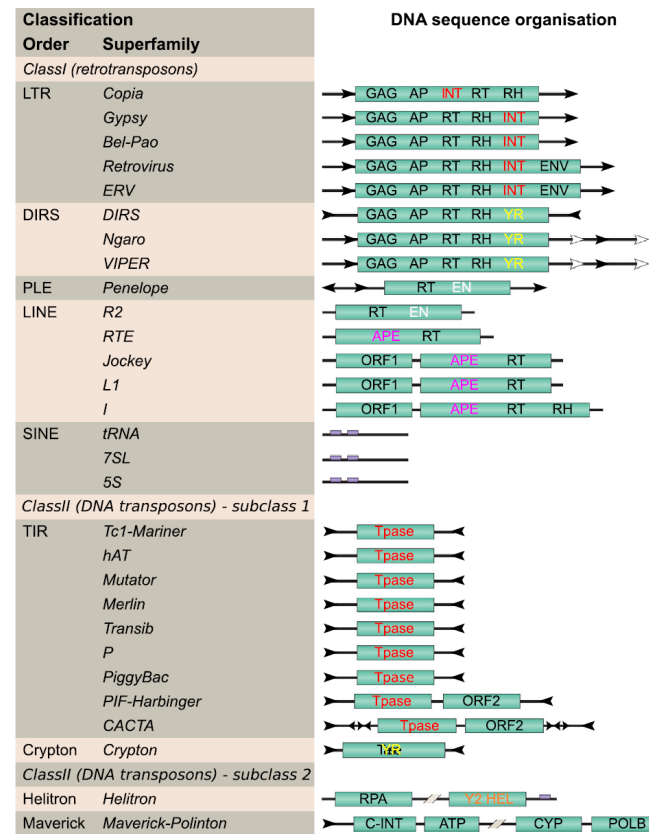
Nature Reviews Genetics **8**, 973–982 (2007) | [Cite this article](#)

Rebase Update, a database of repetitive elements in eukaryotic genomes

Weidong Bao ✉, Kenji K. Kojima ✉ & Oleksiy Kohany

Mobile DNA **6**, Article number: 11 (2015) | [Cite this article](#)

“Unified”



Rebase

Superfamily	Classification Class
<i>Type 2 (retrotransposons)</i>	
Copia	LTR
Gypsy	
BEL	
ERV1, 2 & 3	
DIRS	DIRS
Ngaro	
VIPER	
Penelope	PLE
R2	LINE
RTE	& SINE
Jockey	
L1	
I	
SINE1	
SINE2	
SINE3	
<i>Type 1 (DNA transposons)</i>	
Tc1-Mariner	TIR
hAT	
MuDR	
Merlin	
Transib	(total 15 superfamilies)
P	
PiggyBac	
Harbinger	
En/spm	
Crypton	Crypton
Helitron	Helitron
Maverick-Polinton	Polinton

Piégu et al. Molecular Phylogenetics and Evolution 2015: 90-109

Criteria for classification of transposable elements

Hierarchical levels



class

(e.g. class I)

Type of transposition:

- copy and paste (I)
- cut and paste (II)

order

(e.g. LTR)

Structure:

- type of element termini

Type of replication

Protein domain types

Phylogeny

superfamily

(e.g. Gypsy)

Structure:

- domain order
- type of element termini

family

(e.g. Peabody)

Sequence similarity:

- 80-80-80 rule
- RepeatMasker, CENSOR

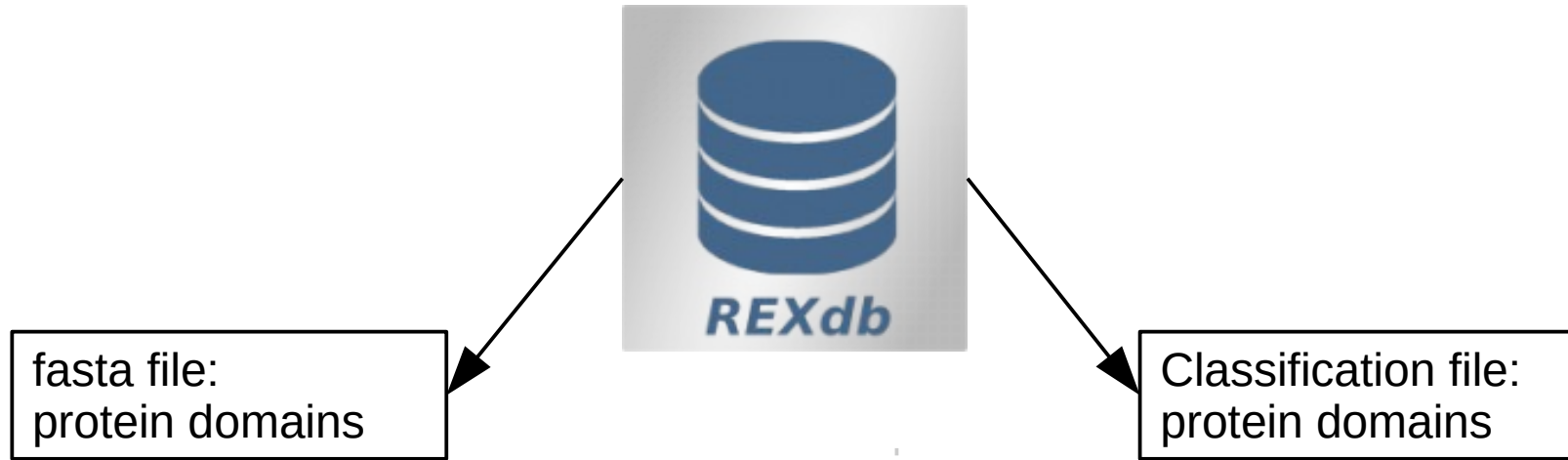
Approaches for identification and classification of transposable elements

- Structural features - LTRs, TIRs
 - various programs (LTR_finder, TIRfinder ...)
- Similarity to sequences in DNA databases
 - Repbase (subscription needed, 1 year - \$1439.00) - Repeatmasker, RepetDB - REPET/PASTEC, DFAM - dfamscan (or nhmmer)
 - TEs are highly divergent in their DNA sequences - DNA databases are of limited use.
- Similarity to sequences in protein databases
 - Pfam (some TE domain types are still missing, the annotations do not follow the established hierarchical classification of TEs)
 - **Protein domains are relatively conserved in sequence, therefore, all types of autonomous TEs can be determined based on the presence and type of protein domains**  **REXdb.**

REXdb database and transposon classification using conserved protein domains

- REXdb covers TEs from a wide range of Viridiplantae species; from Chlorophyta to Spermatophyta).
- A parallel database exists also for Metazoa (not yet released but optional in RepeatExplorer).
- All sequences in REXdb are classified according to the unified classification system (Wicker et al. 2007).
- In RepeatExplorer it serves for automatic classification of TEs. It can also be downloaded from the RepeatExplorer web page (<http://repeatexplorer.org/>).

REXdb database and transposon classification using conserved protein domains



>Ty1-**RT** **REXdb_ID1442**

WRQAMVDEMAALHSNGSWDLVVLPSGKSTVGCRWVYAVKVGPDGQVDRLEKARLVAKGYTQ
VYGSDYGDTFSPVAKIASVRLLLSMAAMCSWPLYQLDIKNAFLHGDLAEEVYMEQPPGFV
AQGESGLVCRLRRSLYGLKQSPRAWFSRFSSVVQEFGLRSTADHSVFYHHNSLGQCIYL
VVYVDDIVITGSDQDGIQKLKQHLFTHFQTKDLGKLKYFLGIEIAQSSSGVVLSQRKYAL
DILEETGMLDCKPVDTP

>Ty1-**INT** **REXdb_ID1442**

HNRLGHPSLSKFQKMVPRFSTLSSLPCECQLGKHTRVSFPKRLNNRAKSPFELVHTDVW
GPCRTASTLGFQYFVTFIDDYSRCTWLFLMKNRAELFSIFQKFYTEIQTQFNISIRVLS
DNAREYFSAQFTSFMSHHGILHQSSCAHTPQQNGVAERKNRHLVETARTLLLHNHVPFRF
WGDAVLTACYLIN

REXdb_ID1442	Class_I	LTR	Ty1/copia	Ale
REXdb_ID1443	Class_I	LTR	Ty1/copia	Ale
REXdb_ID1444	Class_I	LTR	Ty1/copia	Ale
REXdb_ID1445	Class_I	LTR	Ty1/copia	Ale
REXdb_ID1446	Class_I	LTR	Ty1/copia	Ale
REXdb_ID1447	Class_I	LTR	Ty1/copia	Ale

REXdb database and transposon classification using conserved protein domains

- **Viridiplantae v3.0:** 80446 protein domain sequences from a total of 17634 elements from 241 species
 - 13863 LTR retrotransposons (5410 Ty1/copia and 8453 Ty3/gypsy)
 - GAG, PROT, RT, RH, aRH, INT, ChDII, CHDCR domains
 - 852 LINE elements
 - RT, RH, ENDO domains
 - 23 DIRS elements
 - RT, RH, YR (Tyrosine recombinase)
 - 2 Penelope elements
 - RT
 - 65 pararetroviruses
 - PROT, RT, RH domains
 - 2829 Class II transposons
 - TPase or Helicase domain

REXdb database and transposon classification using conserved protein domains

- **Metazoa v3.0:** 11192 protein domain sequences from a total of 5462 elements
 - 2161 LTR retroelements (245 Ty1/copia, 1298 Ty3/gypsy, 564 Bel-Pao, 54 Retroviruses)
 - GAG, PROT, RT, RH, INT domains
 - 1905 LINE elements
 - RT, RH, ENDO domains
 - 209 DIRS elements
 - RT, RH, YR (Tyrosine recombinase)
 - 90 Penelope elements
 - RT, ENDO
 - 1097 Class II transposons
 - TPase or Helicase domain

REXdb database and transposon classification using conserved protein domains

Class_I|LTR|Ty1/copia
Class_I|LTR|Ty3/gypsy
Class_I|DIRS
Class_I|LINE
Class_I|Penelope
Class_I|pararetrovirus
Class_I|LTR|Bel-Pao
Class_I|LTR|Retrovirus

Class_II|Subclass_1|TIR|Academ
Class_II|Subclass_1|TIR|EnSpm/CACTA
Class_II|Subclass_1|TIR|Ginger
Class_II|Subclass_1|TIR|Kolobok
Class_II|Subclass_1|TIR|Merlin
Class_II|Subclass_1|TIR|MuDR/Mutator
Class_II|Subclass_1|TIR|Novosib
Class_II|Subclass_1|TIR|P
Class_II|Subclass_1|TIR|PIF/Harbinger
Class_II|Subclass_1|TIR|PiggyBac
Class_II|Subclass_1|TIR|Sola1
Class_II|Subclass_1|TIR|Sola2
Class_II|Subclass_1|TIR|Sola3
Class_II|Subclass_1|TIR|Transib
Class_II|Subclass_1|TIR|Zator
Class_II|Subclass_1|TIR|Tc1/Mariner
Class_II|Subclass_1|TIR|hAT
Class_II|Subclass_2|Helitron
Class_II|Subclass_2|Maverick

Viridiplantae + Metazoa

Viridiplantae

Metazoa

REXdb: classification of plant LTR retrotransposons into lineages

superfamily
(Ty1/Copia and Ty3/Gypsy)

Domain order

- INT-RT-RH
- RT-RH-INT

Phylogeny

- Bel/Pao

Phylogeny/ENV

- Retrovirus/ERV

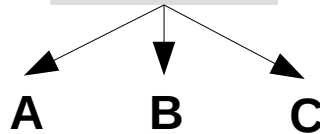


lineage

Phylogeny of polyprotein domains (RT, RH, INT)

Additional support in lineage-specific features

common ancestor



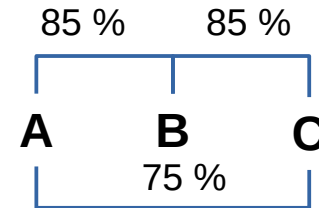
all the same lineage

family
(e.g. Peabody)

DNA sequence similarity

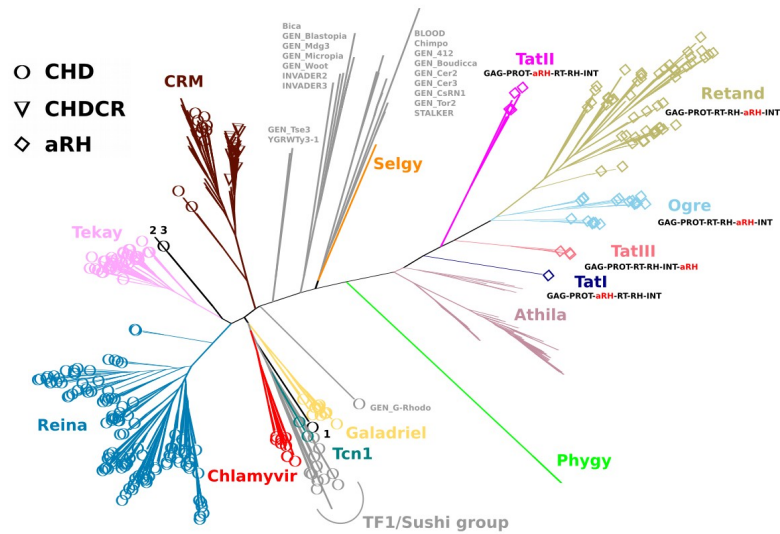
- 80-80-80 rule
- RepeatMasker, CENSOR
- species-specific
- naming problem (synonyms, homonyms)

- Classification problem:



A + B
B + C ?

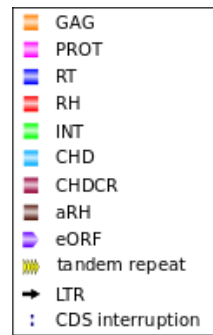
REXdb: classification of plant Ty3/gypsy retrotransposons into lineages



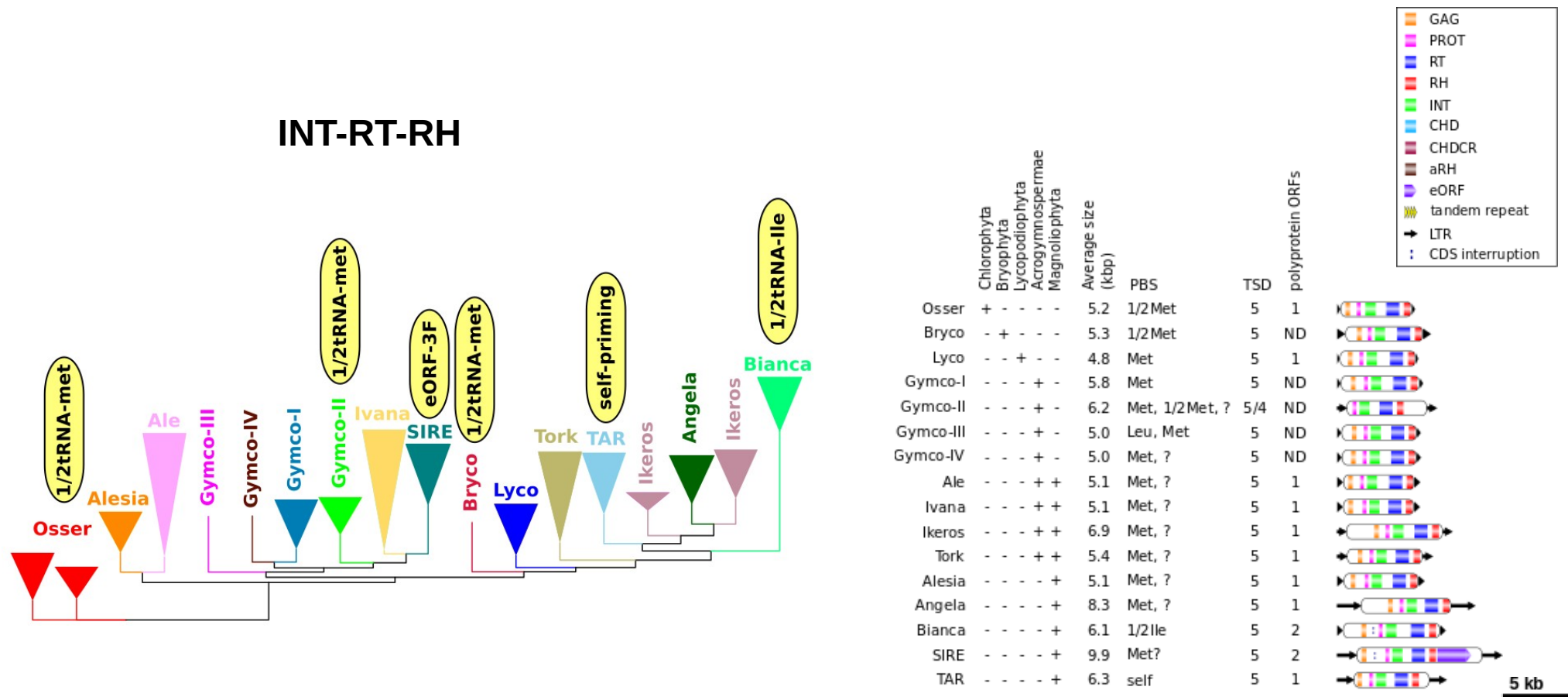
The classification is based on phylogenies inferred from RT, RH, INT, RT-RH-INT sequences

It is strongly supported by structural and sequence features (chromodomain, aRH, eORF, PBS)

	Chlorophyta	Bryophyta	Lycopodiophyta	Acrogymnospermae	Magnoliophyta	Average size (kbp)	PBS	TSD	polyprotein ORFs	
Chlamyvir	+	-	-	-	-	6.7	?, Met, Ile	5	2	
Tcn1	-	+	+	-	-	6.4	self	5	2	
Galadriel	-	-	+	+	+	6.6	Met	5/4	1	
Tekay	-	-	-	+	+	11.5	Met, ?, Ile	5	1	
Reina	-	-	-	+	+	5.5	Met, ?, Thr	5	1	
CRM	-	-	-	+	+	6.8	Met, ?	5	1	
Phygy	-	+	-	-	-	8.6	Trp, Thr, ?	4	ND	
Selgy	-	-	+	-	-	5.0	Ala, ?	4	2	
Athila	-	-	+	+	+	12.3	Asp, Asn, Ala	5	1	
TatI	-	-	+	-	-	9.1	Trp	5	ND	
TatII	-	-	-	+	-	12.8	Asp, Asn, ?	5/4	ND	
TatIII	-	-	-	+	-	14.7	Arg, His, ?	5	ND	
Ogre	-	-	-	-	+	15.0	Arg	5	2	
Retand	-	-	-	-	+	11.5	Lys, Arg, Asn	5	2	



REXdb: classification of plant Ty1/copia retrotransposons into lineages



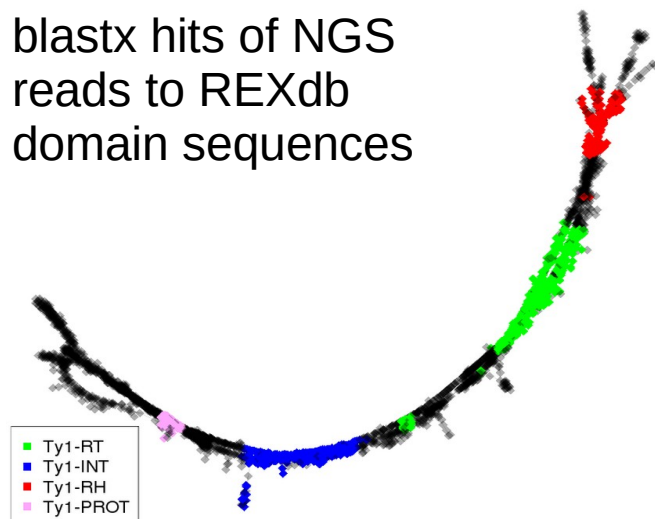
REXdb: classification of transposable elements in RepeatExplorer



Cluster characteristics:

size	3344
size_real	3344
ecount	40179
supercluster	11
annotations_summary	16.54% Class_I/LTR/Ty1_copia/SIRE:Ty1-RT 12.56% Class_I/LTR/Ty1_copia/SIRE:Ty1-INT 5.38% Class_I/LTR/Ty1_copia/SIRE:Ty1-RH 3.08% Class_I/LTR/Ty1_copia/SIRE:Ty1-PROT 0.12% Class_I/LTR/Ty1_copia/Ivana:Ty1-RH
pair_completeness	0.872340425531915
pbs_score	None
TR_score	None
TR_monomer_length	None
loop_index	0.00239234449760766
satellite_probability	6.1246173690846e-24
consensus	None
TAREAN_annotation	Other
orientation_score	1

blastx hits of NGS reads to REXdb domain sequences



supercluster_report.html

SC	size	best_hit	Similarity_based_annotation			Tarean_annotation	clusters
11	5809	SIRE			domains_string	1 (Ty1-GAG), 4 (Ty1-RH), 147 (Ty1-GAG), 420 (Ty1-INT), 103 (Ty1-PROT), 180 (Ty1-RH), 553 (Ty1-RT),	107 , 183 , 372 , 53
				nhits	proportion		
			All	1408	0.24		
			*--repeat	1408	0.24		
			*--mobile element	1408	0.24		
			*--Class_I	1408	0.24		
			*--LTR	1408	0.24		
			*--Ty1 copia	1408	0.24		
			--Ivana	5	0.00086		
			*--SIRE	1403	0.24		

DANTE: a tool for annotation of transposable elements based on the similarity to conserved protein domains database

Galaxy

Tools

search tools

Get Data

RepeatExplorer2

TAREAN

ChIP-Seq Mapper

RepeatExplorer Utilities

DANTE

Protein Domains Filter Tool for filtering of gff3 output from DANTE. Filtering can be performed based on domain type and alignment quality.

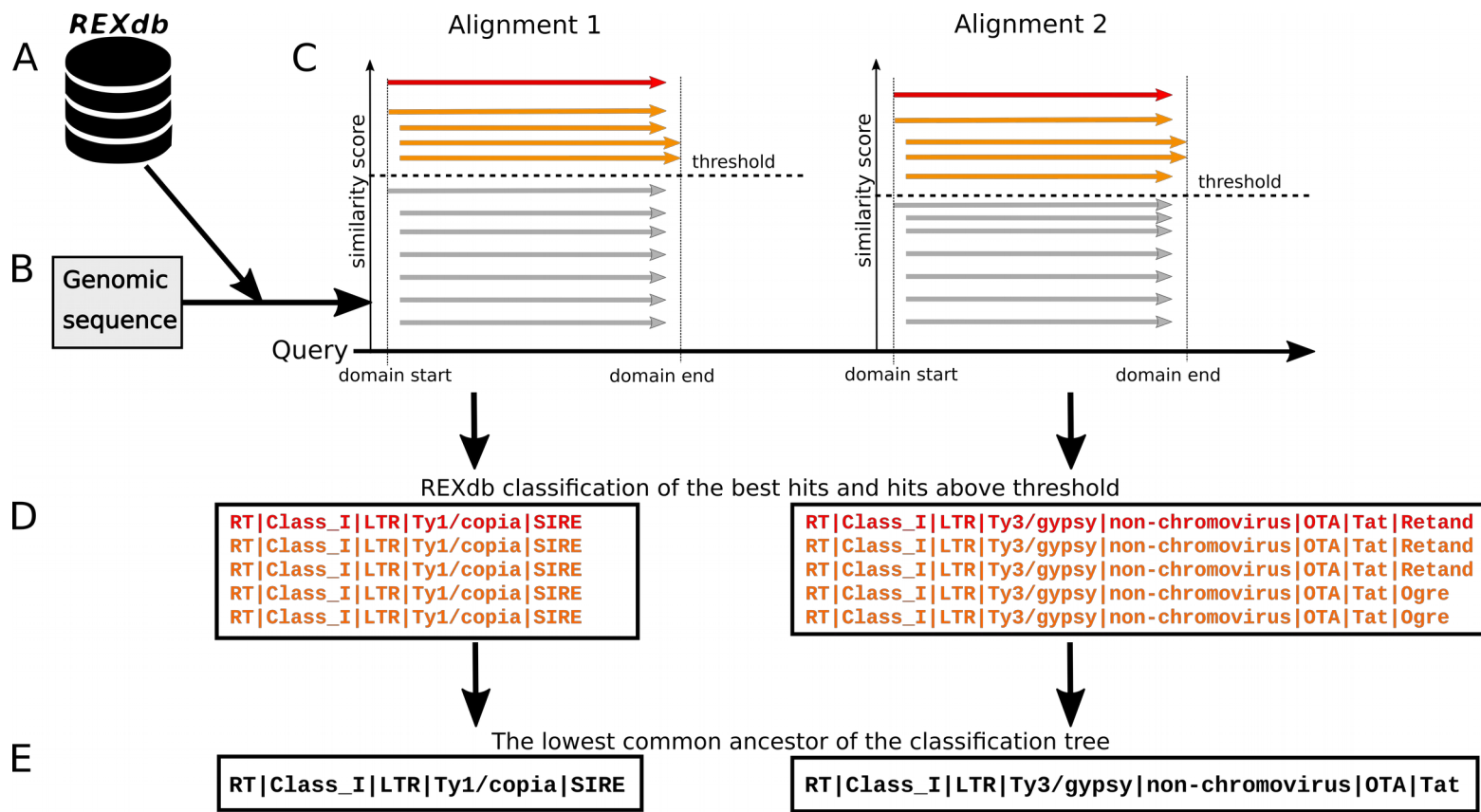
Domain based ANnotation of Transposable Elements - DANTE

Tool for annotation of transposable elements based on the similarity to conserved protein domains database.

Extract Domains Nucleotide Sequences Tool to extract nucleotide sequences of protein domains found by DANTE

Summarize gff3 output from DANTE

Convert dante gff3 to tab delimited file



Final remarks

- Always select the appropriate database of protein domains (either for Viridiplantae or Metazoa spp.).
- Seed-free vascular plants (lycopods, mosses, ferns, horsetails) and more primitive plants are not yet sufficiently represented in the database and they are likely to have unique lineages of some types of TEs.
- The REXdb-based classification of TEs in RepeatExplorer and DANTE works reliably for higher plants but if you encounter conflicts or problems remember that:
 - It is better to classify TEs on the level which is reliable than to classify them incorrectly; pay attention to conflicts (e.g. due to chimerical clusters, nested insertions).
 - Non-autonomous TEs, possessing truncated CDS, and old/mutated TEs are difficult or impossible to classify using protein domain sequences.
 - Analyze all protein domains found in the TE to get the highest confidence of the classification.
 - **Let us know**