RepeatExplorer

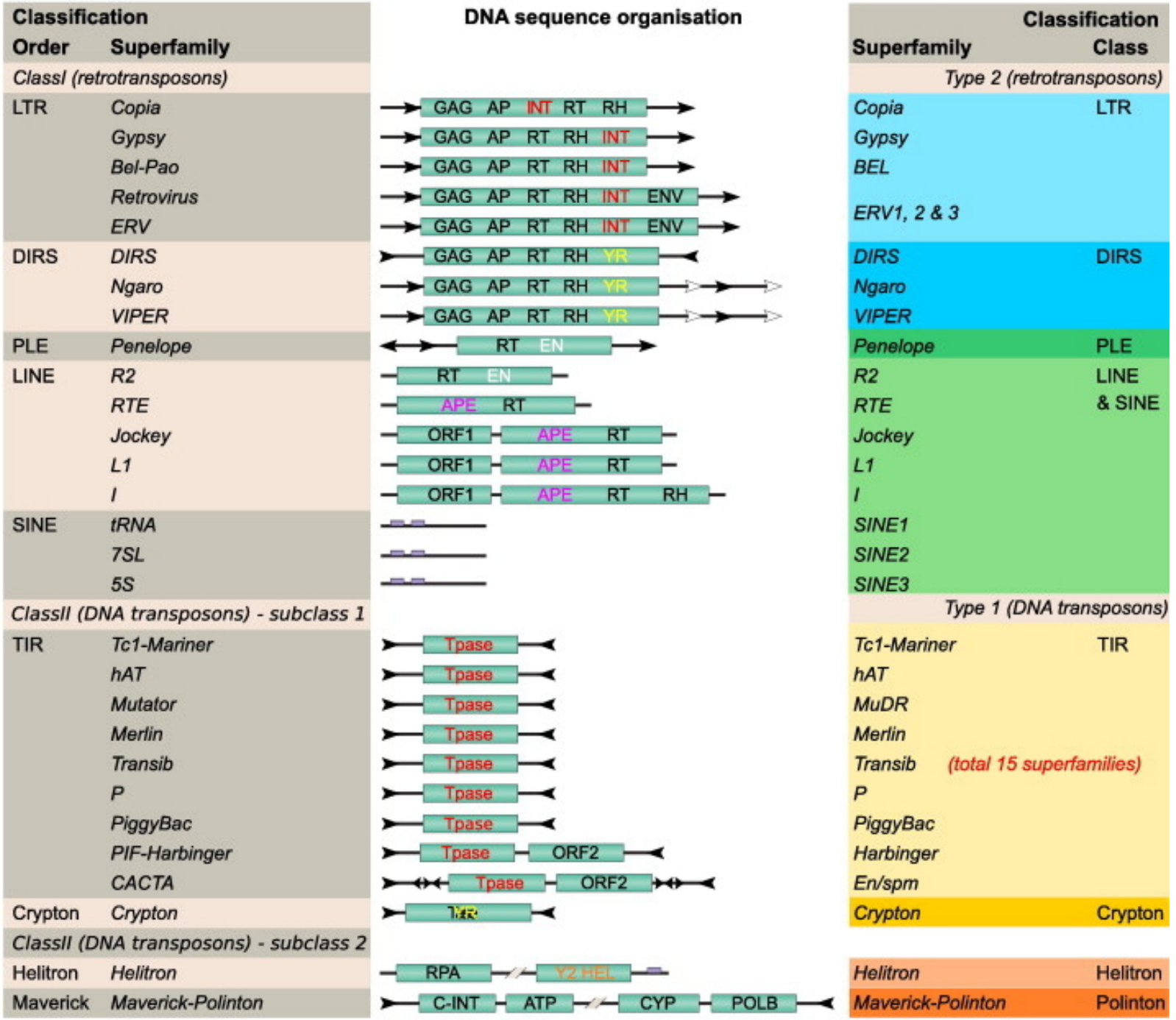
Discover repeats in your next generation sequencing data

REXdb database and transposon classification using conserved protein domains

Pavel Neumann

May 2019

Classification of transposable elements



"Unified"; Wicker et al. 2007

RepBase; Bao et al. 2015

giri REPBASE

WikiPoson

?

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

Criteria for classification of TEs

class (e.g. class I)	order (e.g. LTR)	superfamily (e.g. Gypsy)	family (e.g. Peabody)
Type of transposition: • copy and paste (I) • cut and paste (II)	Structure: • type of element termini Type of replication Protein domain types Phylogeny	Structure: • domain order • type of element termini	Sequence similarity: • 80-80-80 rule • RepeatMasker, CENSOR

- Although there is a consensus that the classification should be hierarchical it is not widely agreed what the hierarchy should reflect (structure, phylogeny....; see Piégu et al. Molecular Phylogenetics and Evolution 2015: 90-109).
    - Examples: Ginger (Gypsy INteGrasE Related) or Polintons/Mavericks (Gypsy-like INT + self-synthesizing = cut(ssDNA)-replicate(dsDNA)-paste).
- TEs are highly divergent in their DNA sequences - DNA databases are of limited use.
- **All types of autonomous TEs can be determined based on the presence and type of protein domains which are relatively conserved in sequence.**

REXdb: a reference database of transposable element protein domains

- Although not exhaustive, the database of protein domains from plant TEs is the largest data-set of its kind (it covers TEs from a wide range of Viridiplantae species; from Chlorophyta to Spermatophyta).
- A parallel databases for **Viridiplantae** and **Metazoa** (not yet released but optional in RepeatExplorer).
- **All sequences in REXdb are classified** according to the unified classification system (Wicker et al. 2007).
- It is implemented in RepeatExplorer where it serves for automatic classification of TEs. It can also be downloaded from the RepeatExplorer web page (http://repeatexplorer.org/).
- REXdb = fasta file of protein domain sequences + classification of each REXdb element in tab-delimited file

>Ty1-RT__**REXdb_ID1442**
WRQAMVDEMAALHSNGSWDLVVLPSGKSTVGCRWVYAVKVGPDGQVDRLKARLVAKGYTQ
VYGSDYGDTFSPVAKIASVRLLLSMAAMCSWPLYQLDIKNAFLHGDLAEEVYMEQPPGFV
AQGESGLVCRLRRSLYGLKQSPRAWFSRFSSVVQEFGMLRSTADHSVFYHHNSLGQCIYL
VVYVDDIVITGSDQDGIQKLKQHLFTHFQTKDLGKLKYFLGIEIAQSSSGVVLSQRKYAL
DILEETGMLDCKPVDTP

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: a reference database of transposable element protein domains

Latest REXdb release: 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: a reference database of transposable element 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: a reference database of transposable element 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

Neumann et al. Mobile DNA 2019 10:1

REXdb

Unified classification, Repbase

superfamily
(Ty1/Copia and Ty3/Gypsy)

domain order
- INT-RT-RH
- RT-RH-INT

phylogeny
- Bel/Pao

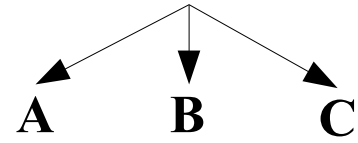
Phylogeny/ENV
- Retrovirus/ERV

lineage

phylogeny of polyprotein domains (RT, RH, INT)

additional support in lineage-specific fetaures

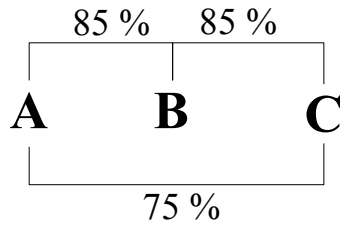
common ancestor



A, B and C belong to the same lineage

family
(e.g. Peabody)

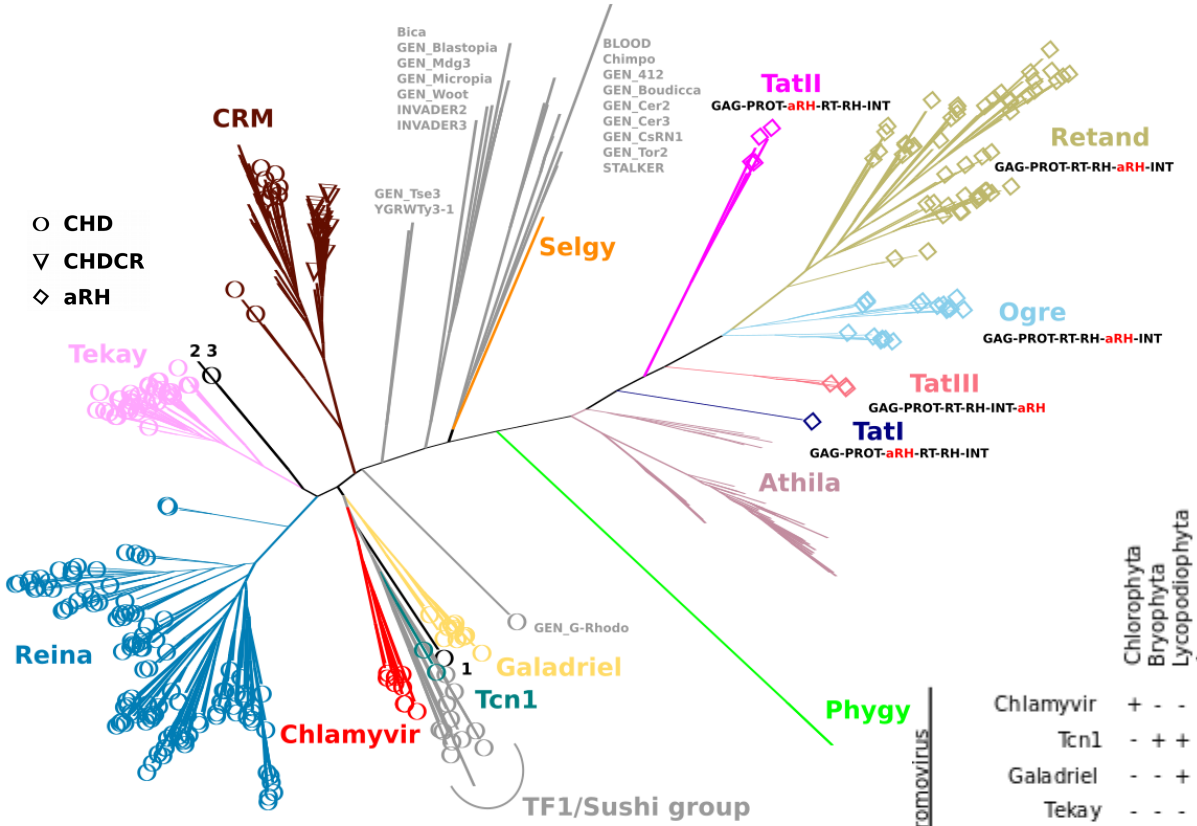
DNA sequence similarity
- 80-80-80 rule
- RepeatMasker, CENSOR
- species-specific
- naming problem (synonyms, homonyms)
- classification problem:



?

REXdb: classification of plant Ty3/gypsy retrotransposons into lineages

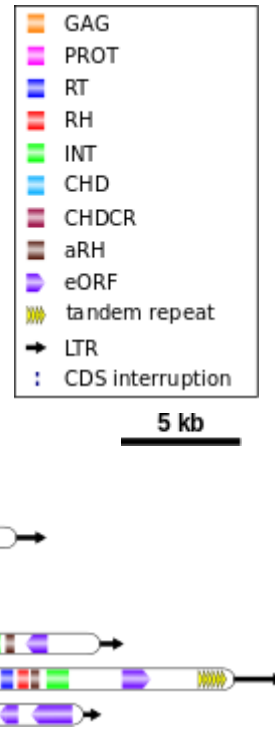
Neumann et al. Mobile DNA 2019 10:1



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)

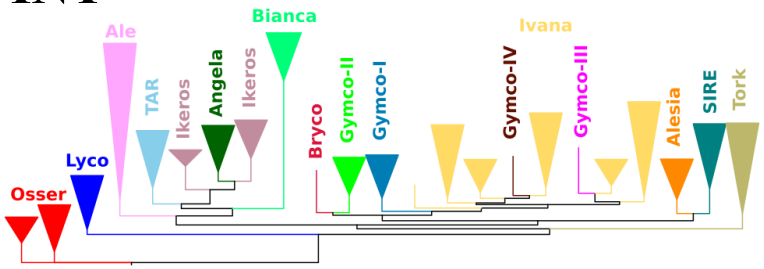
		Chlorophyta	Bryophyta	Lycopodiophyta	Acrogymnospermae	Magnoliophyta	Average size (kbp)	PBS	TSD	polyprotein ORFs
Chromovirus	Chlamyvir	+	-	-	-	-	6.7	?, Met, Ile	5	2
	Tcn1	-	+	+	-	-	6.4	self	5	2
	Galadriel	-	-	+	+	+	6.6	Met	5/4	1
	Tekay	-	-	-	+	+	11.5	Met, ?, Ile	5	1
	Reina	-	-	-	+	+	5.5	Met, ?, Thr	5	1
	CRM	-	-	-	+	+	6.8	Met, ?	5	1
Non-chromovirus OTA	Phygy	-	+	-	-	-	8.6	Trp, Thr, ?	4	ND
	Selgy	-	-	+	-	-	5.0	Ala, ?	4	2
	Athila	-	-	+	+	+	12.3	Asp, Asn, Ala	5	1
Tat	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



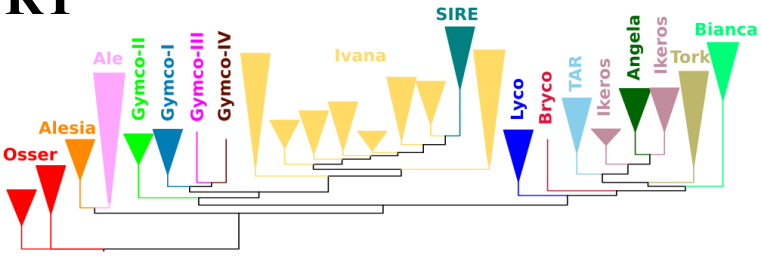
REXdb: classification of plant Ty1/copia retrotransposons into lineages

Neumann et al. Mobile DNA 2019 10:1

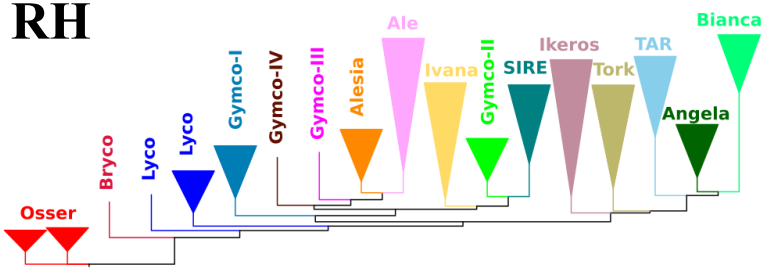
INT



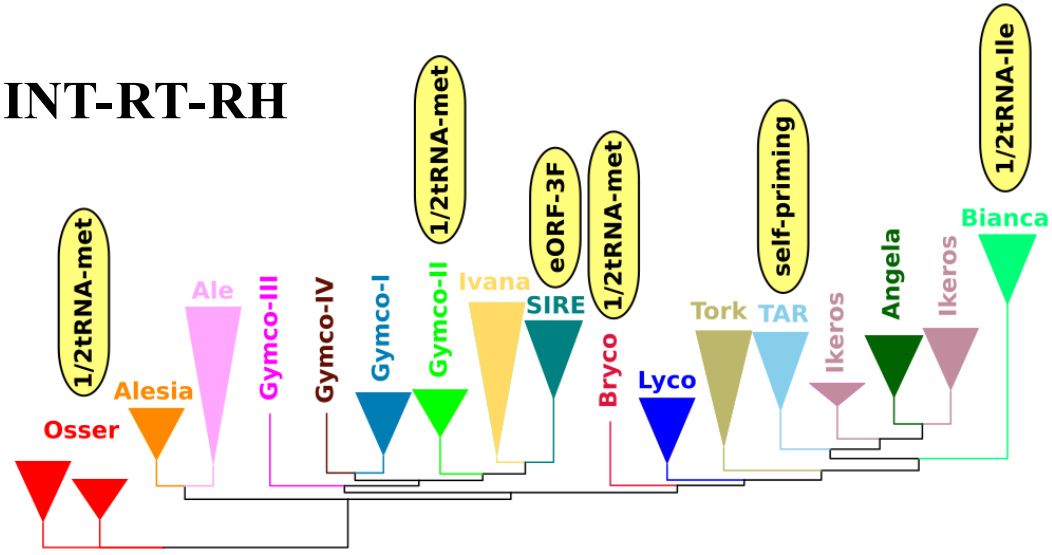
RT



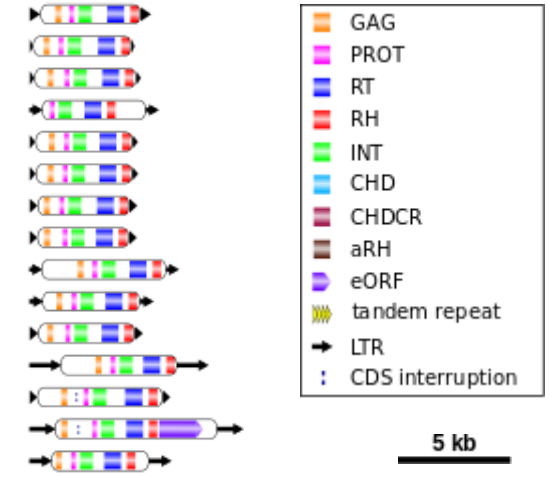
RH



INT-RT-RH



	Chlorophyta	Bryophyta	Lycopodiophyta	Acrogymnospermae	Magnoliophyta	Average size (kbp)	PBS	TSD	polyprotein ORFs
Osser	+	-	-	-	-	5.2	1/2Met	5	1
Bryco	-	+	-	-	-	5.3	1/2Met	5	ND
Lyco	-	-	+	-	-	4.8	Met	5	1
Gymco-I	-	-	-	+	-	5.8	Met	5	ND
Gymco-II	-	-	-	+	-	6.2	Met, 1/2Met, ?	5/4	ND
Gymco-III	-	-	-	+	-	5.0	Leu, Met	5	ND
Gymco-IV	-	-	-	+	-	5.0	Met, ?	5	ND
Ale	-	-	-	+	+	5.1	Met, ?	5	1
Ivana	-	-	-	+	+	5.1	Met, ?	5	1
Ikeros	-	-	-	+	+	6.9	Met, ?	5	1
Tork	-	-	-	+	+	5.4	Met, ?	5	1
Alesia	-	-	-	-	+	5.1	Met, ?	5	1
Angela	-	-	-	-	+	8.3	Met, ?	5	1
Bianca	-	-	-	-	+	6.1	1/2Ile	5	2
SIRE	-	-	-	-	+	9.9	Met?	5	2
TAR	-	-	-	-	+	6.3	self	5	1



RepeatExplorer2 clustering: Improved version or repeat discovery and characterization using graph based sequence clustering (Galaxy Version 1.0.0)

Options

NGS reads

7: Pasted Entry

Input file must contain fasta-formatted NGS reads. If paired end reads are used, reads must be interlaced and all pairs must be complete. Example of input data format is provided in the help below.

paired-end reads

Yes | No

Check if you are using pair reads and input sequences contain both read mates and left mates alternate with their right mates

Sample size

500000

Select taxon and protein domain database version (REXdb)

Viridiplantae version 3.0

- Viridiplantae version 3.0
- Viridiplantae version 2.2
- Metazoa version 3.0
- Metazoa version 2.0

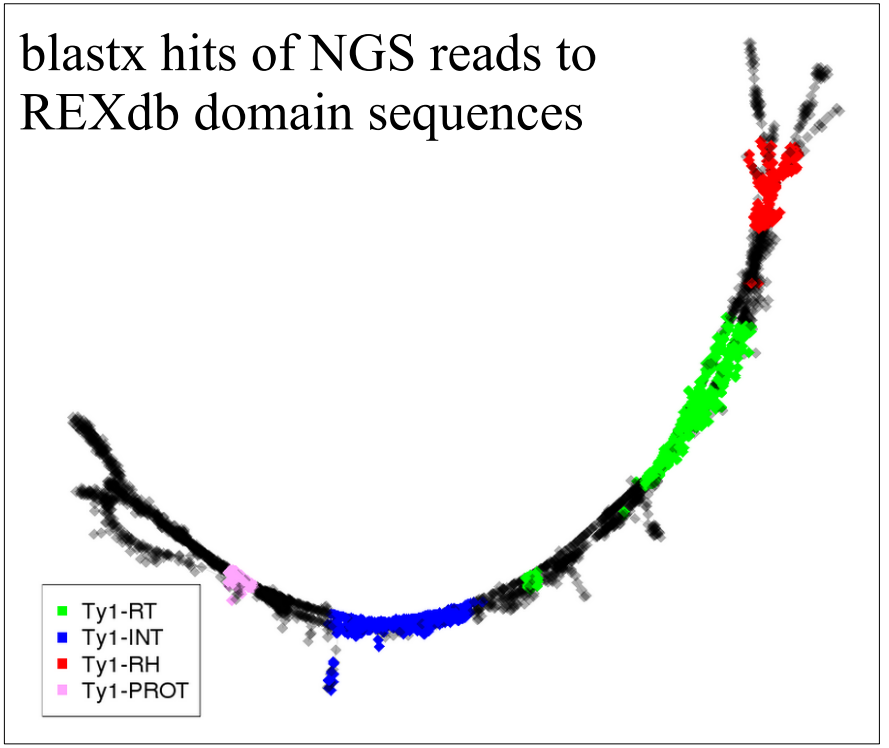
Modify parameters (optional)

-l select=1:ncpus=10:mem=32gb:scratch_local=50gb -l walltime=48:00:00 -q elixirre@pbs.elixir-czech.cz -v TAREAN_MAX_MEM=4000000,T

Execute

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



supercluster_report.html

SC	size	best_hit	Similarity_based_annotation	Tarean_annotation	clusters
11	11	5809 SIRE	| nhits | proportion | domains_string 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 | 1 (Ty1-GAG), 4 (Ty1-RH), °--SIRE | 1403 | 0.24 | 147 (Ty1-GAG), 420 (Ty1-INT), 103 (Ty1-PROT), 180 (Ty1-RH), 553 (Ty1-RT),		107, 183, 372, 53

Assembly annotation

DANTE

Protein Domains Finder Primary tool to find and classify ALL transposable elements protein domains in a DNA sequence (NO QUALITY FILTER USED)

Protein Domains Filter Filter Protein Domains Finder output according to a certain domain type and several alignment quality criteria

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

Protein Domains Finder Primary tool to find and classify ALL transposable elements protein domains in a DNA sequence (NO QUALITY FILTER USED) (Galaxy Version 1.0.0) Options

Choose your input sequence

7: Pasted Entry

Input DNA must be in proper fasta format, multi-fasta containing more sequences is allowed

Select taxon and protein domain database version (REXdb)

Viridiplantae_version_3.0

Execute

Protein Domains Filter Filter Protein Domains Finder output according to a certain domain type and several alignment quality criteria (Galaxy Version 1.0.0) Options

Choose primary GFF3 file of all domains from Protein Domains Finder

8: Unfiltered GFF3 file of ALL protein domains from dataset 7

Minimum identity

0.35

Protein sequence indentity threshold between input and mapped protein from db [0-1]

Minimum similarity

0.45

Protein sequence similarity threshold between input and mapped protein from db [0-1]

Minimum alignment length

0.8

Proportion of the hit length without gaps to the length of the database sequence [0-1]

Interruptions [frameshifts + stop codons]

3

Tolerance threshold per every starting 100 amino acids of alignment sequence

Maximal length proportion

1.2

Maximal proportion of alignment length to the original length of protein domain from database (including indels)

Select protein domain type

All

Filter a custom classification substring

You can type in an arbitrary string to filter a certain repetitive element type of any level. It must be a continuous substring in a proper format of Final_Classification attribute of GFF3 file. Classification levels are separated by | character

Execute

- based on **last** program
- classification is based on **multiple** top hits (the best hit + all other hits with score >= 80% of the score of the best hit)
- sequences are classified **on the deepest level** showing **no conflict** among hits:
  Class_I|LTR|Ty3/gypsy|chromovirus|**Reina**
  Class_I|LTR|Ty3/gypsy|chromovirus|**Tekay**
  = Class_I|LTR|Ty3/gypsy|chromovirus

Keep in mind

- 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.
- It is better to classify TEs on the level which is reliable than to classify them incorrectly; pay attention to conflicts (e.g. in nested insertions, chimerical clusters).
- Non-autonomous TEs, possessing truncated CDS, and old/mutated TEs are difficult or impossible to classify using protein domain sequences.
- Analyze all found protein domains to get the highest confidence of the classification.
- If you are not sure how to classify a given TE take a look at other features (pbs, introns, extra ORF).
- You should be the one who makes the final decision; do not blindly rely on the automatic outputs.