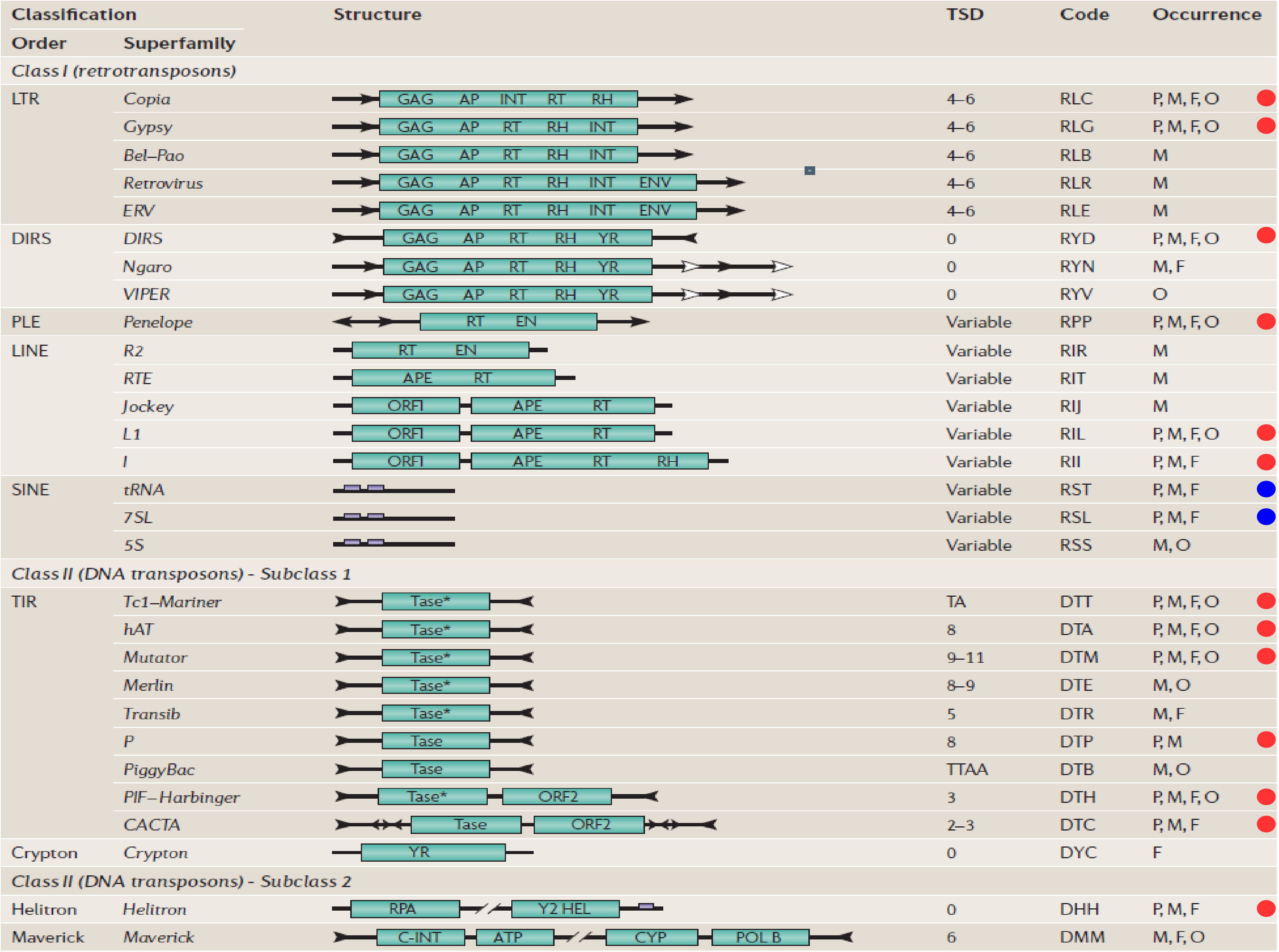
RepeatExplorer

Discover repeats in your next generation sequencing data

Classification of repetitive elements based on the analysis of protein domains

Pavel Neumann
May 2018

A unified classification system for eukaryotic transposable elements (Wicker et al. 2007)



Classification Order	 Superfamily	Structure	TSD	Code	Occurrence	
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	

Repbase classification system (Bao et al. 2015)

Group	Superfamily/clade
DNA transposon	Academ, Crypton (CryptonA, CryptonF, CryptonI, CryptonS, CryptonV), Dada, EnSpm/CACTA, Ginger1, Ginger2, Harbinger, hAT, Helitron, IS3EU, ISL2EU, Kolobok, Mariner/Tc1, Merlin, MuDR, Novosib, P, piggyBac, Polinton, Sola (Sola1, Sola2, Sola3), Transib, Zator, Zisupton
LTR retrotransposon	BEL, Copia, DIRS, Gypsy, ERV1, ERV2, ERV3, ERV4, Lentivirus
Non-LTR retrotransposon	Ambal, CR1, CRE, Crack, Daphne, Hero, I, Ingi, Jockey, Kiria, L1, L2, L2A, L2B, Loa, NeSL, Nimb, Outcast, Penelope, Proto1, Proto2, R1, R2, R4, RandI/Dualen, Rex1, RTE, RTETP, RTEX, Tad1, Tx1, Vingi SINE (SINE1/7SL, SINE2/tRNA, SINE3/5S, SINE4, SINEU)

Criteria for classification of TEs

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

- Although there is a consensus that the classification should be hierarchical it is not widely agreed what the hierarchy should reflect (structure, phylogeny)
- All types of autonomous TEs can be determined based on the presence of conserved protein domains

Databases of protein domains in RepeatExplorer

- Although not exhaustive, the database of protein domains from plant TEs **is the most comprehensive** data-set of its kind (it covers TEs from a wide range of Viridiplantae species; from Chlorophyta to Spermatophyta)

- A parallel database for **Metazoa**. It is being tested ...

- **All sequences in the database are classified** into groups (superfamilies), following the unified classification system

RepeatExplorer: database of protein domains (Viridiplantae)

- **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

RepeatExplorer: database of protein domains (Metazoa)

- **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

RepeatExplorer: basic classification of TEs into superfamilies

- 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

Viridiplantae + Metazoa

Viridiplantae

Metazoa

- 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

Sub-classification of plant LTR retrotransposons

superfamily
(Ty1/Copia and Ty3/Gypsy)

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

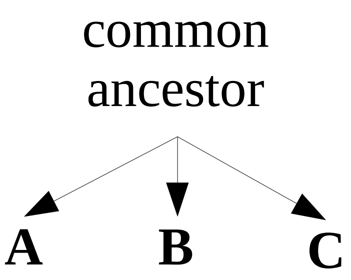
phylogeny
- Bel/Pao

Phylogeny/ENV
- Retrovirus/ERV

REXdb: lineage

phylogeny of polyprotein domains (RT, RH, INT)

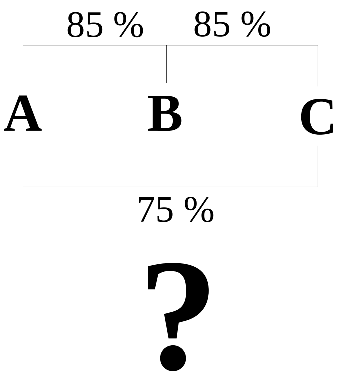
additional support in lineage-specific fetaures



A, B and C belong to the same lineage

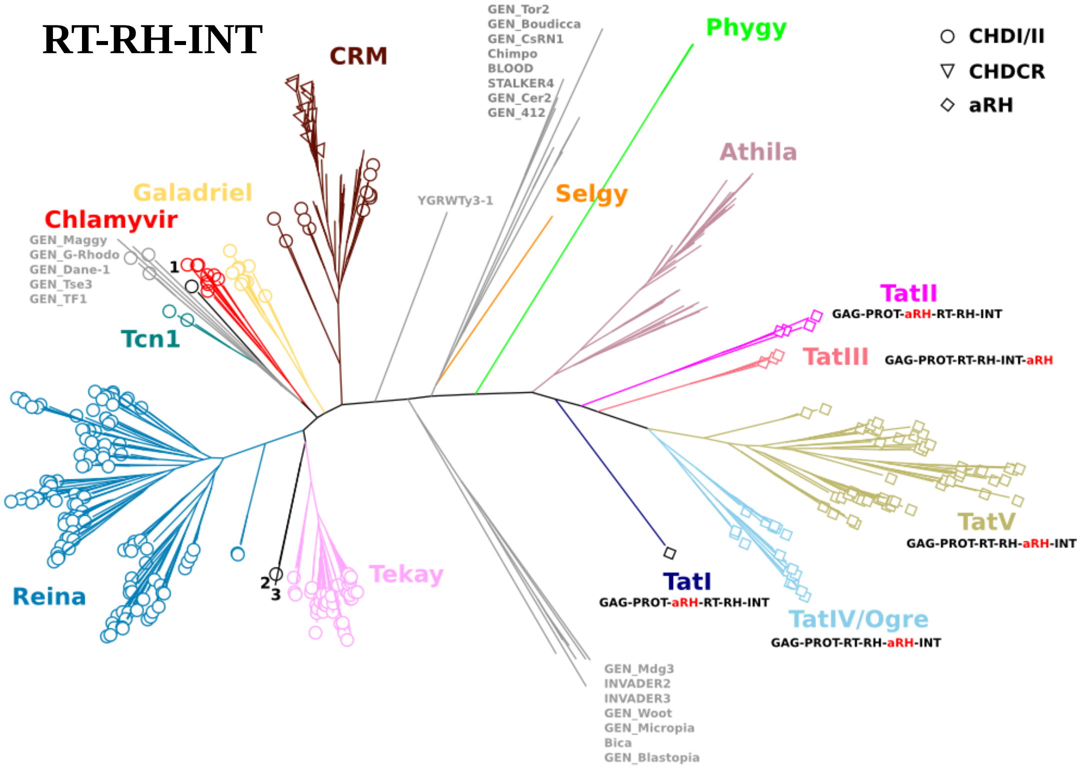
family
(e.g. Peabody)

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

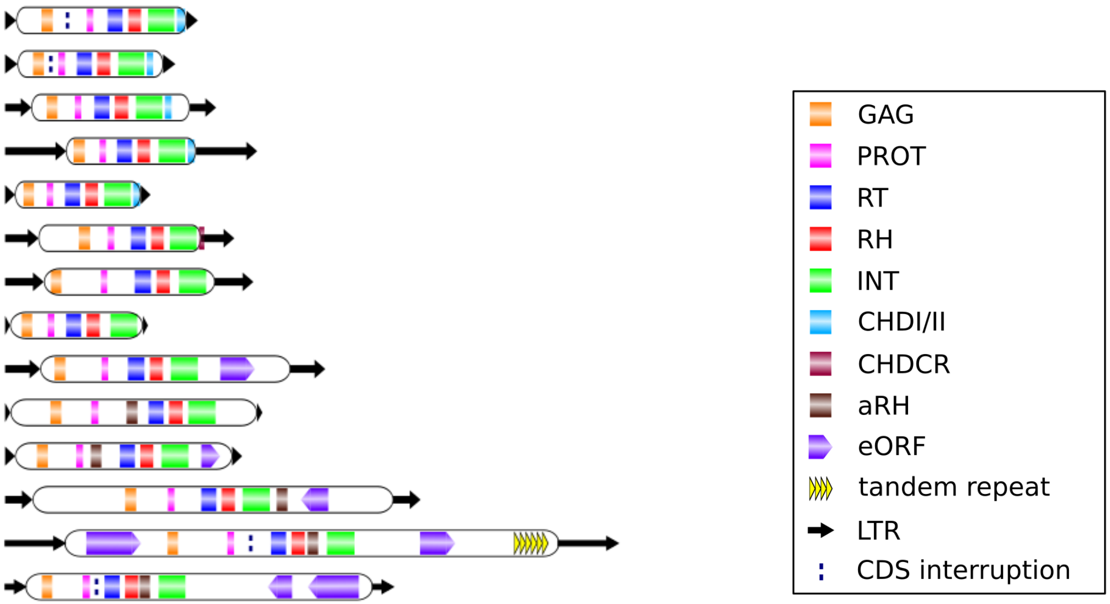


Sub-classification of Ty3/gypsy retrotransposons

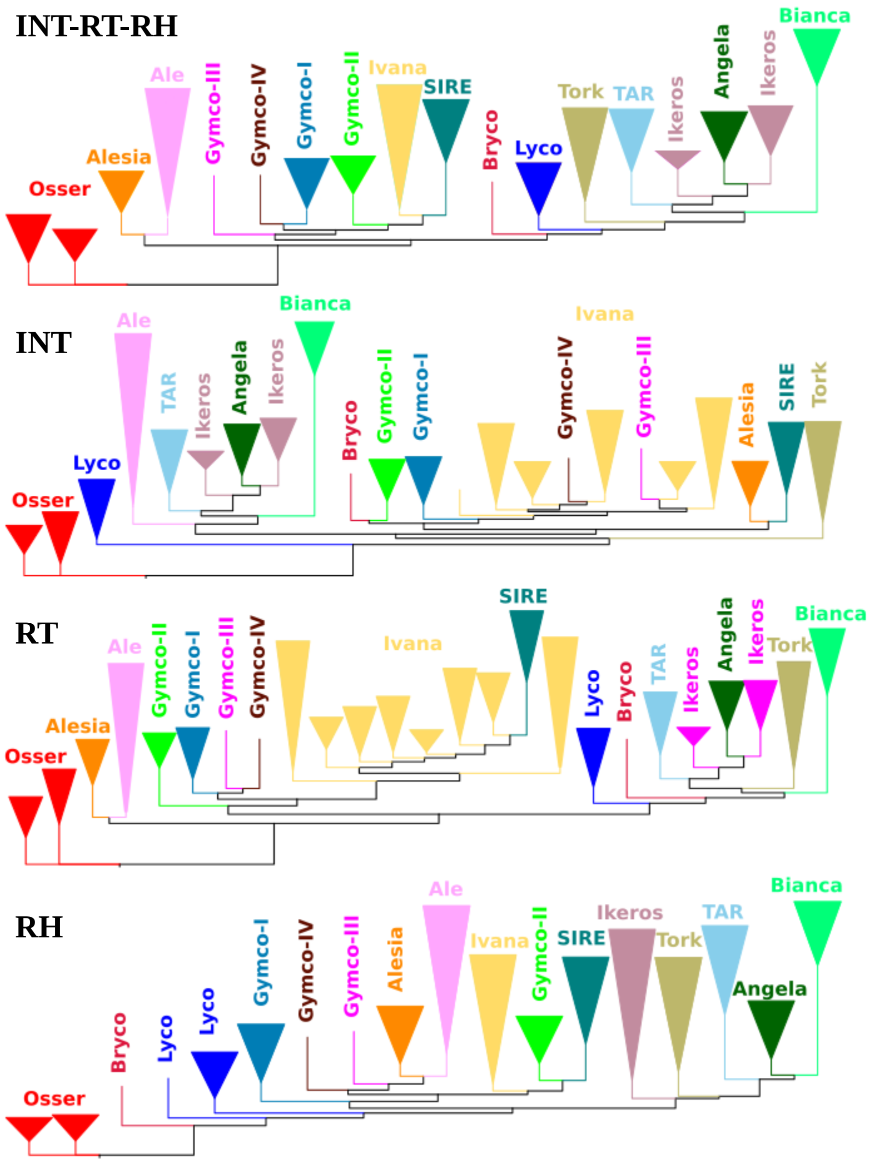
- phylogenies inferred based on RT, RH, INT, RT-RH-INT are in good agreement
- strong support in lineage-specific features (chromodomain, aRH, eORF, PBS)

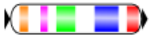
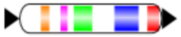
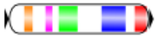
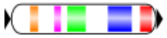
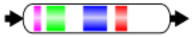
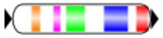
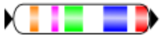
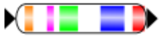
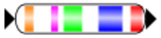
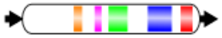
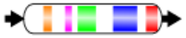
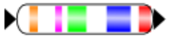
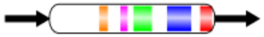
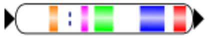
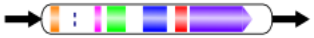
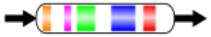


				Chlorophyta	Bryophyta	Lycopodiophyta	Acrogymnospermae	Magnoliophyta	Average size (kbp)	PBS	TSD	polyprotein ORFs
Chromovirus			Chlamyvir	+	-	-	-	-	6.7	?, Met, Ile	5	2
			Tcn1	-	+	+	-	-	6.4	self	5	2
			Galadriel	-	-	+	+	+	6.6	Met	5/4	1
			Tekay	-	-	-	+	+	11.5	Met, ?, Ile	5	1
			Reina	-	-	-	+	+	5.5	Met,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



Sub-classification of Ty1/copia retrotransposons

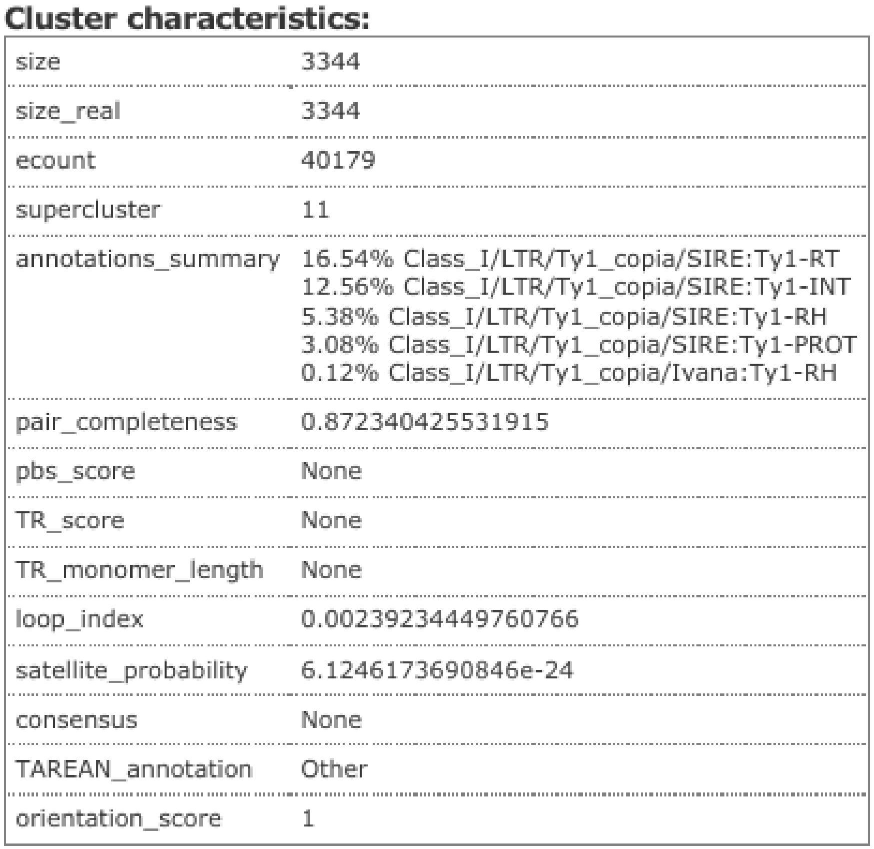


	Chlorophyta	Bryophyta	Lycopodiophyta	Acrogymnospermae	Magnoliophyta	Average size (kbp)	PBS	TSD	polyprotein ORFs	
Osser	+	-	-	-	-	5.2	1/2Met	5	1	
Bryco	-	+	-	-	-	5.3	1/2Met	5	ND	
Lyco	-	-	+	-	-	4.8	Met	5	1	
Gymco-I	-	-	-	+	-	5.8	Met	5	ND	
Gymco-II	-	-	-	+	-	6.2	Met, 1/2Met, ?	5/4	ND	
Gymco-III	-	-	-	+	-	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	

🟧 GAG 🟩 INT 🟥 RH → LTR

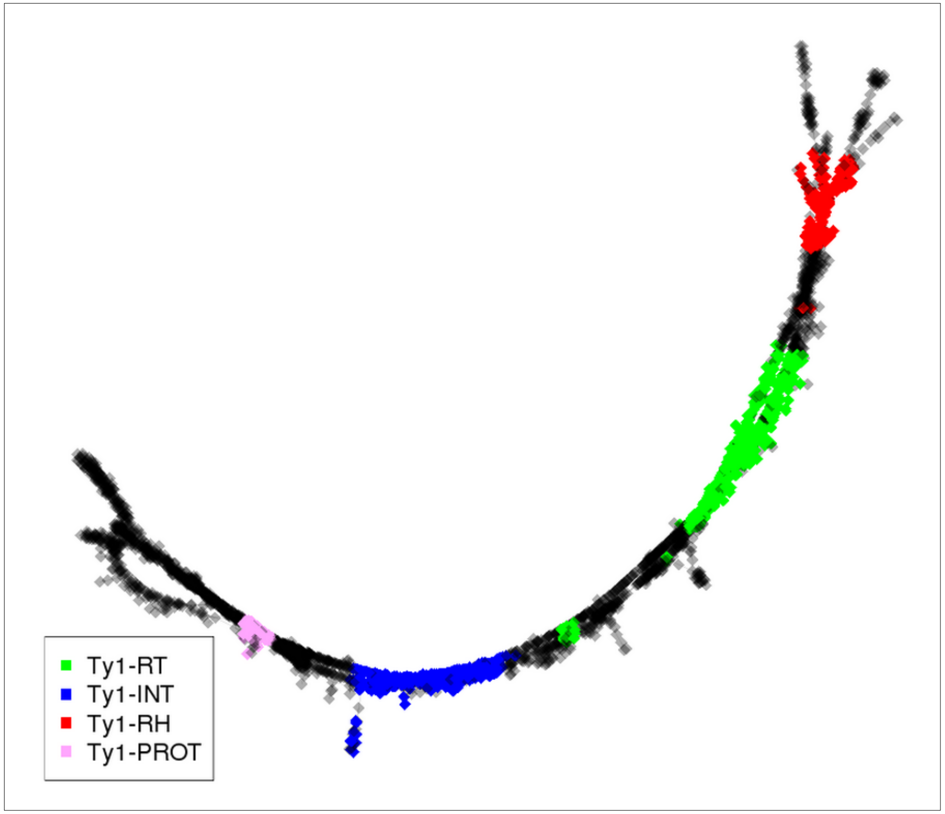
🟪 PROT 🟦 RT ➤ eORF : CDS interruption

RepeatExplorer: Automatic analysis of TE protein domains (blastx using NGS reads)

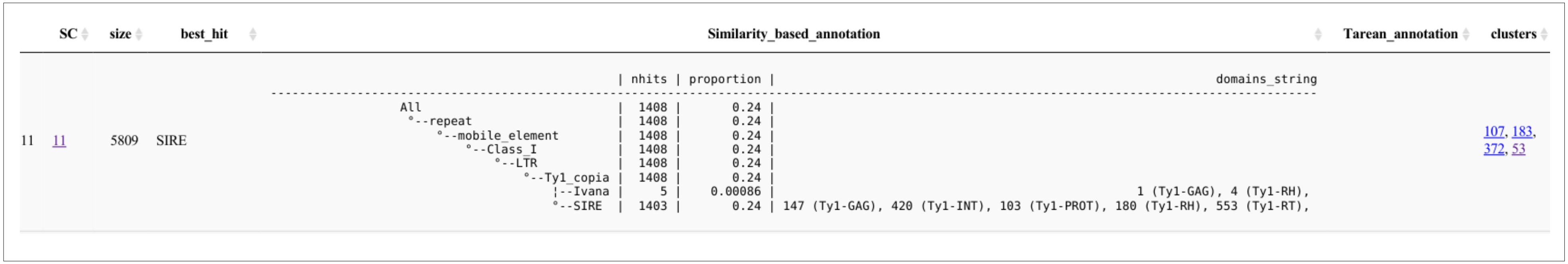


Cluster characteristics:

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

RepeatExplorer: DANTE

- Protein domains search
  - optional
  - based on **last** program (fasty in the previous version)
  - classification is based on **multiple** top hits (80% of the best score)
  - sequences are classified **on the deepest level** showing **no conflict** among hits (Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat|TatV)
  - output is data-rich **gff3** file which can be used in genome browsers

- Protein domains filter
  - multiple criteria for filtering
  - generates filtered gff3 file and protein domain sequences in fasta file
  - protein sequences of reference elements are not included in the fasta file (they are present in the gff3 file)

Note: DANTE tool can be used not only for the analysis of contigs generated by RepeatExplorer but also for any other kind of DNA sequences including whole genome assemblies.

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