RepeatExplorer

Classification of repetitive elements based on the analysis of protein domains

Pavel Neumann
May 2017

A unified classification system for eukaryotic transposable elements (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

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, Kiri a, 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 current 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

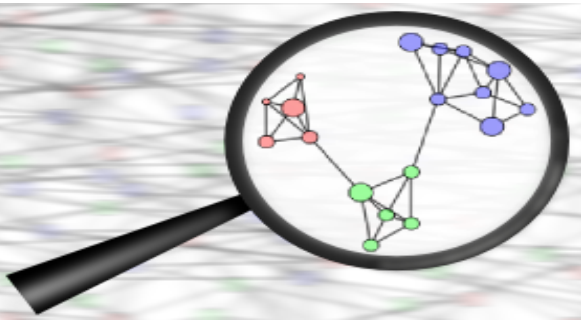
- Although there is a consensus that the classification should be hierarchical it is not widely agreed what the hierarchy should reflect (structure, phylogeny, protein versus DNA sequences)

- There is a huge a gap in classification of LTR retrotransposons on the level between superfamilies and families (studies exist but are ignored)

- RepeatExplorer classification is based on protein domains typical for individual types (superfamilies) of TEs

RepeatExplorer

Discover repeats in your next generation sequencing data



Database of protein domains

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

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

- LTR retrotransposons are further classified into **phylogenetic lineages** (this level fills the gap between superfamilies and families)

RepeatExplorer: database of protein domains

- **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: standard classification of TEs

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

- Class_II|Subclass_1|TIR|EnSpm/CACTA
- 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|Tc1/Mariner
- Class_II|Subclass_1|TIR|hAT
- Class_II|Subclass_2|Helitron

Subclassification of LTR retrotransposons

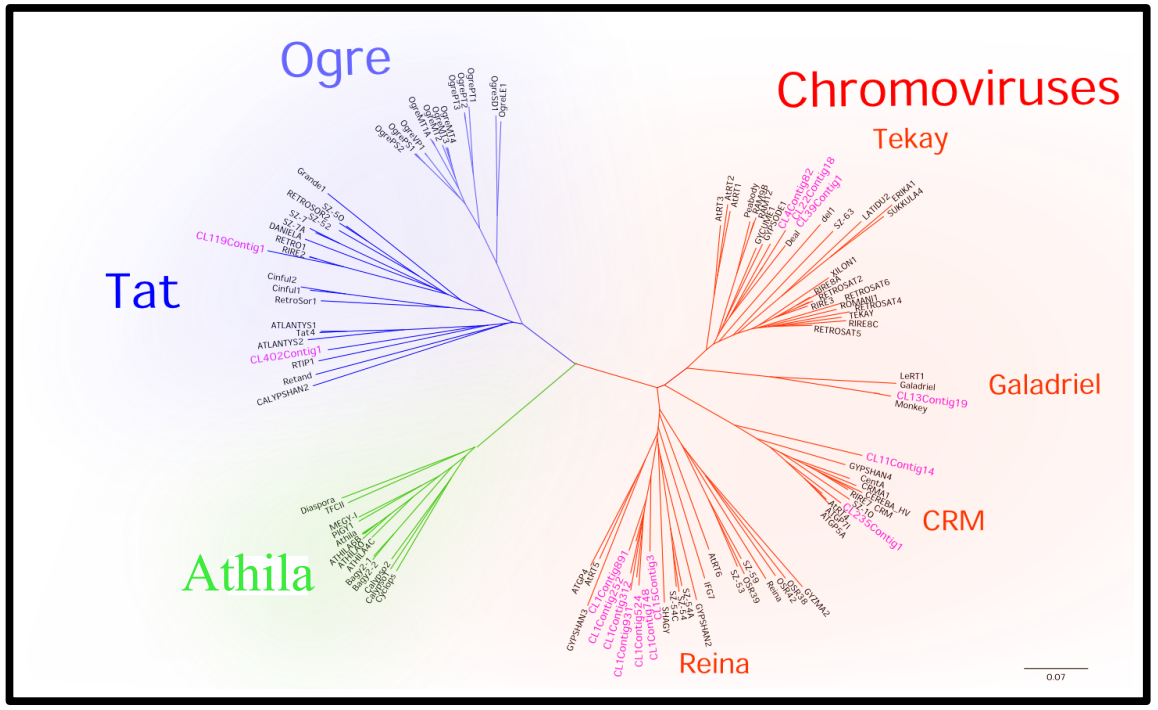
- Current classification is based on phylogenetic analyzes of RT, RH, and INT domains followed by a calculation of “galled network” (in the version presented the last year it was inferred from concatenated PROT-RT-RH-INT domains)
- Structural features (pbs, aRH, chromodomain) are used as secondary criteria (e.g. chromovirus without chromodomain is still chromovirus!)

- Class_I|LTR|Ty1/copia|
  - Ale 1787
  - Alesia 31
  - Angela 540
  - Bianca 260
  - Bryco 17
  - Gymco-I 14
  - Gymco-II 43
  - Ikeros 314
  - Ivana 851
  - Osser 19
  - SIRE 734
  - TAR 203
  - Tork 563
  - Ty1-outgroup 34

- Class_I|LTR|Ty3/gypsy|
  - chromovirus|CRM 736
  - chromovirus|Chlamyvir 44
  - chromovirus|Galadriel 270
  - chromovirus|Reina 708
  - chromovirus|Tcn1 1500
  - chromovirus|Tekay 782
  - chromovirus|chromo-outgroup 7
  - chromovirus|chromo-unclass 51

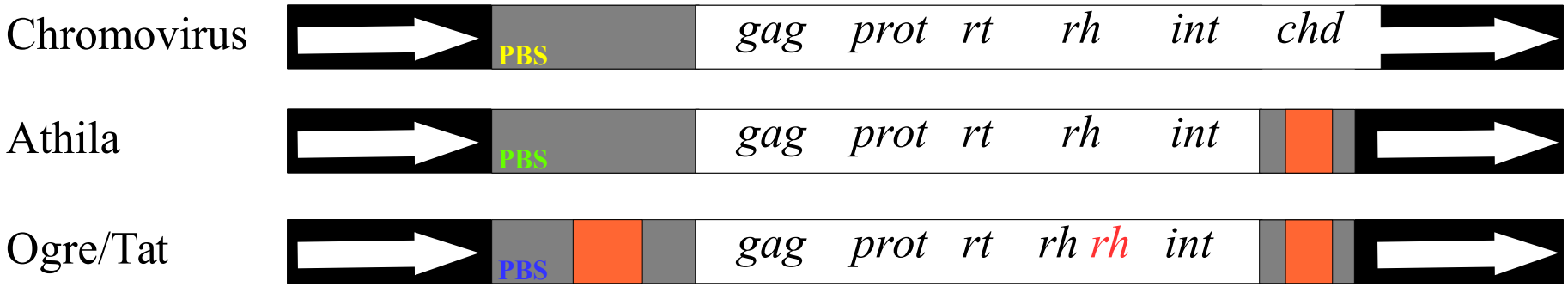
  - non-chromovirus|OTA|Athila 1046
  - non-chromovirus|OTA|Ogre/Tat|TatI 4
  - non-chromovirus|OTA|Ogre/Tat|TatII 27
  - non-chromovirus|OTA|Ogre/Tat|TatIII 39
  - non-chromovirus|OTA|Ogre/Tat|TatIV/Ogre 766
  - non-chromovirus|OTA|Ogre/Tat|TatV 2155
  - non-chromovirus|Phygy 186
  - non-chromovirus|Selgy 114
  - non-chromovirus|nonchromo-outgroup 18

Previous classification of plant Ty3/Gypsy retrotransposons



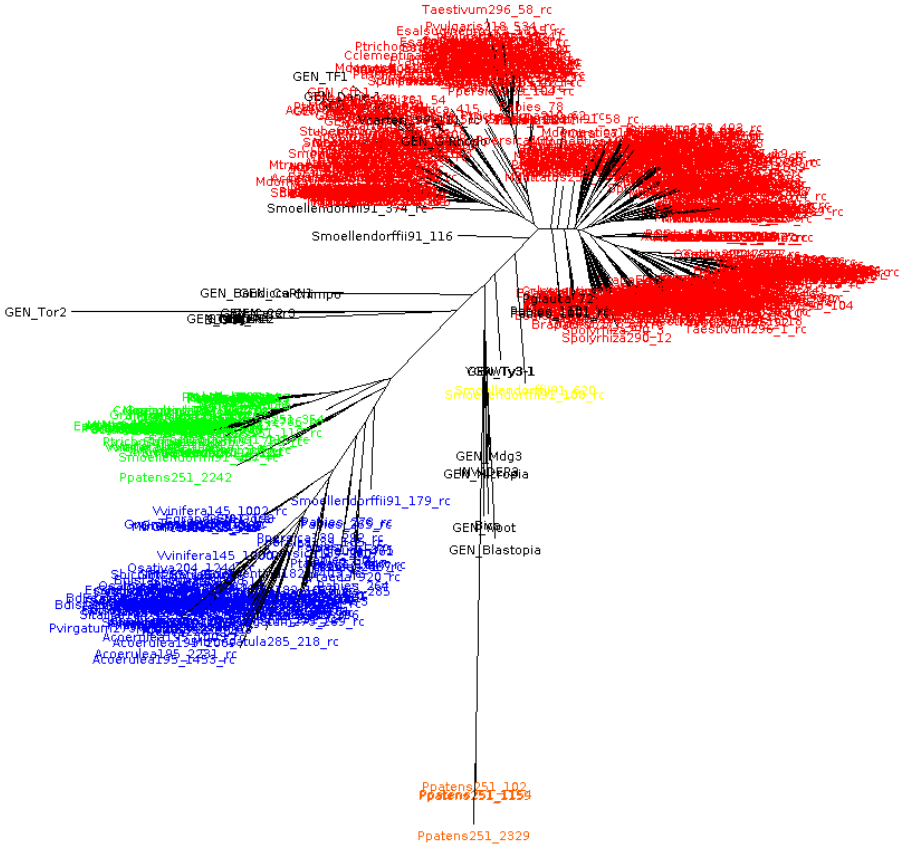
Classification according Llorens et al. (2011)

- It is based on phylogenetic analyzes of protein domain sequences
- It is supported by differences in the structure of the elements

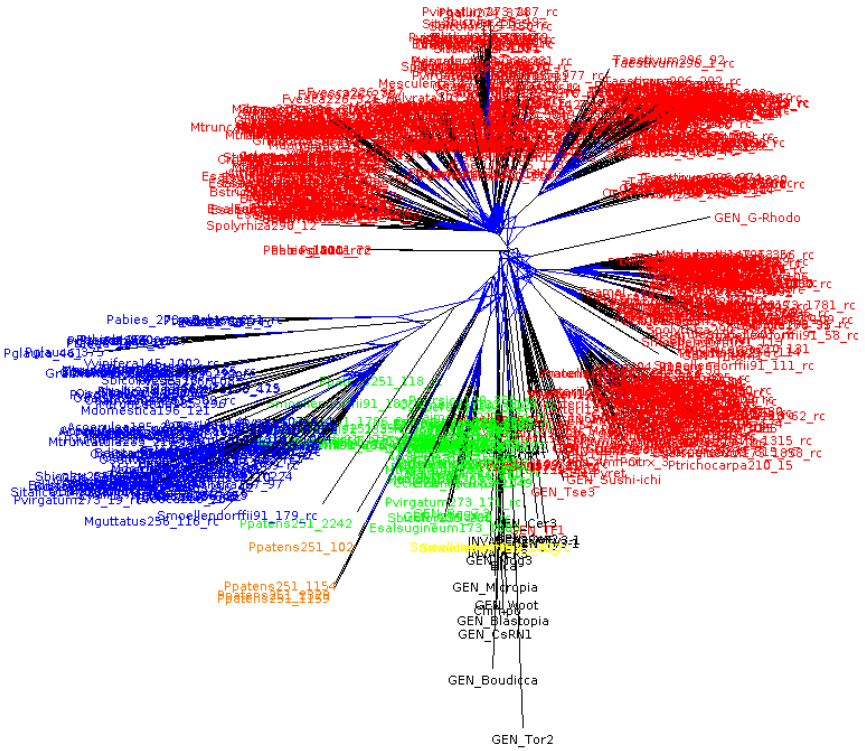


Current classification of Ty3/gypsy elements

RT domain



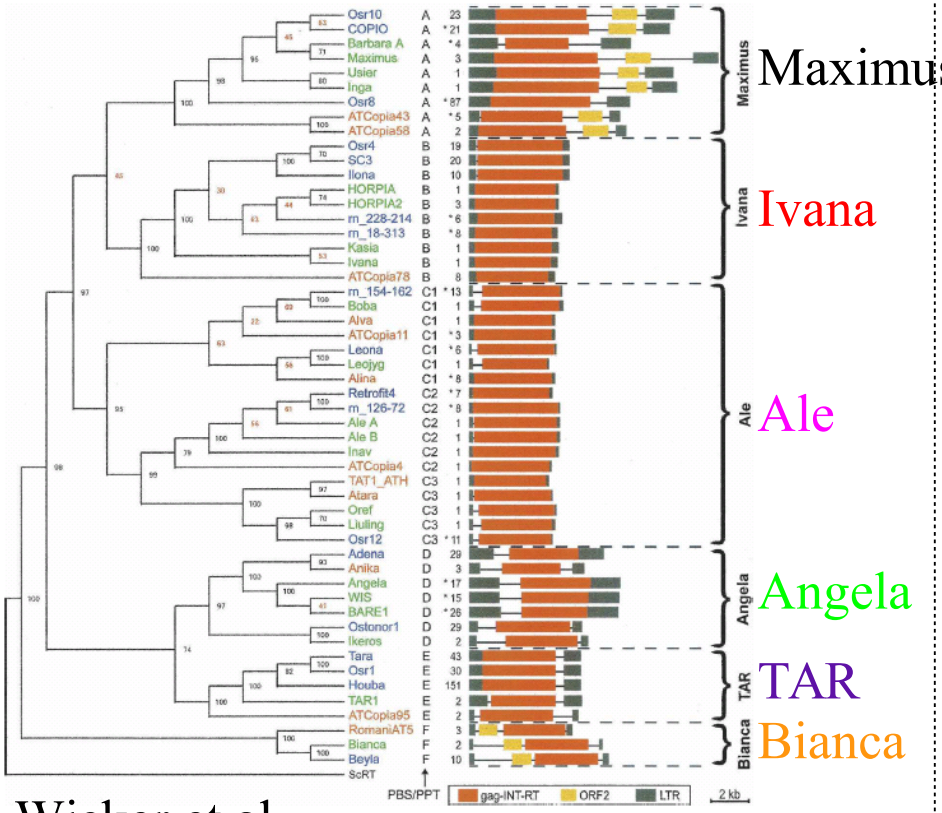
galled network from RT, RH and INT



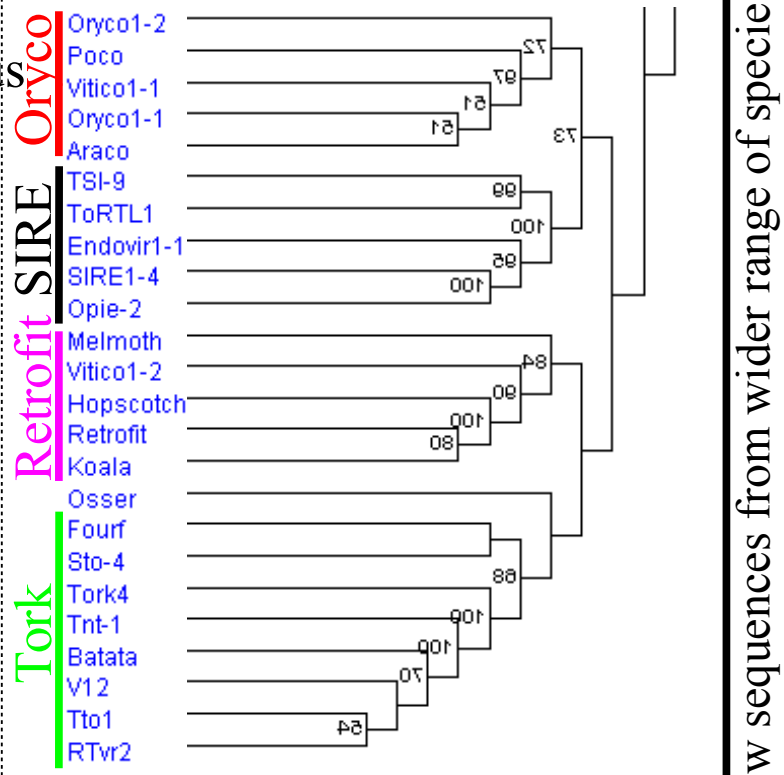
chromovirus Selgy Phygy Athila Ogre/Tat

Previous classification of plant Ty1/Copia retrotransposons

many sequences from a few species



Wicker et al.



Llorens et al.

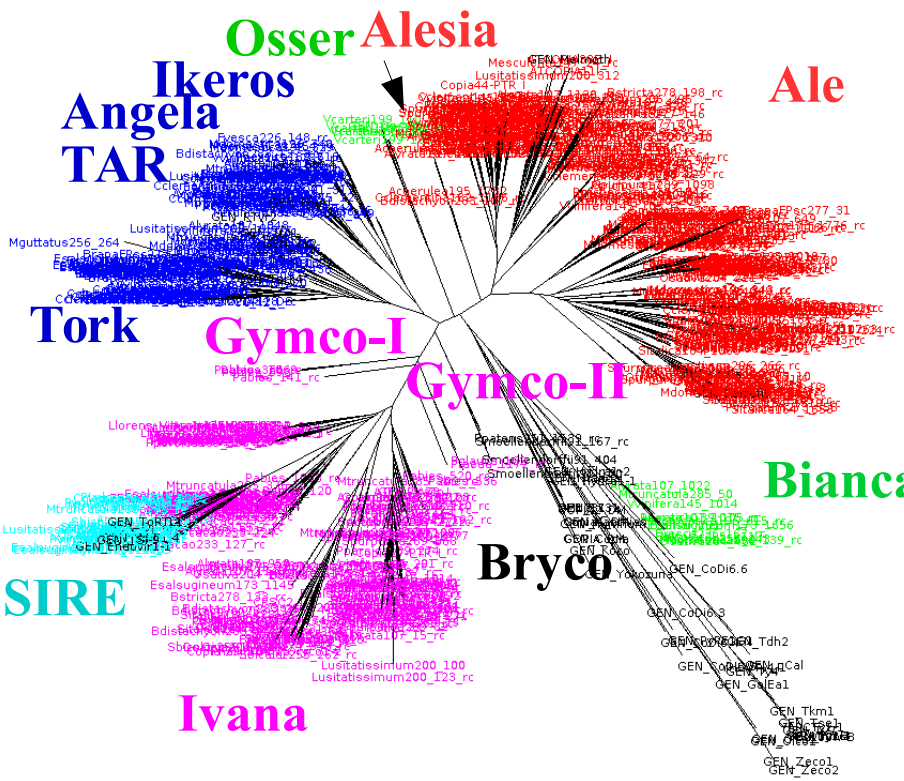
a few sequences from wider range of species

Classification according Wicker et al. (2007) and Llorens et al. (2011)

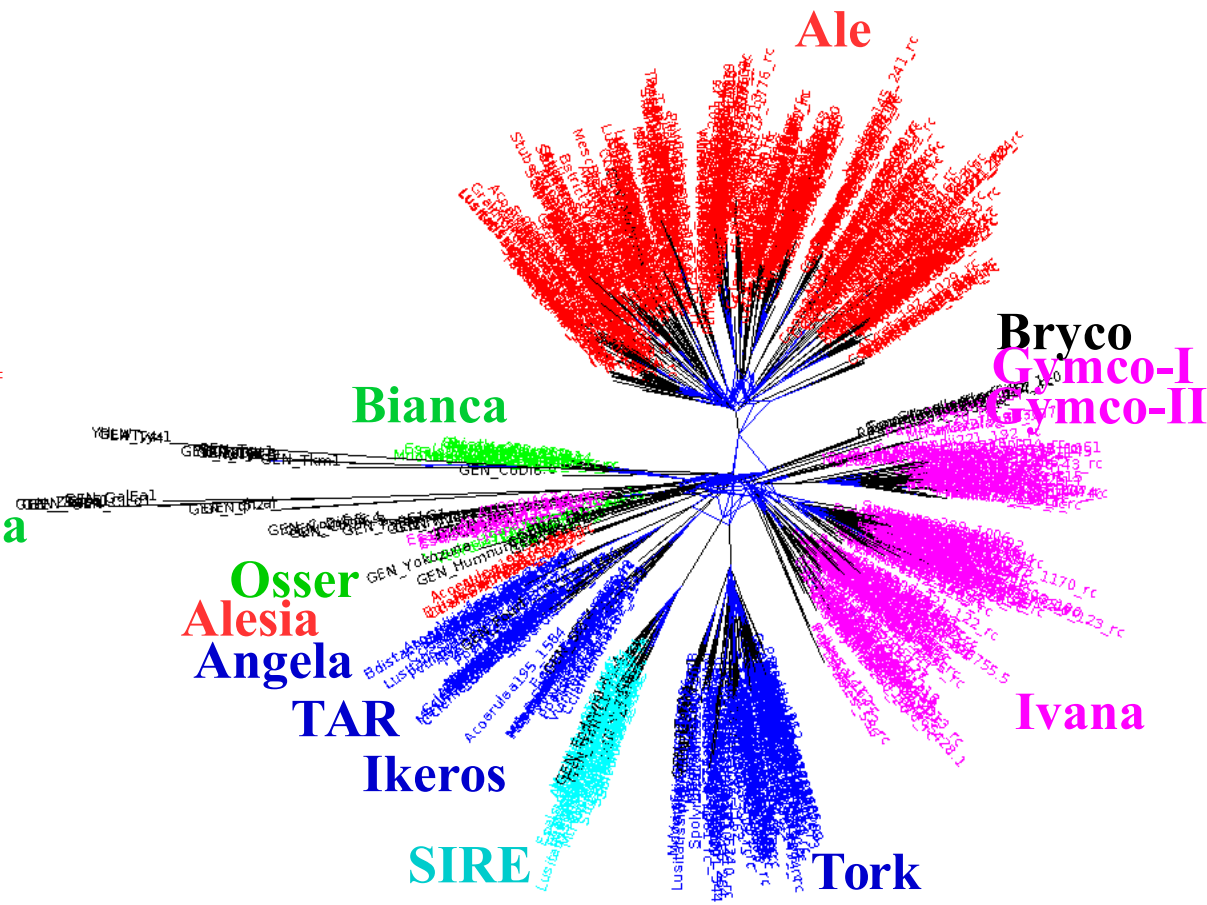
- It is based on phylogenetic analysis of protein domain sequences
- Structural differences among the lineages are less dramatic than in Ty3/Gypsy

Current classification of Ty1/copia elements

RT domain



galled network from RT, RH and INT



Reticulate evolution?

RepeatExplorer: classification based on protein domains

Automatic

- integrated in the clustering pipeline
- based on blastx using sequence reads
- the result is used for classification of clusters
- hits are short (100 bp = 33 aa)

Optional (protein domains tool)

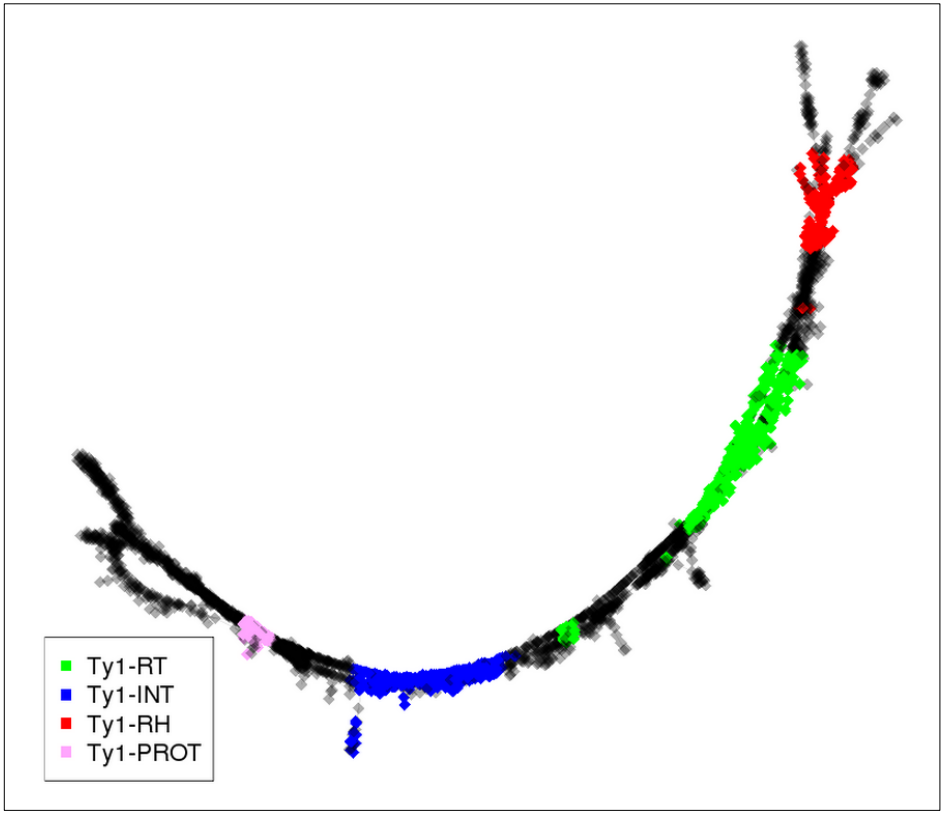
- it cannot be used in the clustering pipeline
- flexible (it accepts any type of DNA sequences in fasta)
- the result can be used to verify and/or to refine the automatic classification
- hits can cover entire domains (potentially more sensitive and accurate)

We need your feedback. If the classification does not work well for your plant species, let us know.

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
			| 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),		
11	11	5809 SIRE			107, 183, 372, 53

RepeatExplorer: protein domains tools

- Protein domains search
  - optional
  - based on **last** program (fasty in the previous version)
  - **one database** of all protein domain sequences
  - classification is based on **multiple** top hits (80% of the best score)
  - a region with hit to a protein domain is 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)
  - phylogenetic analysis is not performed (a difference from the previous version)

Note that protein domain tools 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

- the database includes mostly plant TEs, therefore its use for classification of non-plant elements is very limited
- 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 TEs

optional protein domains tools

- 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)
- 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