Diagnostic features of repetitive elements

Part I: Classification of repetitive elements and their specific features

Diagnostic features of repetitive elements

Organization in the genome

Tandem repeats

Dispersed repeats

rep. unit = monomer

HOR

if highly amplified to large arrays
=> satellite DNA

Transposable elements

- various types
- enormous number of divergent families

Diagnostic features of repetitive elements

Transposable elements

They are conserved in structure but not in DNA sequence - DNA sequence databases (Repbase, Dfam) are of limited use.

Termini: direct repeats, inverted repeats, specific sequence (e.g. TG/CA, poly-A, CACTA)

Internal part: coding region - structure (single ORF, multiple ORFs, exon/introns), protein types, domains
specific sequences (e.g. pbs, ppt)

Target site: sequence, duplication, duplication length

Diagnostic features of repetitive elements

Transposable (mobile) elements: structure and coding capacity is related to the mechanism of transposition

Class I (copy and paste)
LTR retrotransposons
Non-LTR retrotransposons

Class II (cut and paste)
DNA transposons
Helitrons

Diagnostic features of repetitive elements

Class I: LTR retrotransposons

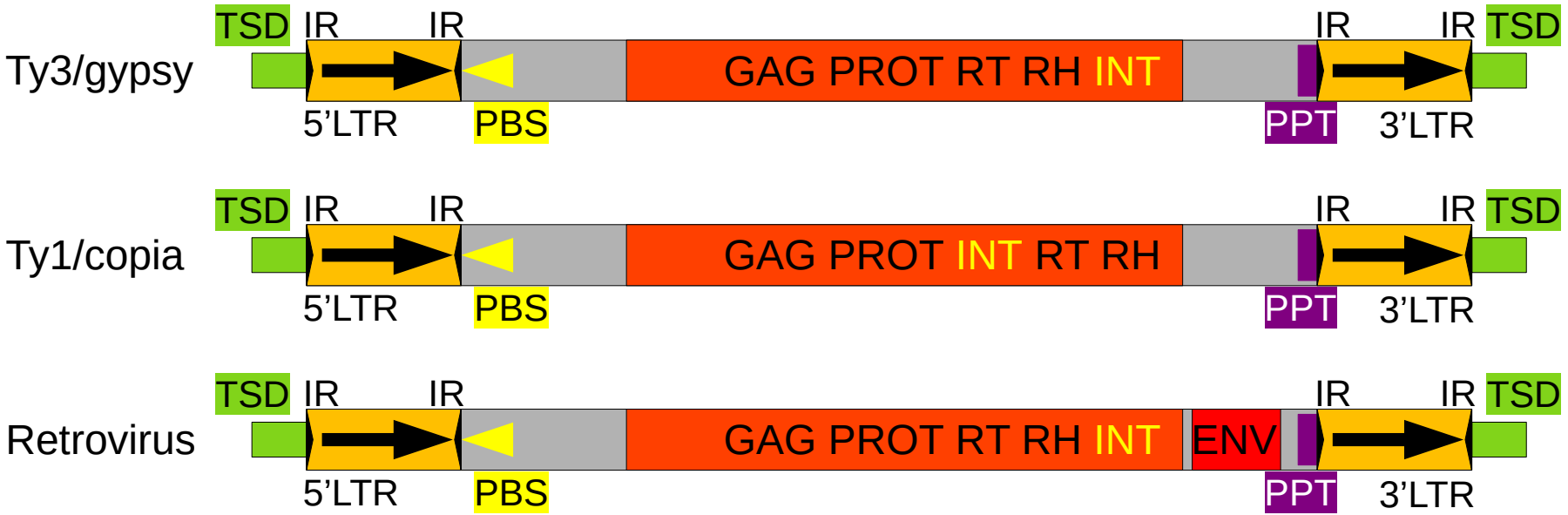
1. Autonomous element

Proteins
GAG: Matrix
Capsid
Nucleocapsid
Pol: Protease (Prot)
Reverse transcriptase (RT)
RNAse H (RH)
Integrase (INT)

Diagnostic features of repetitive elements

Class I: LTR retrotransposons

1. Autonomous elements



Diagnostic features of repetitive elements

Class I: LTR retrotransposons

Type of PBS
- 3’ end of tRNA (various types), 3’ end of half tRNA or self-priming

Extra domains in Pol
- aRH, chromodomain, CR chromodomain

Extra ORF
upstream or downstream of Gag-Pol, + or - orientation

Additional features
structure of coding region, element length, LTR length, presence of tandem repeats

Diagnostic features of repetitive elements

Class I: LTR retrotransposons

Neumann *et al. Mobile DNA* (2019) 10:1
https://doi.org/10.1186/s13100-018-0144-1

Mobile DNA

RESEARCH Open Access

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 Hoštáková and Jiří Macas

Occurrence and/or type of distinct sequence and structural features correlates with phylogenies inferred from the three pol protein domains.

Protein domain sequences can be used for fine classification of LTR-retrotransposons in plants

REXdb

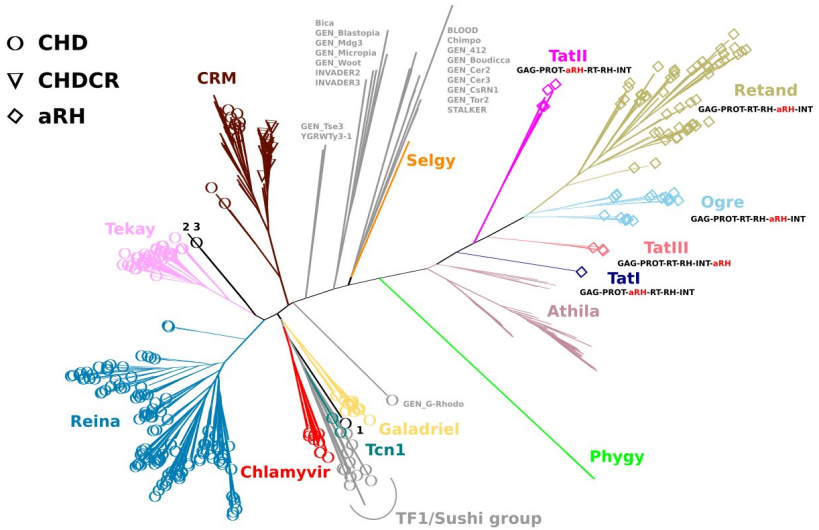
RepeatExplorer

Discover repeats in your next generation sequencing data

Domain based ANnotation of Transposable Elements - DANTE Tool for annotation of transposable elements based on the similarity to conserved protein domains database.

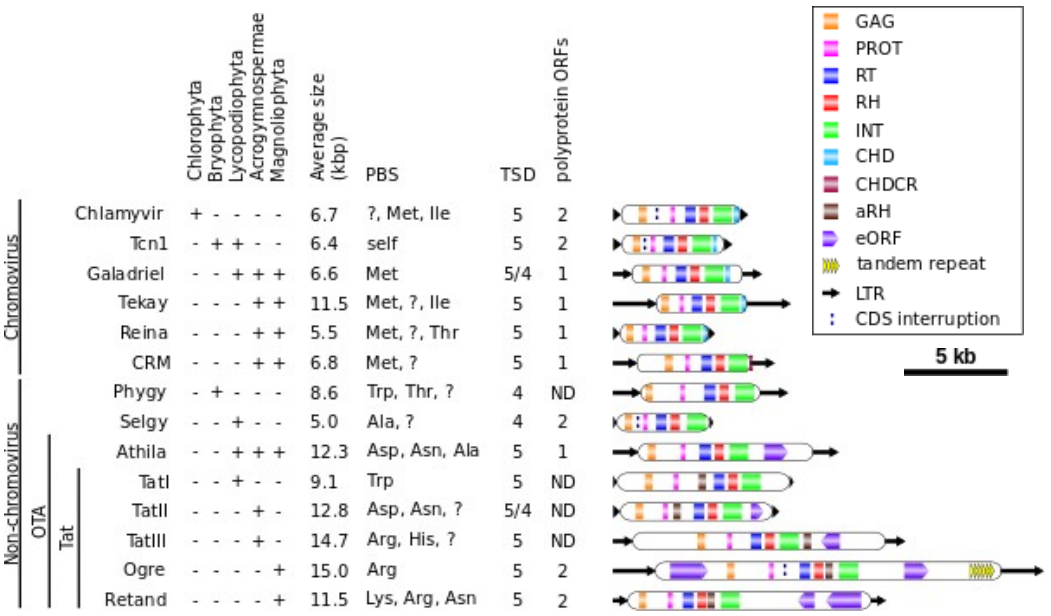
Diagnostic features of repetitive elements: part II

LTR retrotransposons: classification into lineages



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

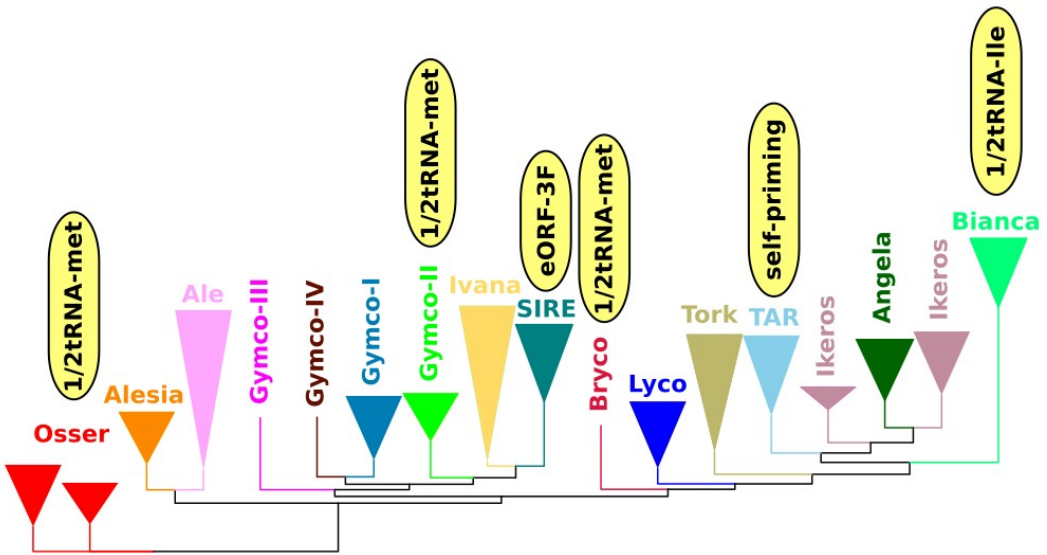
It is strongly supported by structural and sequence features (chromodomain, aRH, eORF, PBS)

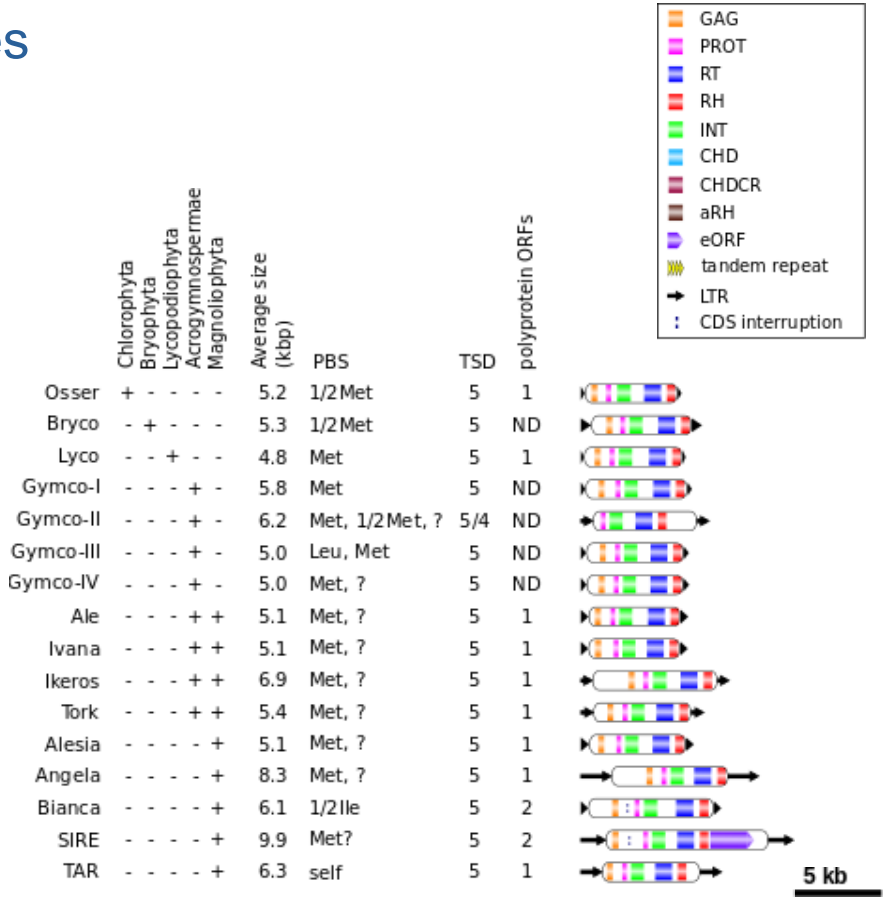


Diagnostic features of repetitive elements: part II

LTR retrotransposons: classification into lineages

A phylogenetic tree inferred from 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

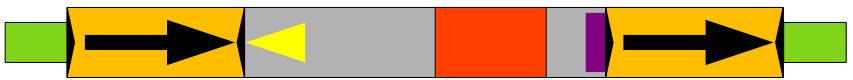
5 kb

Diagnostic features of repetitive elements

Class I: LTR retrotransposons

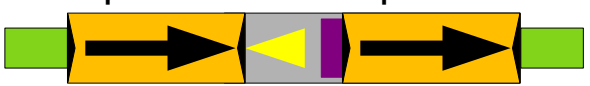
1. Autonomous element

2. Non-autonomous elements

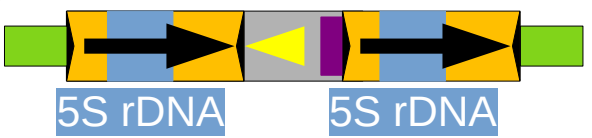


Large retrotransposon derivative (LARD)

Terminal-repeat retrotransposons in miniature (TRIM)



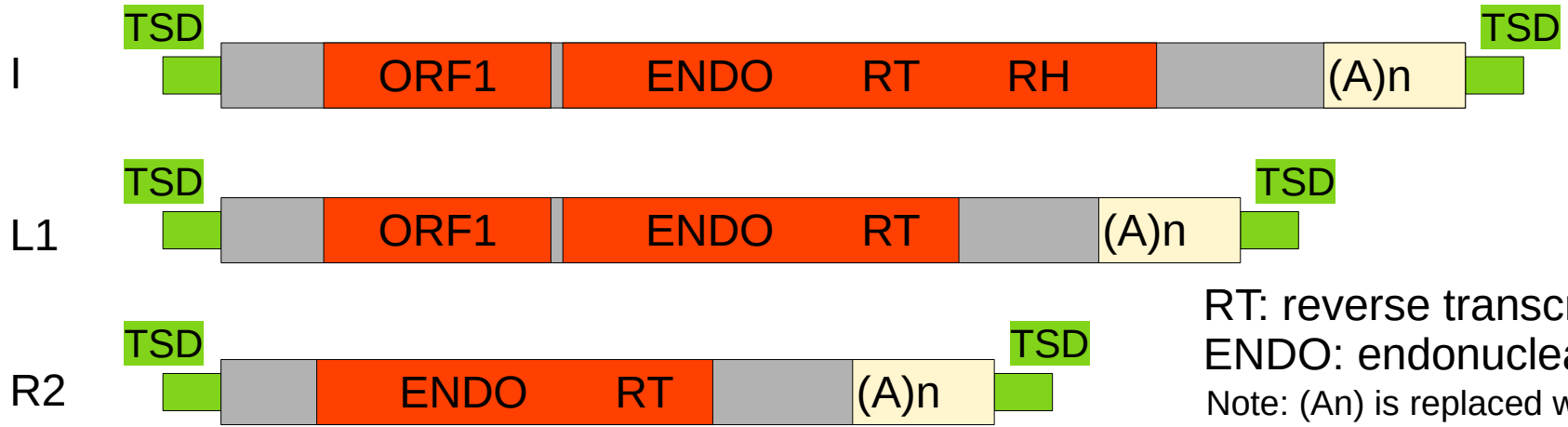
Cassandra



Diagnostic features of repetitive elements

Class I: non-LTR retrotransposons

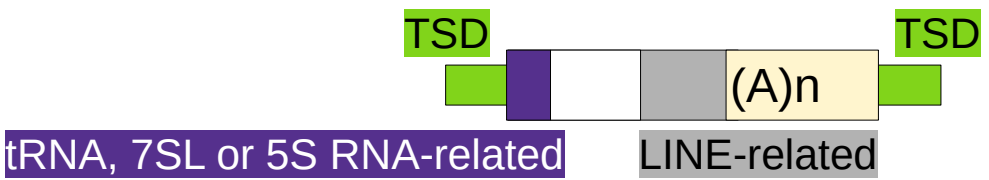
1. autonomous: long interspersed nuclear elements (LINEs)



RT: reverse transcriptase
ENDO: endonuclease

Note: (An) is replaced with simple sequence repeat (e.g.TAAn) in some elements

1. nonautonomous: short interspersed nuclear elements (SINEs)



Diagnostic features of repetitive elements

Class II: subclass I

1. Autonomous element

various types, conserved DDE/D motif

Superfamily	*Termini*	*TSD*	*Superfamily*	*Termini*	*TSD*
Mariner/Tc1	YR..YR	TA	*piggyBac*	YY..RR	TTAA
Zator	GG..CC	3	*Harbinger*	RR..YY	3
Ginger1	TGT..ACA	4	*ISL2EU*	RR..YY	2
Ginger2/TDD	TGT..ACA	4–5	*EnSpm/CACTA*	CAC..GTG	2–4
IS3EU	TAY..RTA	6	*Transib*	CAC..GTG	5
Merlin	GG..CC	8–9	Sola1	?	4
hAT	YA..TR	5–8	Sola2	GRG..CYC	4
MuDR	GR..YC	8–9	Sola3	GAG..CTC	TTAA
P	CA..TG	7–8	*Academ*	YR..YR	3–4
Kolobok	RR..YY	TTAA	*Novosib*	CA..TG	8

https://doi.org/10.1266/ggs.18-00024

Diagnostic features of repetitive elements

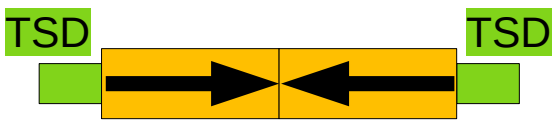
Class II: subclass I

1. Autonomous element

2. Non-autonomous elements

Miniature Inverted-repeat Transposable Elements (MITEs)

Foldback

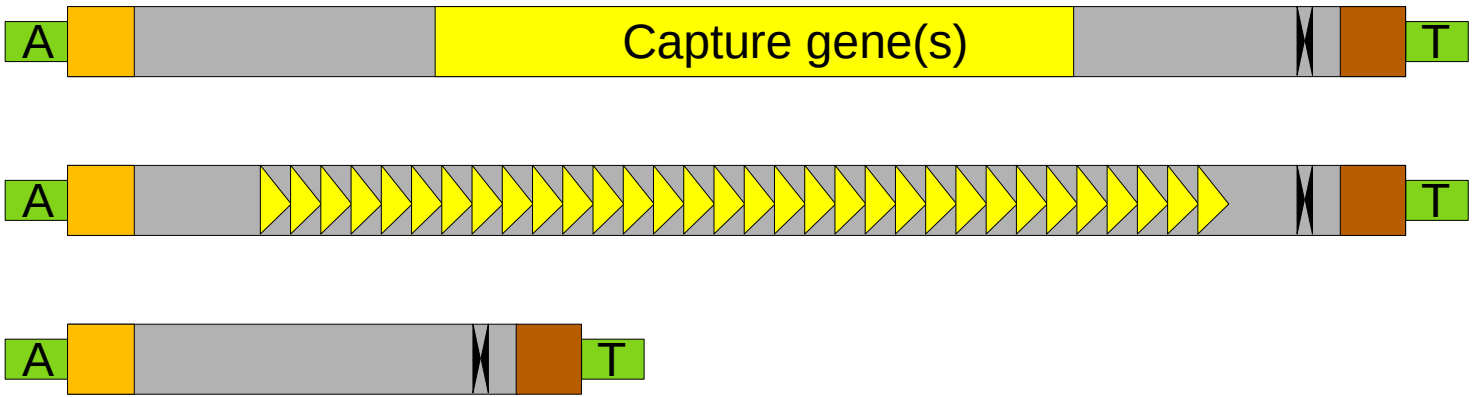


Diagnostic features of repetitive elements

Class II: subclass II - Helitron

1. Autonomous element

2. Non-autonomous elements



Diagnostic features of repetitive elements

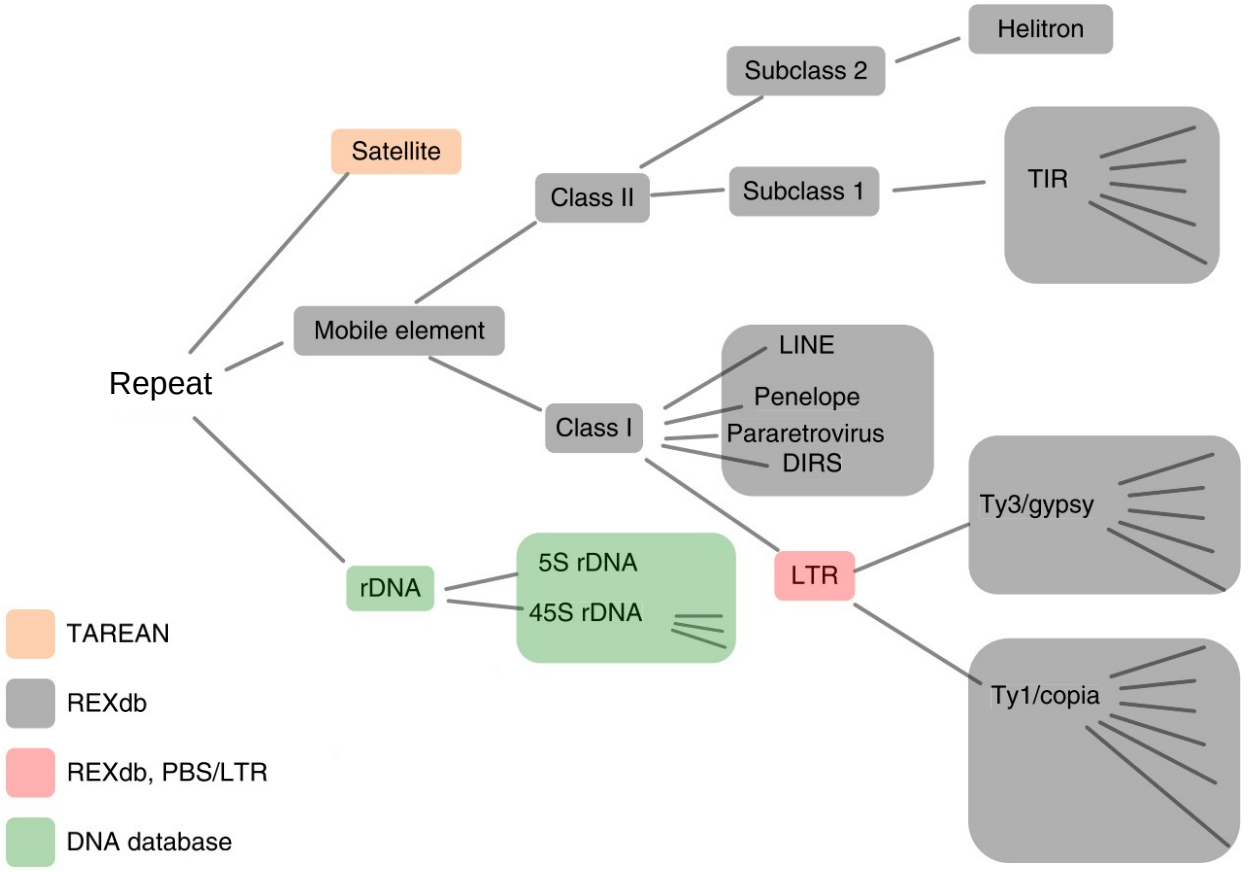
Diagnostic features of repetitive elements

Part II: Features used for automatic and manual annotation of RepeatExplorer clusters

Diagnostic features of repetitive elements: part II

Decision tree for automatic annotation.

- hierarchical
- supercluster level



Diagnostic features of repetitive elements: part II

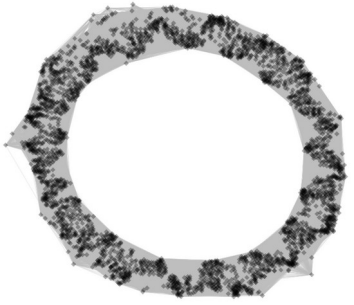
satDNA: detected by TAREAN

Cluster characteristics:

number of reads in the graph	22763
the total number of reads in the cluster	81695
number of edges of the graph:	20000401
supercluster	3
similarity based annotation	
PBS/LTR	None
pair_completeness	0.990667413923341 high
TR_score	0.529088372093023
TR_monomer_length	69
loop_index	0.991273304244347 high
satellite_probability	0.986367215261502 high
TAREAN consensus	TAAAAGTCACGAAGTTCGTTAAACTTGCAAAATTTGGTATTTTGGAAGATTTTCGTGTGCTATACTACA
TAREAN_annotation	Putative satellites (high confidence)
orientation_score	1

Cluster characteristics:

number of reads in the graph	3641
the total number of reads in the cluster	3641
number of edges of the graph:	970429
supercluster	34
similarity based annotation	
PBS/LTR	None
pair_completeness	0.992884510125889
TR_score	0.954855233995261
TR_monomer_length	602
loop_index	0.999725350178522
satellite_probability	0.978853345659724
TAREAN consensus	TTATTTTAGCACTATTTTTTAATGGAATTCCGAGATAAGACAATATTCAAAGATTAAATACAATAATGAATTTAATATGT GTTCTACAACCTTTAAAATGTTTATTTTATTTTTCATAAATTTCTAGATATATTTAAAGAAGAAATAGTTATCATAATAT TTAAAATAATCACATTGTCAAAAAATTAAATATTATGAAGTCCCCCGCACCCCTCGTAATGTTTATGGATTTCTATATTA AGATACATGTCTTAACGTGGCTGTATGTTGATAAATTAATAATTATAACATACATGTGAAATTGATAACATTTTTTGAAA ACTTGGAATACTTCCGTTAGAGAAAAGAAAATCATAGTGCAAAAACTAACCACTTTCTTTTAAAAAAAAGAAGAGAATTA AGAAGAAATGAAGAAGTTTTTTTAAACTCAATAAATTTTAAGCTATTTTAATAGTAAATGCGCTAAATTTATATTAAATG AACTTGAAAACCCCTAATGACTTTTGATCTAACTCTAGAATAAACCCACGCATTAAACTCGCTACTGTTTCTTATTTTAG CGCTACTTTTGAACTAATAGTAGGACATATAATTAATATTTA
TAREAN_annotation	Putative satellites (high confidence)
orientation_score	1



Diagnostic features of repetitive elements: part II

Transposable elements: protein domains in REXdb

Viridiplantae version 3.0

80446 protein domain sequences from a total of 17634 elements from 241 species

Metazoa version 3.0

11192 protein domain sequences from a total of 5462 elements

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

Diagnostic features of repetitive elements: part II

Fine classification of plant LTR retrotransposons into lineages

Neumann et al. Mobile DNA, 2019

Viridiplantae version 3.0

Eight domain types from 13863 LTR retrotransposons (5410 Ty1/copia and 8453 Ty3/gypsy)

GAG, PROT, RT, RH, aRH, INT, ChDII, CHDCR domains

--**Ty1_copia**
- |--Ale
- |--Alesia
- |--Angela
- |--Bianca
- |--Bryco
- |--Lyco
- |--Gymco-III
- |--Gymco-I
- |--Gymco-II
- |--Ikeros
- |--Ivana
- |--Gymco-IV
- |--Osser
- |--SIRE
- |--TAR
- |--Tork
- '--Ty1-outgroup

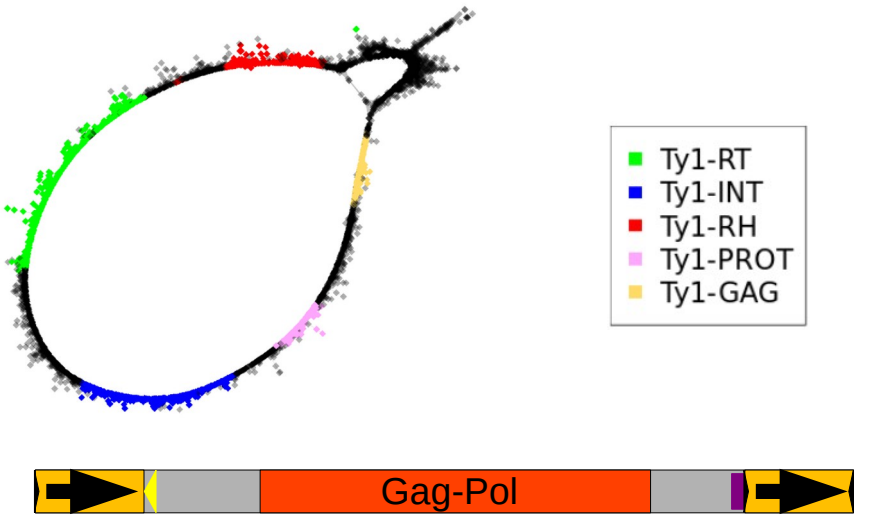
--**Ty3_gypsy**
- |--non-chromovirus
- | |--non-chromo-outgroup
- | |--Phygy
- | |--Selgy
- | '--OTA
- | |--Athila
- | '--Tat
- | |--TatI
- | |--TatII
- | |--TatIII
- | |--Ogre
- | '--Retand
- '--chromovirus
- |--Chlamyvir
- |--Tcn1
- |--chromo-outgroup
- |--CRM
- |--Galadriel
- |--Tekay
- |--Reina
- '--chromo-unclass

Diagnostic features of repetitive elements: part II

LTR retrotransposons: protein domains and LTR/PBS in clusters

Cluster characteristics:

number of reads in the graph	5947
the total number of reads in the cluster	5947
number of edges of the graph:	225728
supercluster	24
similarity based annotation	**12.73% Class_I/LTR/Ty1_copia/Tork:Ty1-RT** **10.61% Class_I/LTR/Ty1_copia/Tork:Ty1-INT** **6.05% Class_I/LTR/Ty1_copia/Tork:Ty1-RH** **4.61% Class_I/LTR/Ty1_copia/Tork:Ty1-GAG** **4.52% Class_I/LTR/Ty1_copia/Tork:Ty1-PROT** 0.44% Class_I/LTR/Ty1_copia/Ikeros:Ty1-RH 0.39% Class_I/LTR/Ty1_copia/Ale:Ty1-RT 0.32% Class_I/LTR/Ty1_copia/Ale:Ty1-INT 0.20% Class_I/LTR/Ty1_copia/Ikeros:Ty1-RT 0.15% Class_I/LTR/Ty1_copia/Ivana:Ty1-RH 0.12% Class_I/LTR/Ty1_copia/Ivana:Ty1-PROT
PBS/LTR	Phe
pair_completeness	0.817542787286064
TR_score	0.530885222566782
TR_monomer_length	4466
loop_index	0.906507482764419
satellite_probability	0.211222441079967
TAREAN consensus	ACAGGTATCAAAGGAGACGACGACATCAGGGTTATGGGTGAAACTTGAAAGTTTGTATATGACCAAATCGCTGGTAAATC GACTCTACCTGAAGCAAGCTTTGTATTCATTCAAGATGATTGAAGACAAAGTGTTGGCTGAGCAGTTGGATATGTTCAAC ..... GAAGACGAATATGATCGAGAAAGCACACAGCGCAATTTTGTTGAGCCTTGGTGATAAGGTTCTACG
TAREAN_annotation	Putative LTR elements
orientation_score	1



Smooth circular graph shapes are actually rare due to high level of sequence divergence and presence of various structural variants

Diagnostic features of repetitive elements: part II

LTR retrotransposons: protein domains and LTR/PBS in superclusters

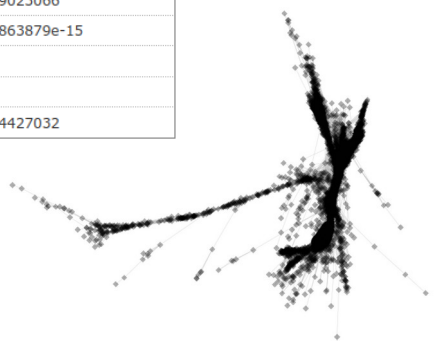
Cluster no. 8

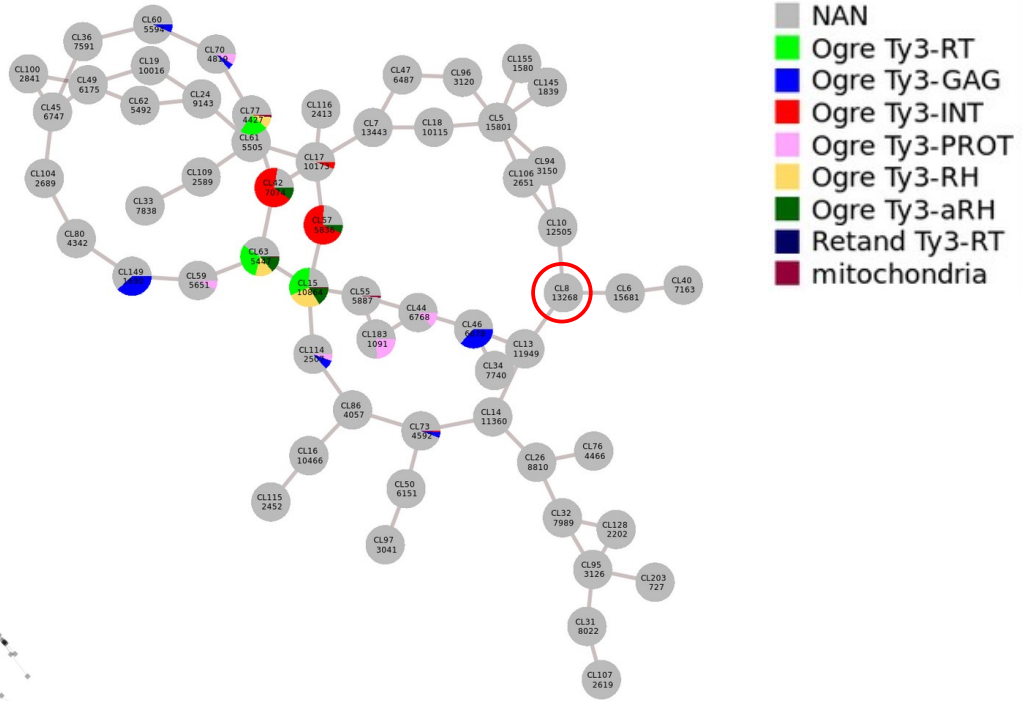
Go back to cluster table

Cluster is part of supercluster: 1

Cluster characteristics:

number of reads in the graph	13268
the total number of reads in the cluster	13268
number of edges of the graph:	3501255
supercluster	1
similarity based annotation	
PBS/LTR	Arg Arg Arg Arg Arg Arg Arg Arg Arg
pair_completeness	0.540104468949507
TR_score	None
TR_monomer_length	None
loop_index	0.118724559023066
satellite_probability	3.51557780863879e-15
TAREAN consensus	None
TAREAN_annotation	Other
orientation_score	0.999999714427032





NAN
Ogre Ty3-RT
Ogre Ty3-GAG
Ogre Ty3-INT
Ogre Ty3-PROT
Ogre Ty3-RH
Ogre Ty3-aRH
Retand Ty3-RT
mitochondria

Diagnostic features of repetitive elements: part II

RepeatExplorer: transposable elements lacking protein-coding sequences

Non-autonomous elements often remains annotated only as repeats.

In many cases the annotation can be improved by analyzing the information and sequence data provided in RepeatExplorer output.

Diagnostic features of repetitive elements: part II

RepeatExplorer: non-autonomous transposable elements

Data useful for manual annotation
/seqclust/clustering/clusters/dir_CL#

1) contigs.ace
- Cap3 assembly file of contigs that can be viewed by using an assembly viewer program, e.g. clview or tablet)

2) contigs.info.minRD5_sort-GR.fasta, contigs.info.minRD5_sort-length.fasta, contigs.info.minRD5_sort-RD.fasta
- Cap3 contigs with average read depth ≥5 sorted by different criteria (genome representation, length and read depth, respectively

3) LTR_info.with_PBS_blast.csv
- information about the detection of LTR and PBS sequences.

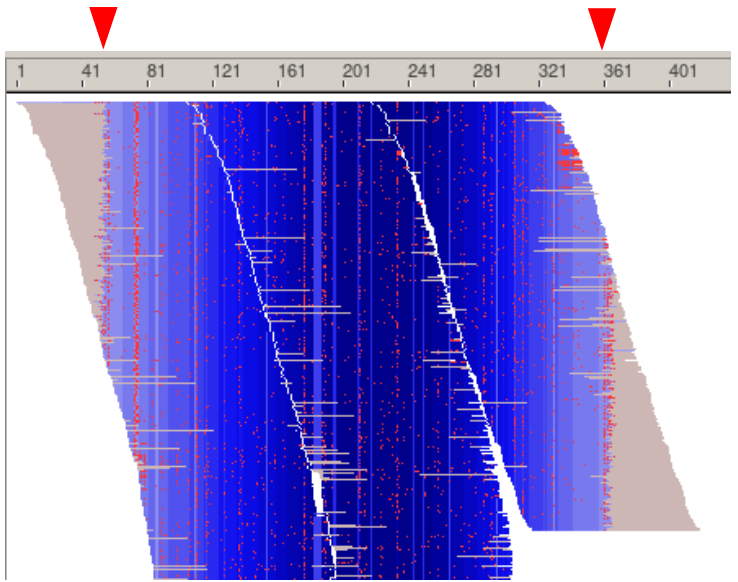
4) index.html
- RepeatExplorer html summary for the cluster (graph, pair_completeness, number of reads in the graph, clusters with similarity, clusters connected through mates etc.)

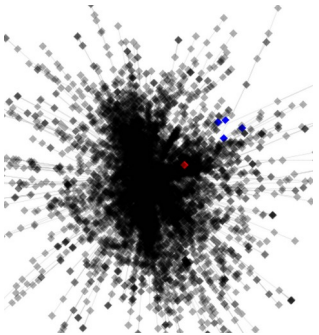
Diagnostic features of repetitive elements: part II

RepeatExplorer: unclassified clusters - MITEs

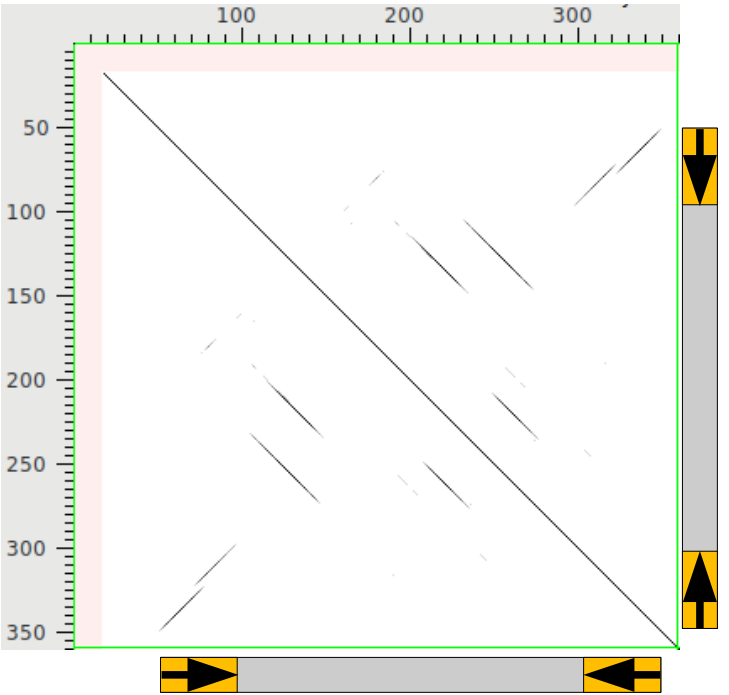
clview:
seqclust/clustering/clusters/dir_CL0007/**reads.fas.CL7.ace**

CL7Contig2365





Dot-plot of CL7Contig2365

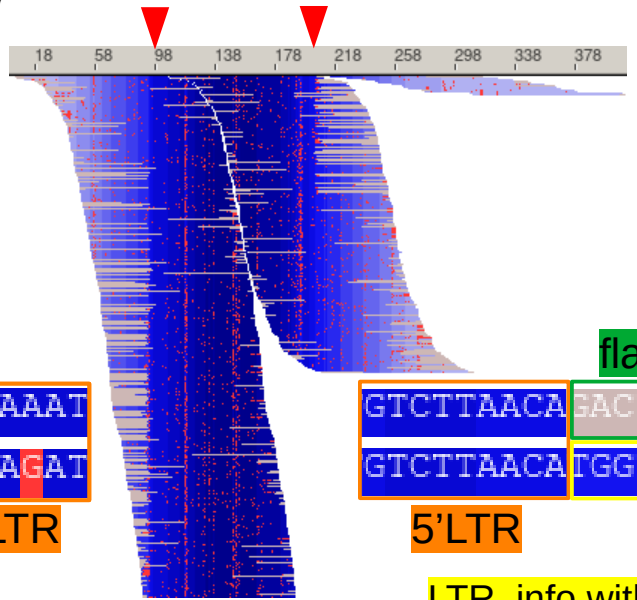


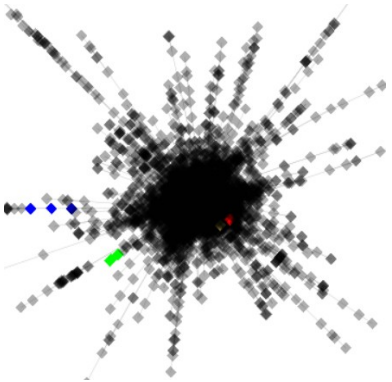
Diagnostic features of repetitive elements: part II

RepeatExplorer: unclassified clusters - TRIMs

clview:
seqclust/clustering/clusters/dir_CL0150/**reads.fas.CL150.ace**

CL150Contig567





flanking seq.

ATCGACTAC TGTTAAAT

GAGGAGGGG TGTTAGAT

PPT

3'LTR

flanking seq.

GTCTTAACA GACTAAAGATGGTTG

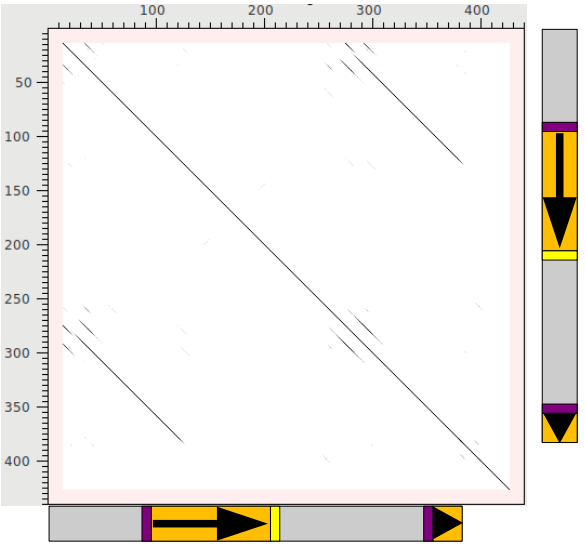
GTCTTAACA TGGTATCAGAGCCAA

5'LTR

PBS

LTR_info.with_PBS_blast.csv

Dot-plot of CL150Contig567



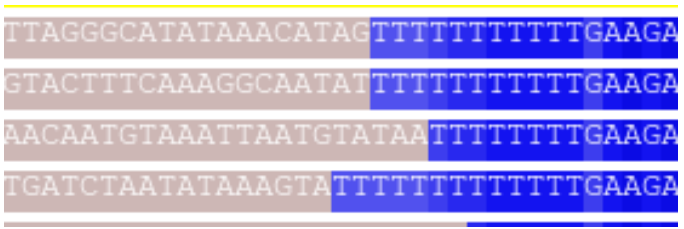
Diagnostic features of repetitive elements: part II

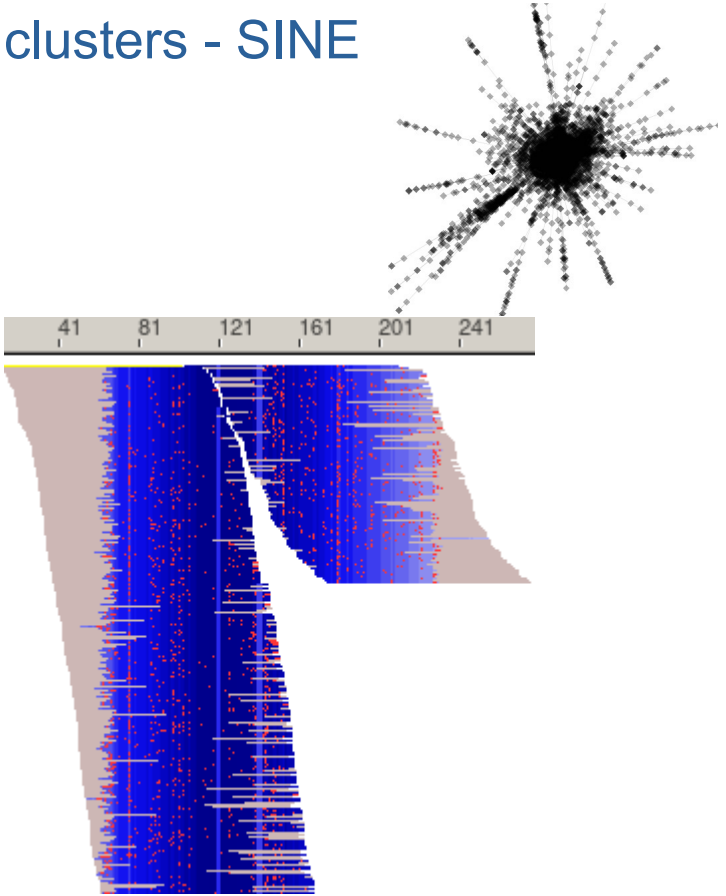
RepeatExplorer: unclassified clusters - SINE

clview:

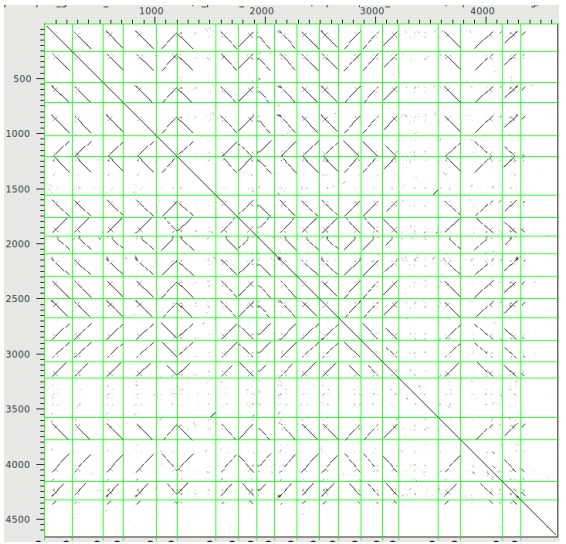
seqclust/clustering/clusters/dir_CL129

CL129Contig978





Dot-plot of top 20 contigs



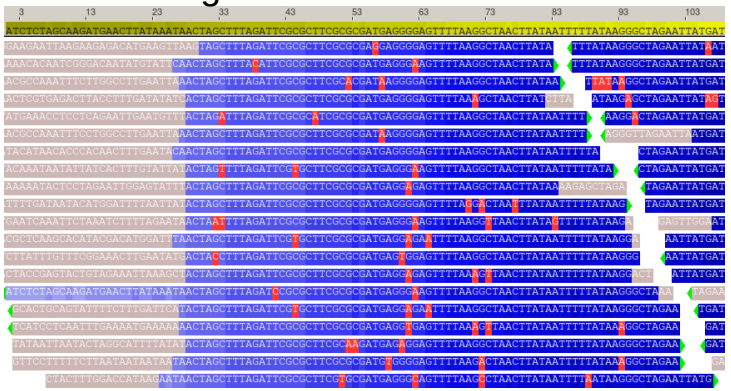
Diagnostic features of repetitive elements: part II

RepeatExplorer: unclassified clusters - putative Helitrons

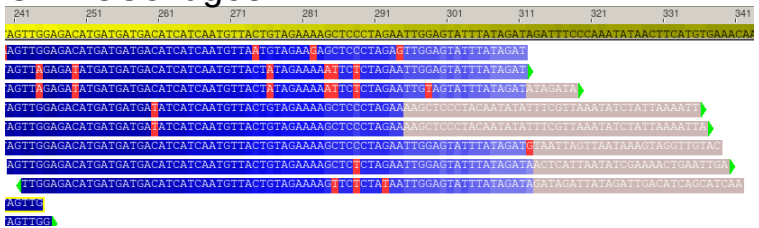
clview:

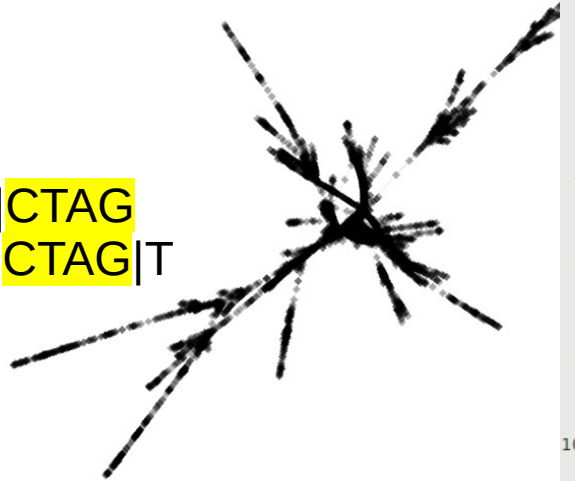
seqclust/clustering/clusters/dir_CL0145/**reads.fas.CL145.ace**

CL145Contig549

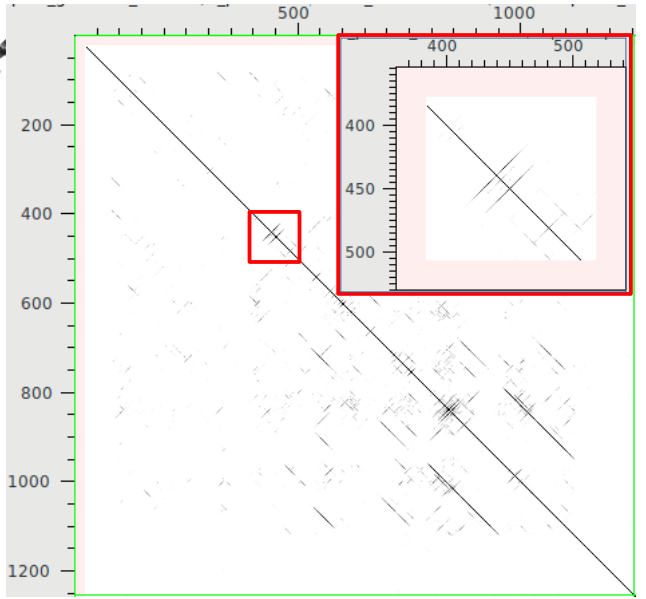


CL145Contig687





Dot-plot of CL145Contig549



Diagnostic features of repetitive elements: part II

RepeatExplorer: annotation of repeats

- Automatic annotation relies on a combination of TAREAN, similarity to sequences in REXdb and DNA sequence databases, and LTR/PBS detection. For plants, it seems highly reliable (feedback needed ...).

- Manual annotation is possible using information provided in RepeatExplorer output.

- Difficult/impossible to annotate in detail:
  - large and/or variable non-autonomous elements,
  - chimerical repeats (e.g. a transposons that captured fragment of satDNA, satDNA that evolved from a transposon)