

RepeatExplorer

**Classification of repetitive elements based on the analyzes of
protein domains**

**Pavel Neumann
May 2016**

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

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	<i>Copia</i>		4-6	RLC	P, M, F, O ●
	<i>Gypsy</i>		4-6	RLG	P, M, F, O ●
	<i>Bel-Pao</i>		4-6	RLB	M
	<i>Retrovirus</i>		4-6	RLR	M
	<i>ERV</i>		4-6	RLE	M
DIRS	<i>DIRS</i>		0	RYD	P, M, F, O ●
	<i>Ngaro</i>		0	RYN	M, F
	<i>VIPER</i>		0	RYV	O
PLE	<i>Penelope</i>		Variable	RPP	P, M, F, O ●
LINE	<i>R2</i>		Variable	RIR	M
	<i>RTE</i>		Variable	RIT	M
	<i>Jockey</i>		Variable	RIJ	M
	<i>L1</i>		Variable	RIL	P, M, F, O ●
	<i>I</i>		Variable	RII	P, M, F ●
SINE	<i>tRNA</i>		Variable	RST	P, M, F ●
	<i>7SL</i>		Variable	RSL	P, M, F ●
	<i>5S</i>		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	<i>Tc1-Mariner</i>		TA	DTT	P, M, F, O ●
	<i>hAT</i>		8	DTA	P, M, F, O ●
	<i>Mutator</i>		9-11	DTM	P, M, F, O ●
	<i>Merlin</i>		8-9	DTE	M, O
	<i>Transib</i>		5	DTR	M, F
	<i>P</i>		8	DTP	P, M
	<i>PiggyBac</i>		TTAA	DTB	M, O
	<i>PIF-Harbinger</i>		3	DTH	P, M, F, O ●
	<i>CACTA</i>		2-3	DTC	P, M, F ●
Crypton	<i>Crypton</i>		0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	<i>Helitron</i>		0	DHH	P, M, F ●
Maverick	<i>Maverick</i>		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 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

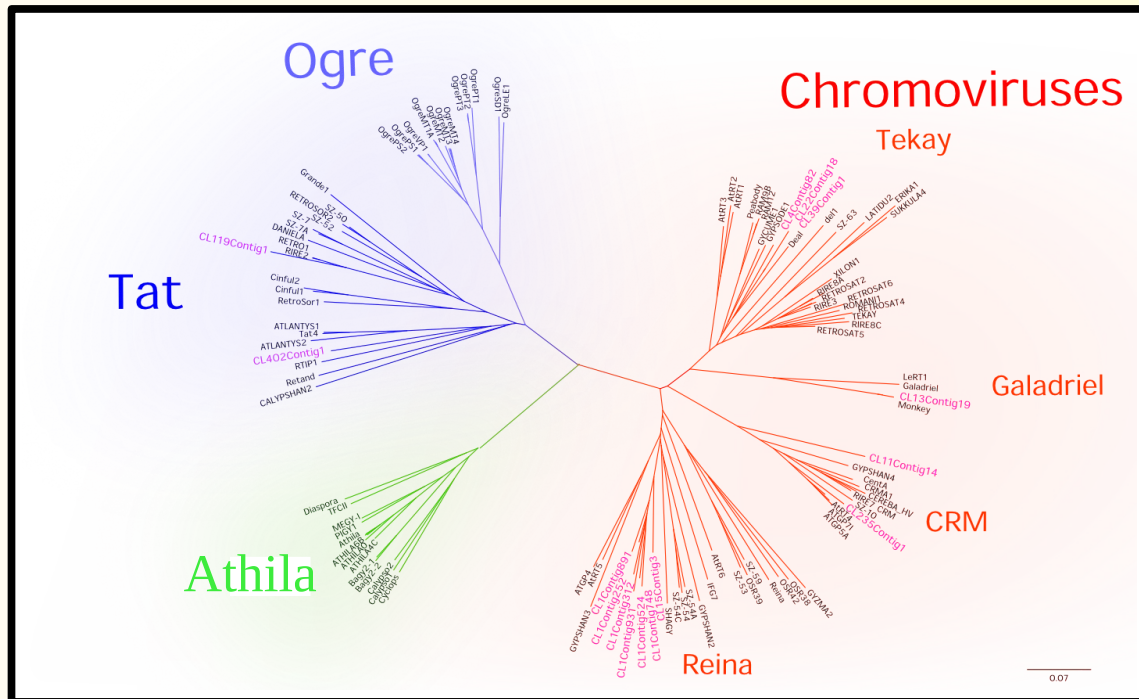
family
(e.g Peabody)

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

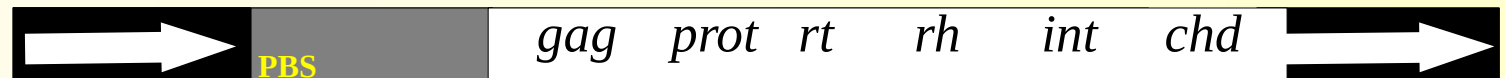
Phylogeny-based classification of plant Ty3/Gypsy retrotransposons



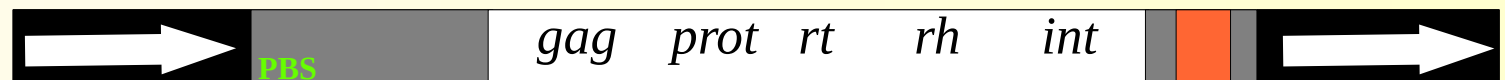
Classification according Llorens et al. (2011)

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

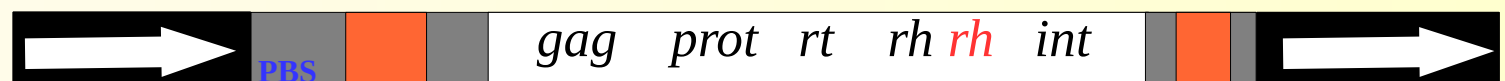
Chromovirus



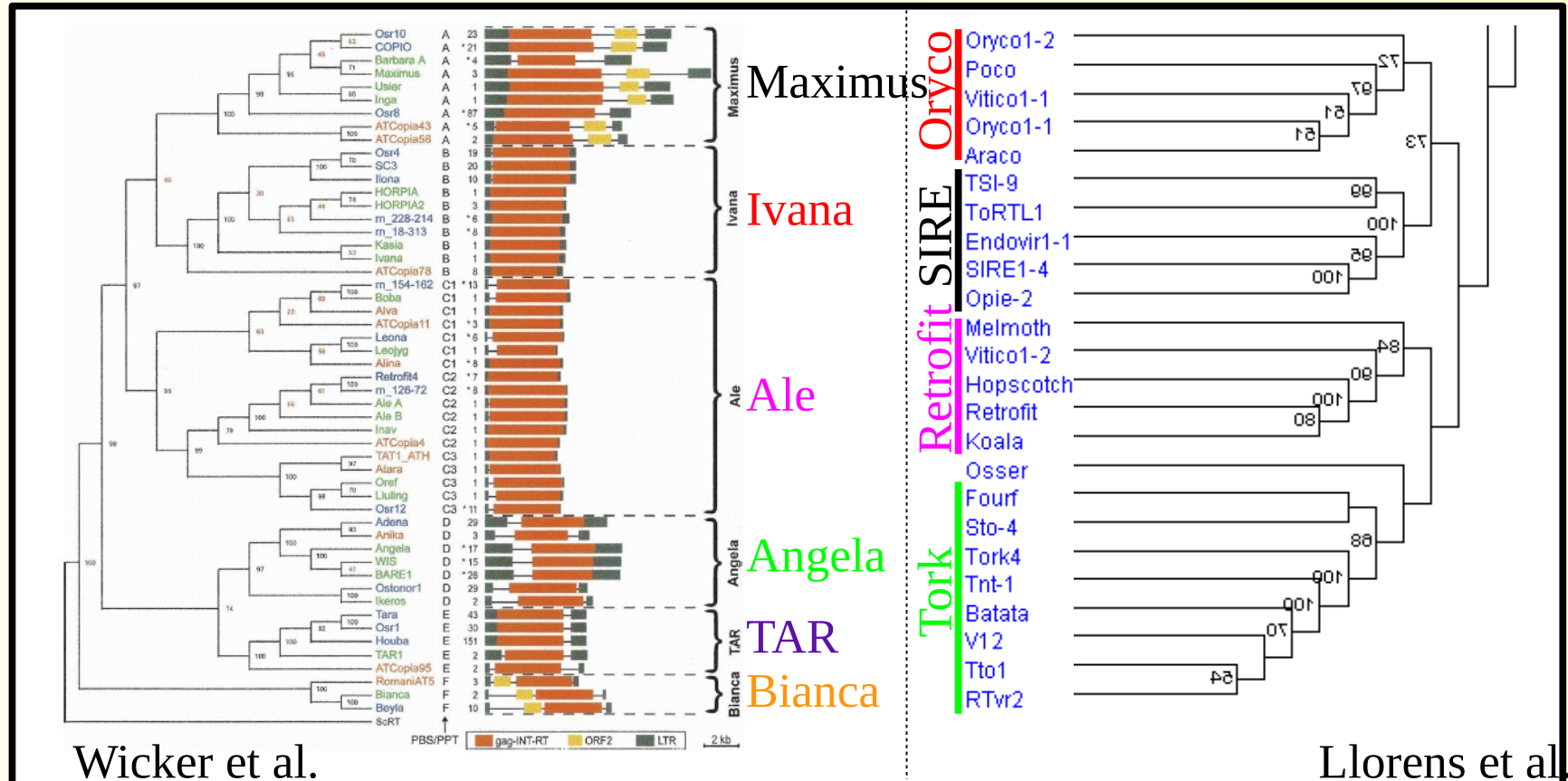
Athila



Ogre/Tat



Phylogeny-based classification of plant Ty1/Copia retrotransposons

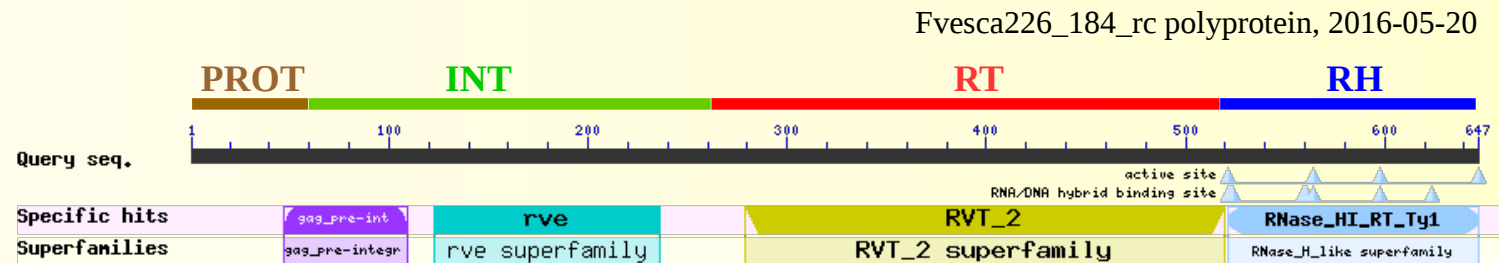


Classification according Wicker et al. (2007) and Llorens et al. (2011)

- It is based on phylogenetic analyzes of protein domain sequences (RepeatExplorer: Ivana/Oryco, Maximus/SIRE, Bianca, Tork, TAR, Angela, Ale/Retrofit)
- Structural differences among the lineages are less dramatic than in Ty3/Gypsy

Major problems of phylogeny-based classification

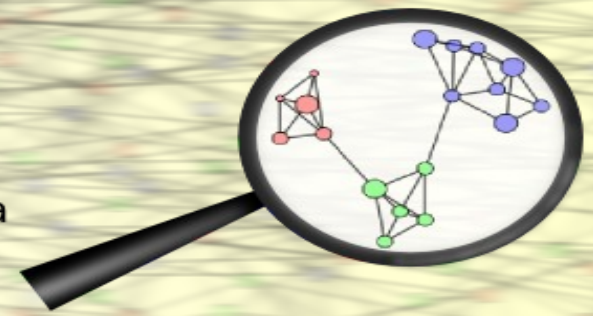
- Databases of TE protein domain sequences exist but are not sufficiently comprehensive (e.g. CDD)



- Sequences in available databases are not classified (or not well enough)

RepeatExplorer

Discover repeats in your next generation sequencing data



Database of protein domains

- Although not exhaustive, it is one of the most comprehensive databases of plant TE protein domains (**it will be updated soon** to cover LTR retrotransposons from a wide range of Viridiplanate species; from Chlorophyta over Tracheophyta and Bryophyta to Spermatophyta)
- All sequences in the database are classified into phylogenetic lineages

RepeatExplorer: database of protein domains

- **2942 domains from a total of 1045 elements**
- 398 LTR retrotransposon elements (255 **(6626 after update)** Ty1/copia and 143 **(8926)** Ty3/gypsy)
 - 386 GAG, 394 PROT, 378 RT, 383 RNaseH, 391 INT, 31 ChDII, and 45 CHDCR domains (totally 2317 **(82401)** domains)
- 15 pararetroviruses
 - 15 PROT, 15 RT, and 15 RNaseH domains (totally 45 domains)
- 228 nonLTR retrotransposon elements
 - 213 RT, 51 RNaseH, and 190 ENDO domains (totally 454 domains)
- 404 DNA transposons
 - totally 435 domains extracted from the following transposon families: DTH (Harbinger), DTM (Mutator/MuDR), DTC (EnSpm/CACTA), DTN (Novosib), DTA (hAT/AC), DTT (Mariner), DHH (Helitron)

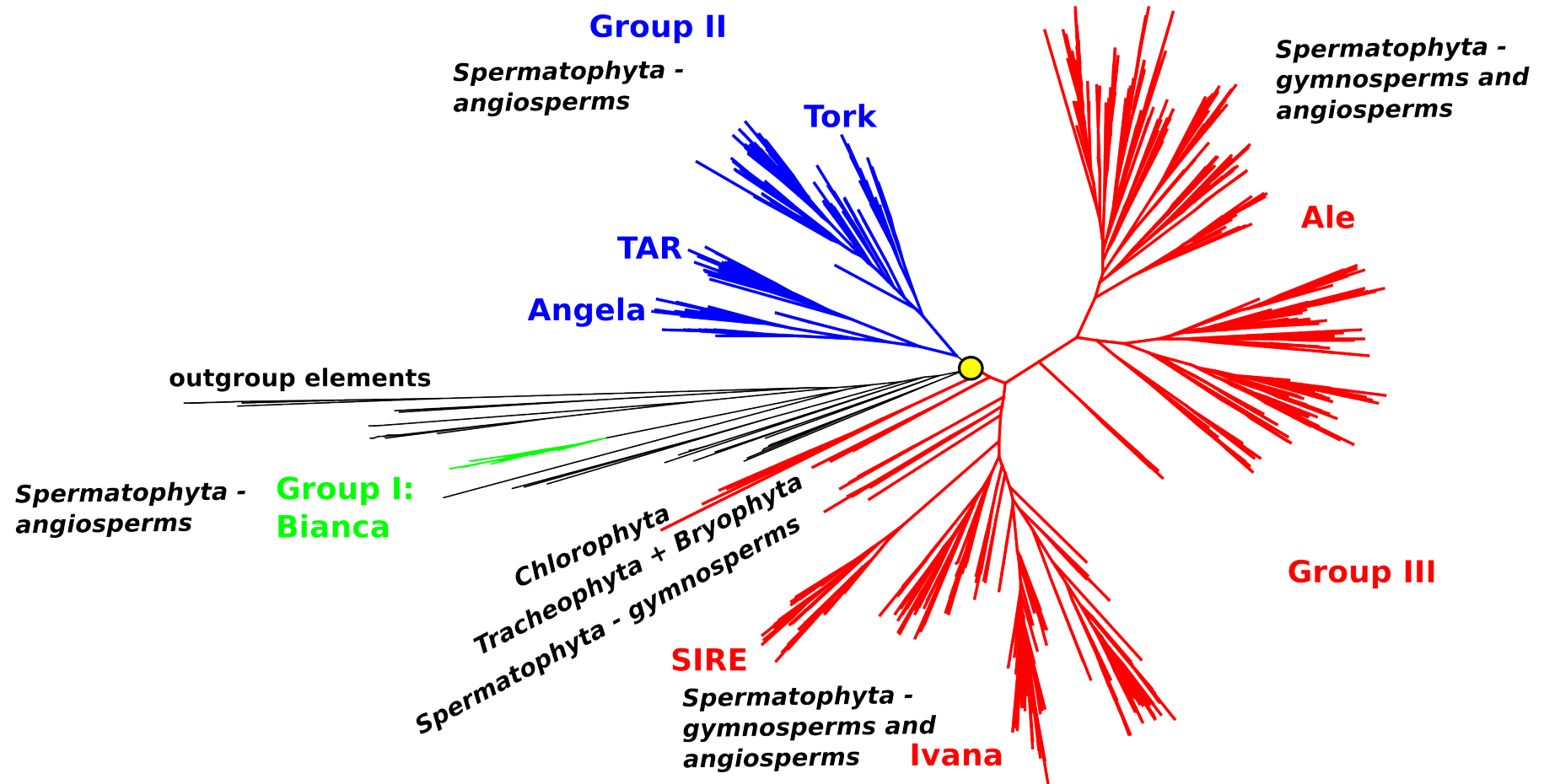
Inevitable change in classification of LTR retrotransposons

- Current classification is mostly based on sequences from angiosperm species
- If the classification is to be hierarchical, reflecting evolutionary history of LTR retrotransposons in plants, then some lineages have to merged

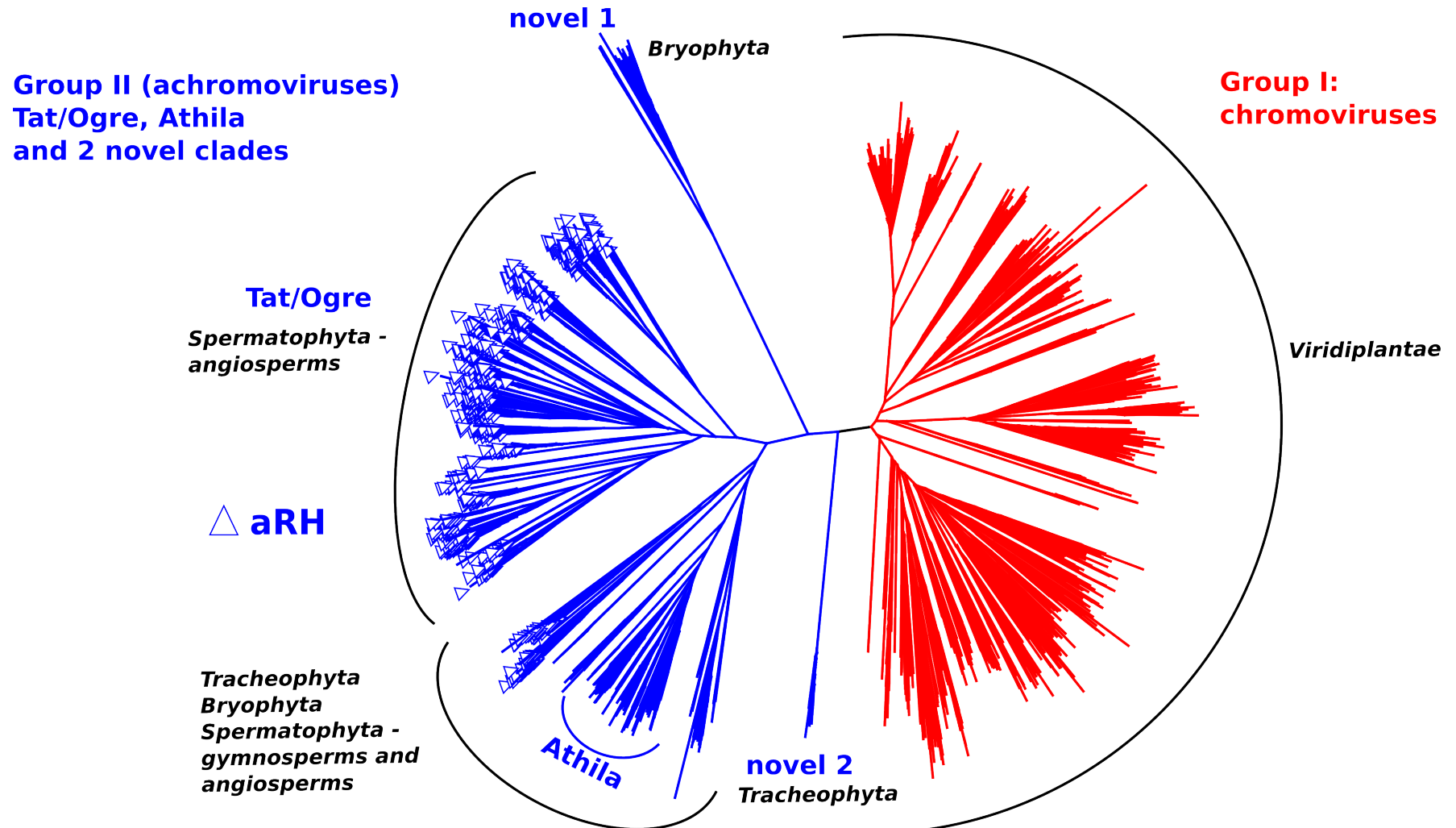
- Ty1/Copia
 - **Group I:** Bianca
 - **Group II:** Angela, TAR, Tork
 - Angela
 - Tar
 - Tork
 - **Group III:** SIRE, Ivana, Ale
 - SIRE, Ivana
 - SIRE
 - Ivana
 - ...
 - Ale
 - Ale I
 - Ale II
 - Ale III

- Ty3/Gypsy
 - **Group I:** Chromoviruses
 - Tekay
 - Reina
 - Galadriel
 - CRM
 -
 - **Group II:** Achromoviruses
 - Athila, Tat/Ogre, + 2 novel groups
 - Tat/Ogre I
 - Tat/Ogre II (+ Athila)
 - novel group I
 - novel group II

Ty1/Copia: Phylogenetic analyzes of PROT-INT-RT-RH



Ty3/Gypsy: preliminary result of phylogenetic analyzes (RT)

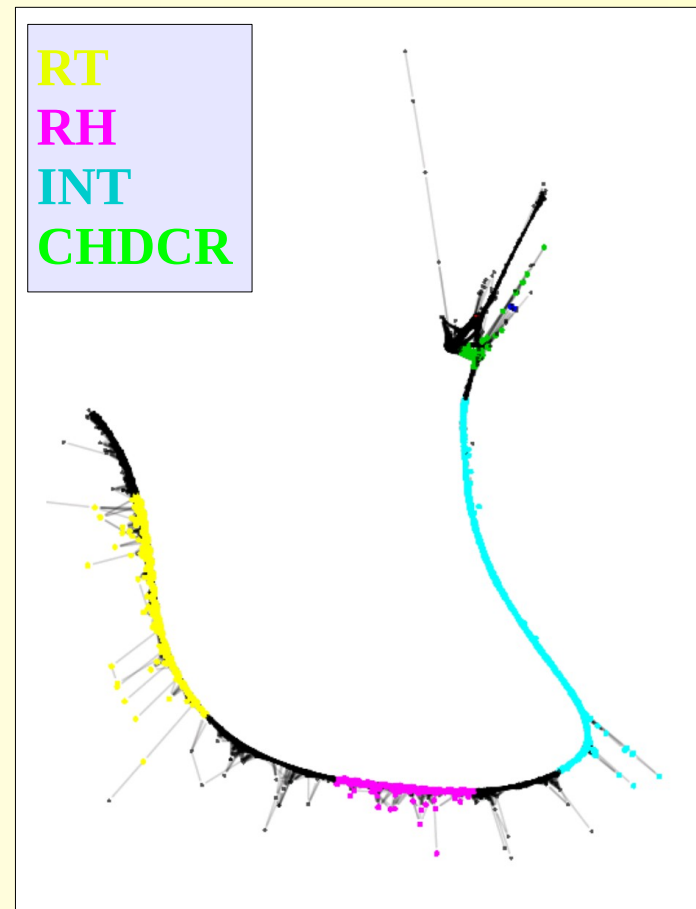


RepeatExplorer: Automatic analyzes of TE protein domains (blast using NGS reads)

HTML summary of graph based clustering

Summary of TE domain hits

	Domain	ID	Type	Lineage	Hits	MeanScore
1	Ty3-INT	PiSat1_ID200	Ty3/gypsy	chromovirus	878	62
2	Ty3-RT	PiSat1_ID200	Ty3/gypsy	chromovirus	810	64
3	Ty3-CHDCR	PiSat1_ID200	Ty3/gypsy	chromovirus	629	55
4	Ty3-RH	PiSat1_ID200	Ty3/gypsy	chromovirus	357	60
5	Ty3-INT	MedT1_ID93	Ty3/gypsy	chromovirus	318	64
6	Ty3-INT	MedT2_ID89	Ty3/gypsy	chromovirus	93	67
7	Ty3-RH	LotJ2_ID65	Ty3/gypsy	chromovirus	67	63
8	Ty3-RH	MedT2_ID89	Ty3/gypsy	chromovirus	66	71
9	Ty3-RT	MedT2_ID89	Ty3/gypsy	chromovirus	54	70
10	Ty3-RT	MedT1_ID93	Ty3/gypsy	chromovirus	43	64
11	Ty3-RH	MedT1_ID93	Ty3/gypsy	chromovirus	11	51
12	Ty3-INT	LotJ2_ID65	Ty3/gypsy	chromovirus	9	64
13	Ty3-RT	CarP1_ID79	Ty3/gypsy	chromovirus	6	43
14	Ty3-RT	BraR5_ID117	Ty3/gypsy	chromovirus	4	71
15	Ty3-INT	CRR2_ID16	Ty3/gypsy	chromovirus	4	66
16	Ty3-RT	PopT1_ID21	Ty3/gypsy	chromovirus	4	67
17	Ty3-INT	CRA3a_ID1	Ty3/gypsy	chromovirus	4	60
18	Ty3-CHDCR	MedT1_ID93	Ty3/gypsy	chromovirus	3	38
19	Ty3-INT	SorB1_ID109	Ty3/gypsy	chromovirus	3	60

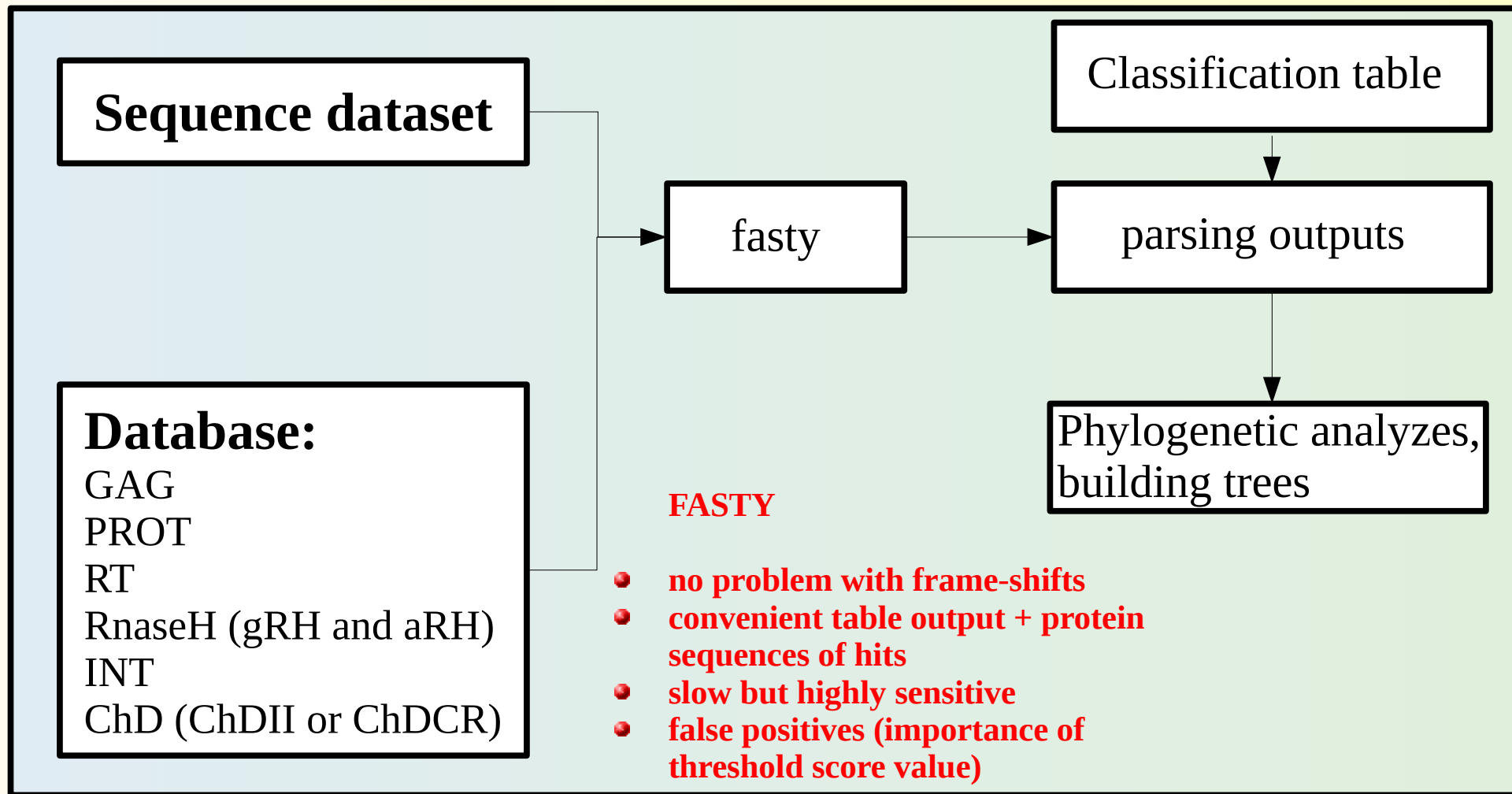


RepeatExplorer: optional analyzes of TE protein domains using “protein domains tools”

1. Protein domain search (it scans DNA sequences with protein domains database using fasty)
2. Filter output (it filters raw fasty output and extracts sequences giving hits above given threshold)
3. Create tree (it calculates and draws phylogenetic tree from a set of sequences)

RepeatExplorer: Protein domains tools

What it does:



Fasty using RT domain sequences

- RT is one of the most conserved domains of the polyprotein
- several conserved motifs are intermingled with less conserved regions and the length is highly conserved among distinct families (about 260 and 170 aa in Ty1/copia and Ty3/gypsy, respectively) - RT domain sequences are relatively easy to align
- RT domain sequences are sufficiently diverse, allowing to distinguish among distinct lineages/clades
- RT domains have already been used to infer phylogenetic relationships of retrotransposon families in plants

Parsing and filtering output of fasty: outputs

output1: table of fasty hits that passed the filter criteria

Type	lineage	query_name	query_length	hit_name	hit_length	strand	opts	bits	e_val	perc_id	perc_sim	sw	alen	an0	ax0	pn0	px0	an1	ax1	pn1	px1	gapq	gapl	fs
Ty3/gypsy	Ogre/Tat	CL1Contig12	4659	Ty3-RT_RIRE2_J	174	f	878	243.6	2.20E-065	0.678	0.874	878	174	3035	3556	1	4659	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig13	4492	Ty3-RT_RIRE2_J	174	f	894	248	1.10E-066	0.69	0.885	894	174	3060	3581	1	4492	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig23	5275	Ty3-RT_RIRE2_J	174	f	878	243.7	2.50E-065	0.678	0.874	878	174	3301	3822	1	5275	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig26	3906	Ty3-RT_RIRE2_J	174	f	897	248.8	5.30E-067	0.695	0.885	897	174	2891	3412	1	3906	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig27	663	Ty3-RT_RIRE2_J	174	r	840	233.3	4.10E-063	0.678	0.879	840	174	584	61	663	1	1	174	1	174	0	0	2
Ty3/gypsy	Ogre/Tat	CL1Contig30	3399	Ty3-RT_RIRE2_J	174	f	872	242	5.10E-065	0.672	0.868	872	174	1298	1819	1	3399	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig31	4756	Ty3-RT_RIRE2_J	174	f	887	246.1	4.20E-066	0.684	0.879	887	174	2494	3015	1	4756	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig33	2778	Ty3-RT_RIRE2_J	174	f	857	237.9	7.00E-064	0.695	0.885	857	174	2030	2551	1	2778	1	174	1	174	0	0	2
Ty3/gypsy	Ogre/Tat	CL1Contig38	2294	Ty3-RT_RIRE2_J	174	f	870	241.4	5.00E-065	0.678	0.862	870	174	1592	2113	1	2294	1	174	1	174	0	0	0
Ty1/copia	SIRE	CL27Contig4	1238	Ty1-RT_Vicker_Maximus_cons	258	f	1314	361.7	2.60E-101	0.888	0.963	1314	215	592	1236	1	1238	1	215	1	258	0	0	0
Ty1/copia	SIRE	CL27Contig15	13784	Ty1-RT_Vicker_Maximus_cons	258	r	1584	435	2.50E-122	0.884	0.969	1584	258	5359	4586	13784	1	1	258	1	258	0	0	0
Ty1/copia	SIRE	CL27Contig36	1114	Ty1-RT_Vicker_Maximus_cons	258	r	1549	425.4	1.50E-120	0.872	0.957	1549	258	967	194	1114	1	1	258	1	258	0	0	0
Ty1/copia	Angela	CL55Contig6	3043	Ty1-RT_Vicker_BARE1_B_cons	258	f	1586	435.5	3.90E-123	0.933	0.992	1586	255	2	766	1	3043	4	258	1	258	0	0	0
Ty1/copia	TAR	CL89Contig4	5745	Ty1-RT_Vicker_Osr1_Os_cons	257	r	1582	434.4	1.50E-122	0.891	0.973	1582	257	4461	3691	5745	1	1	257	1	257	0	0	0

```
>CL27Contig4_592-1236
WIQAMQEELQQFELNNVWELVKRPDPKRKHNIIGTKWIYRNKQDEHGQVVR
NKARLVAQGYTQVEGIDFDETFAPVARLEAIRILLAYANHHNIFLYQMDV
KSAFLNGKIEEEVYVAQPPGFEDPKRPDMVYKLNKALYGLKQAPRAWYDT
LKDFLKSFGKPGSLDPTLFTKTYDGELFVCQIYVDDIIFGCTVKRYSYE
FGYMMQVQYQMSMMG
>CL27Contig15_5359-4586
WIQAMQEELQQFELNNVWELVKRPDPKRKHNIIGTKWIYRNKQDEHGQVVR
NKARLVAQGYTQVEGIDFDETFAPVARLEAIRILLAYANHHNILLYQMDV
KSAFLNGRIEEVYVAQPPGFEDPKRPDMVYKLNKALYGLKQAPRAWYDT
LKDFLKSFGKPGSLDPTLFTKTYDDELFCQIYVDDIIFGCTDKRYSDE
FGHMMQEYQYQMSMMGELKFLLGLQIRQQSNGIFISQEKYLDCLKKFGMQ
DCKGYSTP
>CL27Contig36_967-194
WIQAMHEELQQFELNNVWELVKRPDPKRKHNIIGTKWIYRNKQDENGQVVI
NKARLVAQGYTQVEGIDFDETFALVARLEAIRILLAYANHHNILLYQMDV
KSAFLNGKIEEEVYVAQPPGFEDPKRPDMVYKLNKALYGLKQAPRAWYDT
LKDFLKSFGKPGSLDPTLFTKTYDDELFCQIYVDDIIFGCTDKRYSDE
FGHMMQEYQYQMSMMGELKFLLGLQIRQQSNDIFISQEKYLDCLKKFGMQ
DCKGYTTP
```

output2: Ty1 domains

```
>CL1Contig12_3035-3556
VRGVIHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCELLSFLDAYSQYHQIFMAKEDEEKTAFITPCGTTCYCFIRMPFGLK
NAGSTFARVVYTAFAEPQIHRNVEAYMDDIVVKSksKETLIQDLEETFQNL
RRIQLKLNPEKCVFVPSGKLLGF
>CL1Contig13_3060-3581
VRGVIHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCELLSFLDAYSQYHQIFMAEEDEEKTAFITPCGTTCYCFIRMPFGLK
NAGSTFARVVHTAFAEPQIHRNVEAYMDDIVVKSksKDTLIQDLDETANL
RKIQLKLNPEKCVFVPSGKLLGF
>CL1Contig23_3301-3822
VRGVLHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCDLLSFLDAYSQYHQIFMAEEDEEKTAFITPCGTTCYCFIRMPFGLK
NAGSTFARVVHTAFAEPQIHRNVEAYMDDIVVKSksKDRATLIQDLDETANL
RKINLKLNPEKCVFVPSGKLLGF
>CL1Contig26_2891-3412
VRGVLHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCELLSFLDAYSQYHQIFMAEEDEEKTAFITPCGTTCYCFIRMPFGLK
NAGSTFARVVHTAFAEPQIHRNVEAYMDDIVVKSksKDTLIQDLDETANL
RKIQLKLNPEKCVFVPSGKLLGF
```

output3: Ty3 domains

Classification of LTR retrotransposons: table or phylogenetic tree?

output1: table of fasty hits that passed the filter criteria

Type	lineage	query_name	query_length	hit_name	hit_length	strand	opts	bits	e_val	perc_id	perc_sim	sw	alen	an0	ax0	pn0	px0	an1	ax1	pn1	px1	gapq	gapl	fs
Ty3/gypsy	Ogre/Tat	CL1Contig12	4659	Ty3-RT_RIRE2_J	174	f	878	243.6	2.20E-065	0.678	0.874	878	174	3035	3556	1	4659	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig13	4492	Ty3-RT_RIRE2_J	174	f	894	248	1.10E-066	0.69	0.885	894	174	3060	3581	1	4492	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig23	5275	Ty3-RT_RIRE2_J	174	f	878	243.7	2.50E-065	0.678	0.874	878	174	3301	3822	1	5275	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig26	3906	Ty3-RT_RIRE2_J	174	f	897	248.8	5.30E-067	0.695	0.885	897	174	2891	3412	1	3906	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig27	663	Ty3-RT_RIRE2_J	174	r	840	233.3	4.10E-063	0.678	0.879	840	174	584	61	663	1	1	174	1	174	0	0	2
Ty3/gypsy	Ogre/Tat	CL1Contig30	3399	Ty3-RT_RIRE2_J	174	f	872	242	5.10E-065	0.672	0.868	872	174	1298	1819	1	3399	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig31	4756	Ty3-RT_RIRE2_J	174	f	887	246.1	4.20E-066	0.684	0.879	887	174	2494	3015	1	4756	1	174	1	174	0	0	0
Ty3/gypsy	Ogre/Tat	CL1Contig33	2778	Ty3-RT_RIRE2_J	174	f	857	237.9	7.00E-064	0.695	0.885	857	174	2030	2551	1	2778	1	174	1	174	0	0	2
Ty3/gypsy	Ogre/Tat	CL1Contig38	2294	Ty3-RT_RIRE2_J	174	f	870	241.4	5.00E-065	0.678	0.862	870	174	1592	2113	1	2294	1	174	1	174	0	0	0
Ty1/copia	SIRE	CL27Contig4	1238	Ty1-RT_Wicker_Maximus_cons	258	f	1314	361.7	2.60E-101	0.888	0.963	1314	215	592	1236	1	1238	1	215	1	258	0	0	0
Ty1/copia	SIRE	CL27Contig15	13784	Ty1-RT_Wicker_Maximus_cons	258	r	1584	435	2.50E-122	0.884	0.969	1584	258	5359	4586	13784	1	1	258	1	258	0	0	0
Ty1/copia	SIRE	CL27Contig36	1114	Ty1-RT_Wicker_Maximus_cons	258	r	1549	425.4	1.50E-120	0.872	0.957	1549	258	967	194	1114	1	1	258	1	258	0	0	0
Ty1/copia	Angela	CL55Contig6	3043	Ty1-RT_Wicker_BARE1_B_cons	258	f	1586	435.5	3.90E-123	0.933	0.992	1586	255	2	766	1	3043	4	258	1	258	0	0	0
Ty1/copia	TAR	CL89Contig4	5745	Ty1-RT_Wicker_Osr1_Os_cons	257	r	1582	434.4	1.50E-122	0.891	0.973	1582	257	4461	3691	5745	1	1	257	1	257	0	0	0



- Classification shown in the table is reliable only in case of **high similarity** to the database sequences
- Sequences that have low similarity to those in the database should be classified using phylogenetic analysis

A few tips at the end

- bad alignment leads to a wrong tree
- wrong translation leads to a wrong result
 - continuous translations from multiple frames is convenient but not always correct; consider avoiding old severely mutated sequences - typically but not necessarily those in clusters composed of only a few reads
- analyze Ty3/gypsy and Ty1/copia separately (they are too divergent and difficult to align)
- analyze more domains to get the highest confidence
 - RT, RNaseH, and INT are relatively easy to compare
 - GAG, PROT, CHDII/CHDCR are highly divergent and often difficult to compare but can be used for finer classification of LTR retrotransposons
- if you are not sure take a look at other features (pbs, introns, extra ORF)
- it should be **you** who makes the final decision; do not blindly rely on the automatic outputs