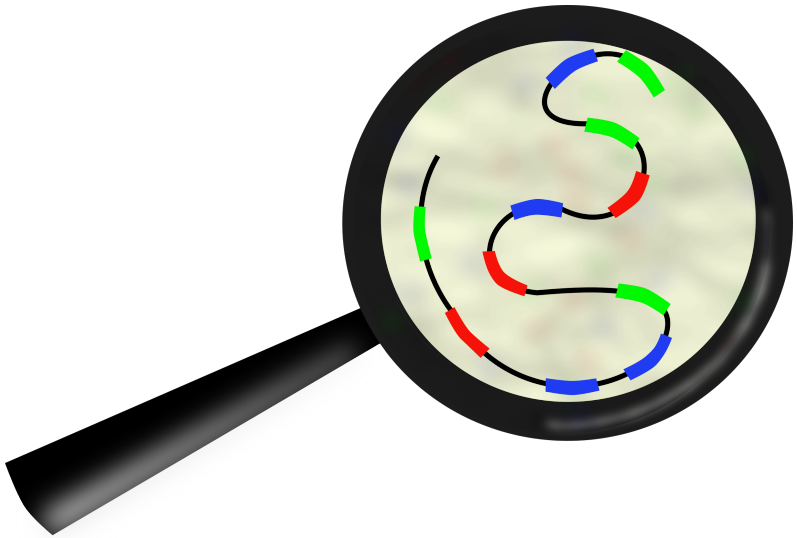
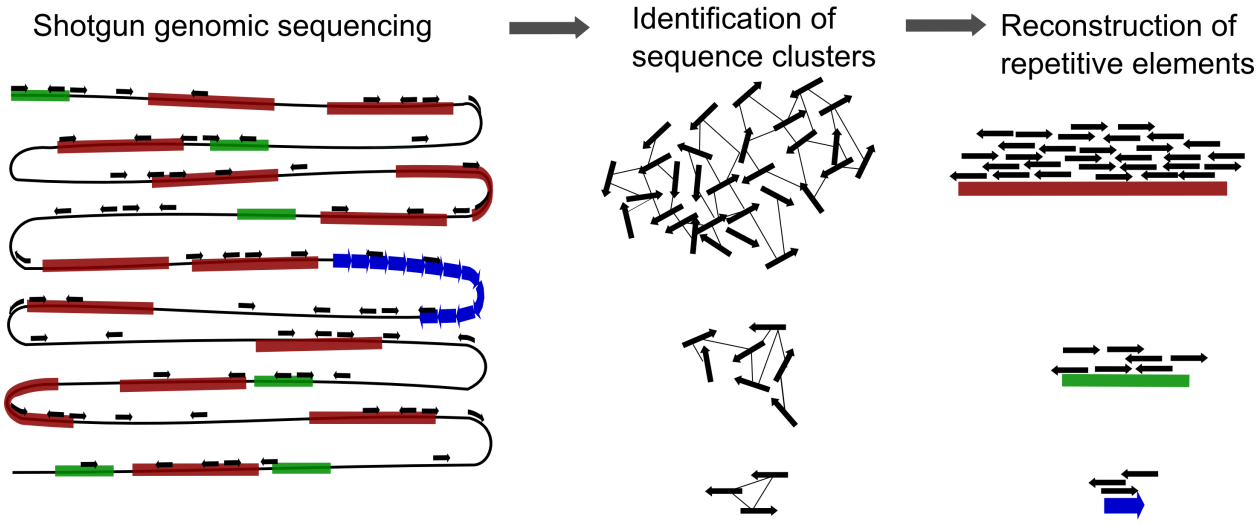
RepeatExplorer tools for genome annotation

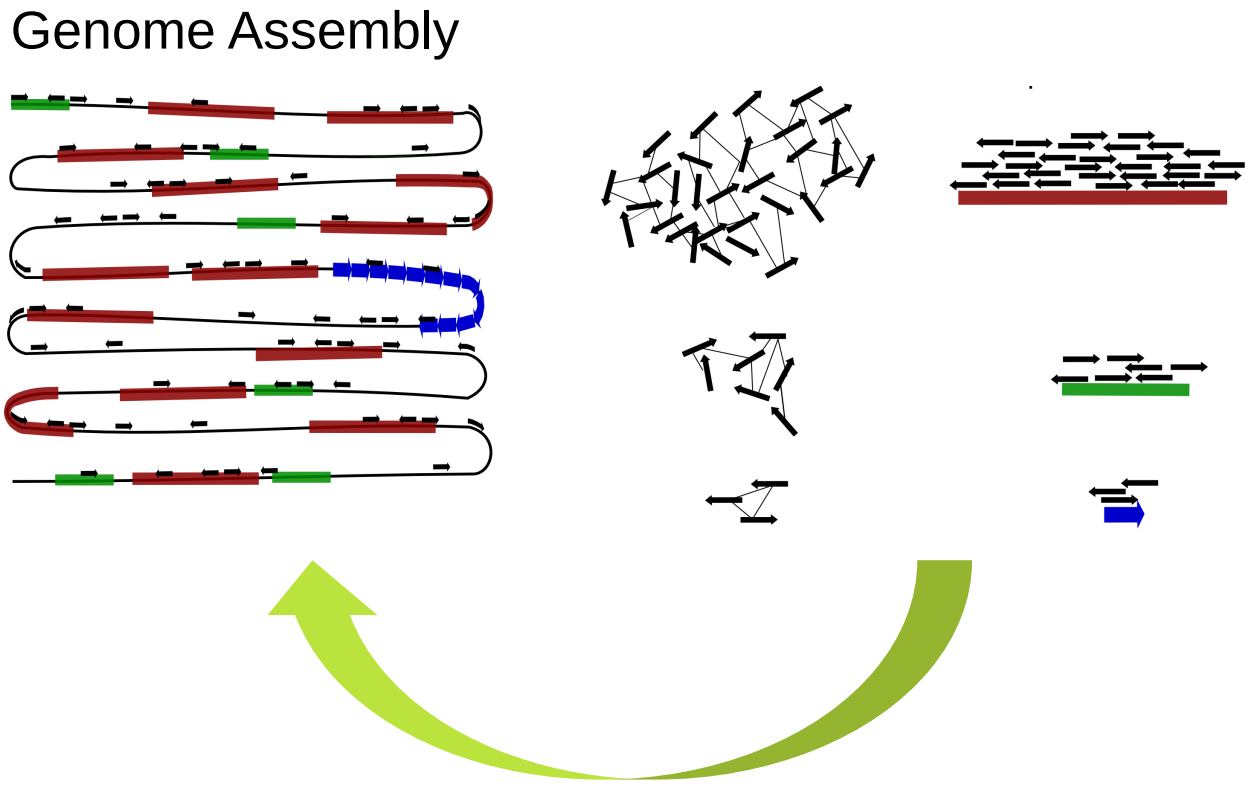


RepeatExplorer tools for genome annotation



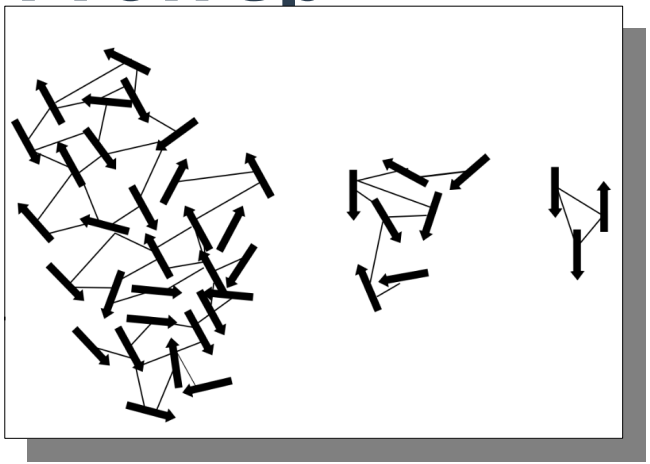
Characterization of repeat from low-pass shogun sequencing

RepeatExplorer tools for genome annotation



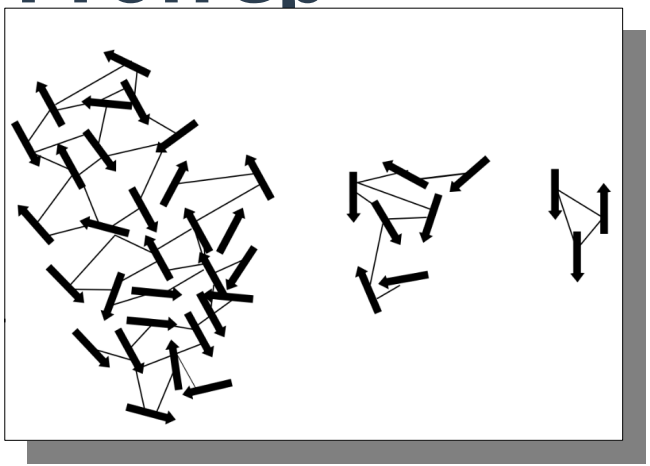
RepeatExplorer server tools for genome annotation

Profrep



RepeatExplorer server tools for genome annotation

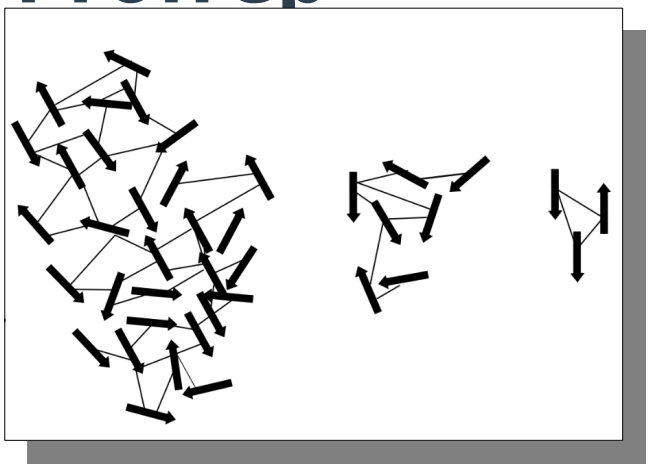
Profrep



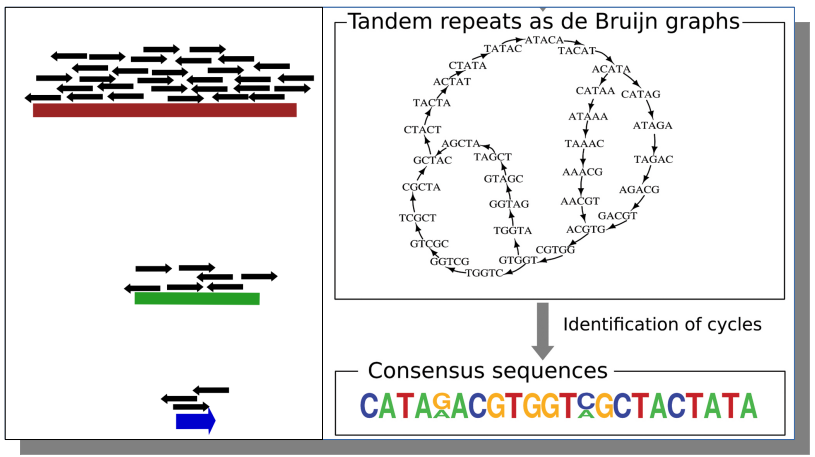
DANTE

RepeatExplorer server tools for genome annotation

Profrep



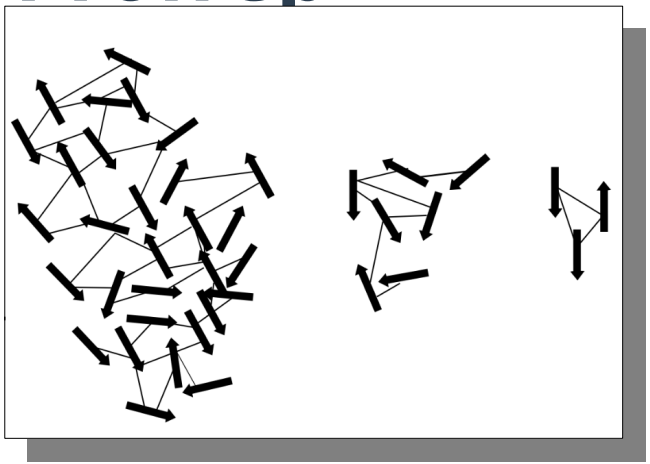
DANTE



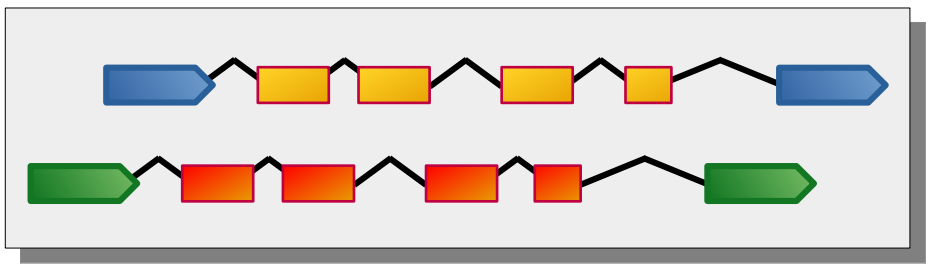
Library based annotation

RepeatExplorer server tools for genome annotation

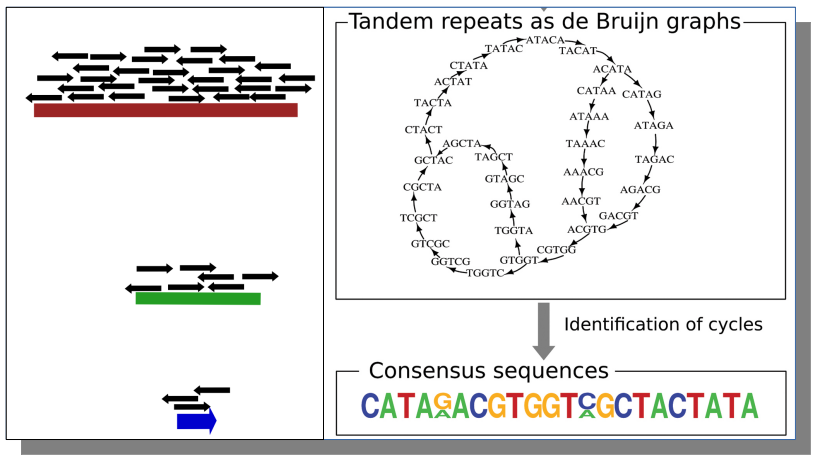
Profrep



DANTE

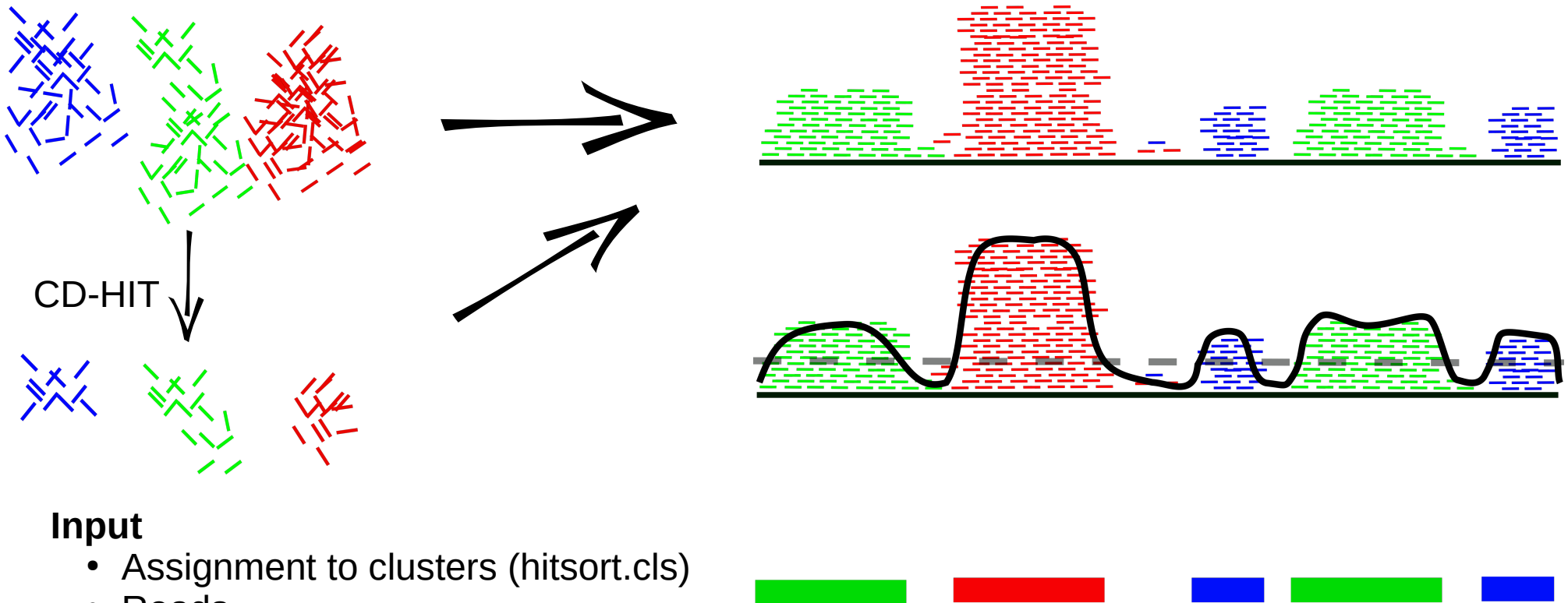


DANTE LTR



Library based annotation

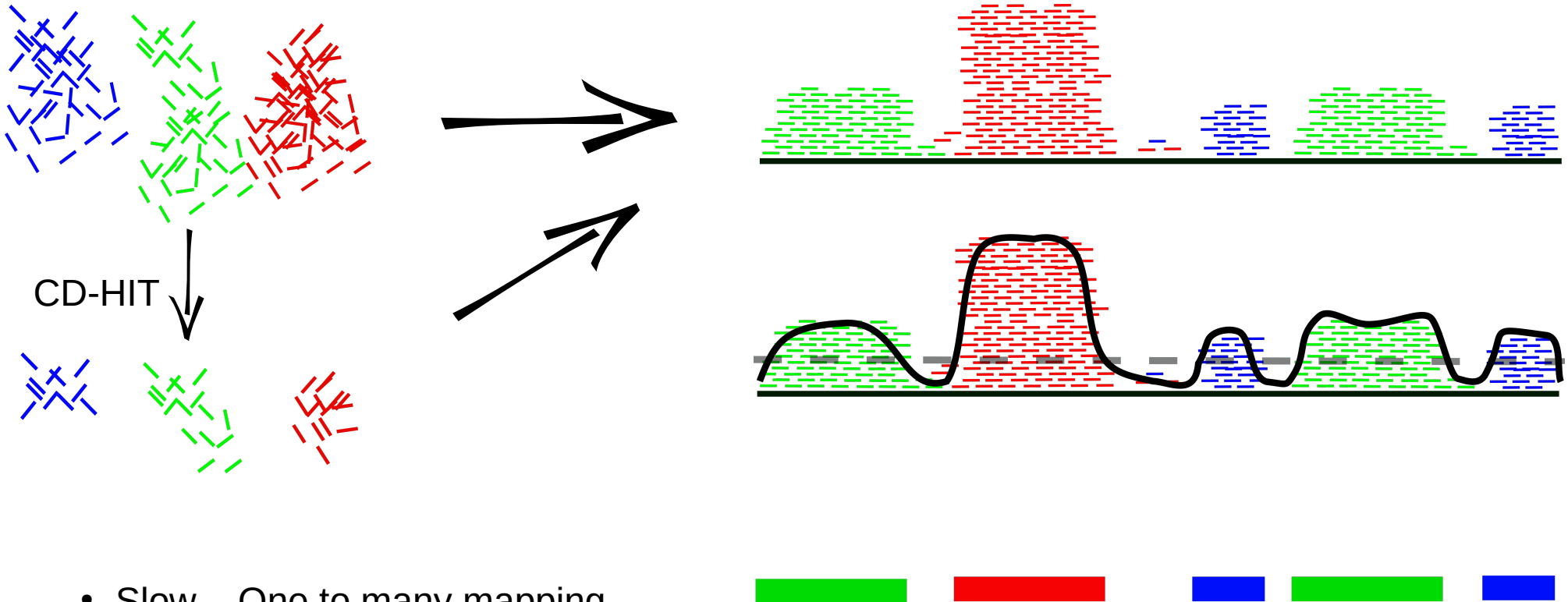
PROFREP



Input

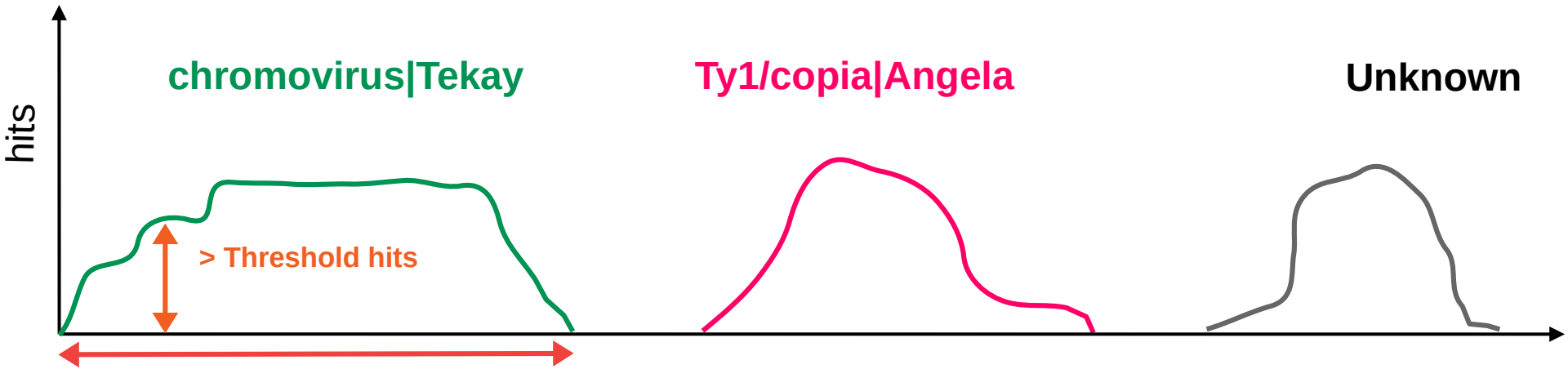
- Assignment to clusters (hitsort.cls)
- Reads
- Cluster annotation – curated
- Genome size (optional)

PROFREP



- Slow – One to many mapping
- Copy Number information

PROFREP



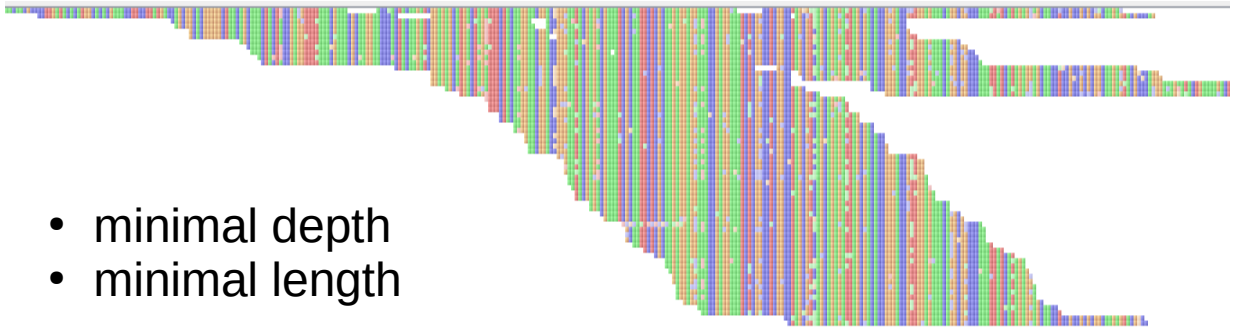

```
##gff-version 3
9 profrep repeat 53   1221 . . . Name=repeat|mobile_element|Class_I|LTR|Ty1/copia|SIRE;Average_PID=83
9 profrep repeat 7819 7921 . . . Name=repeat|mobile_element|Class_I|LTR|Ty3/gypsy|chromovirus|Tekay;Average_PID=77
9 profrep repeat 8365 8449 . . . Name=unknown_CL203;Average_PID=75
9 profrep repeat 9353 9434 . . . Name=unknown_CL209;Average_PID=83
```

Library Based Repeat Annotation

- Repbase
- DFAM
- **Custom library**

Library Based Repeat Annotation

Contigs from clustering results



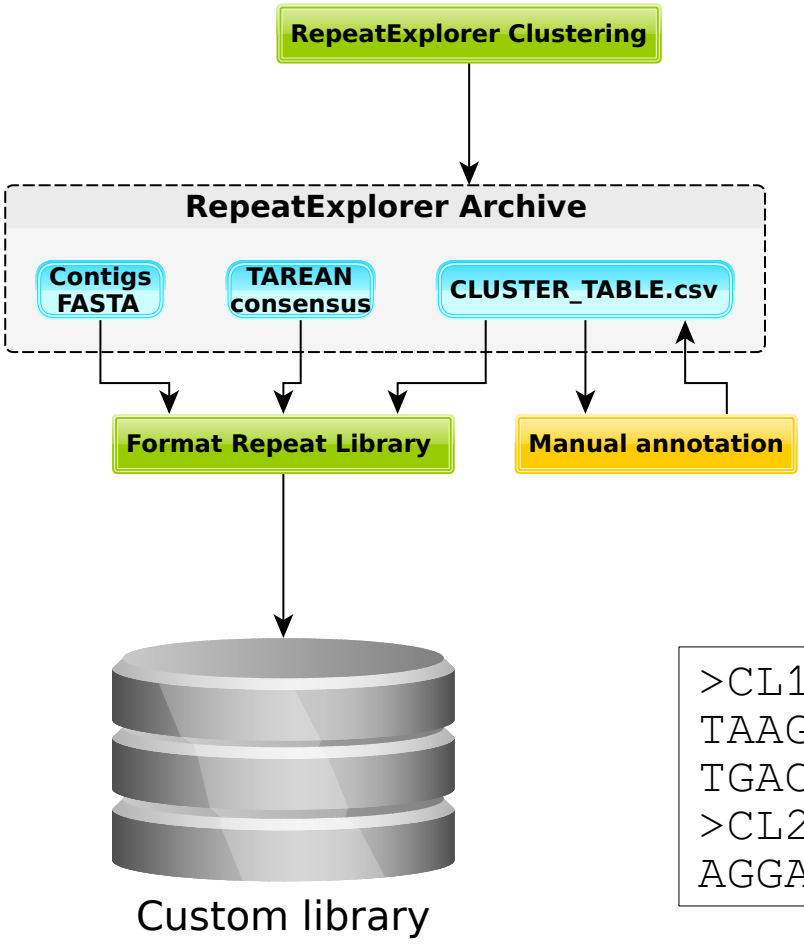
- minimal depth
- minimal length

- **Custom library**

TAREAN consensus

- as dimers

Library Based Repeat Annotation



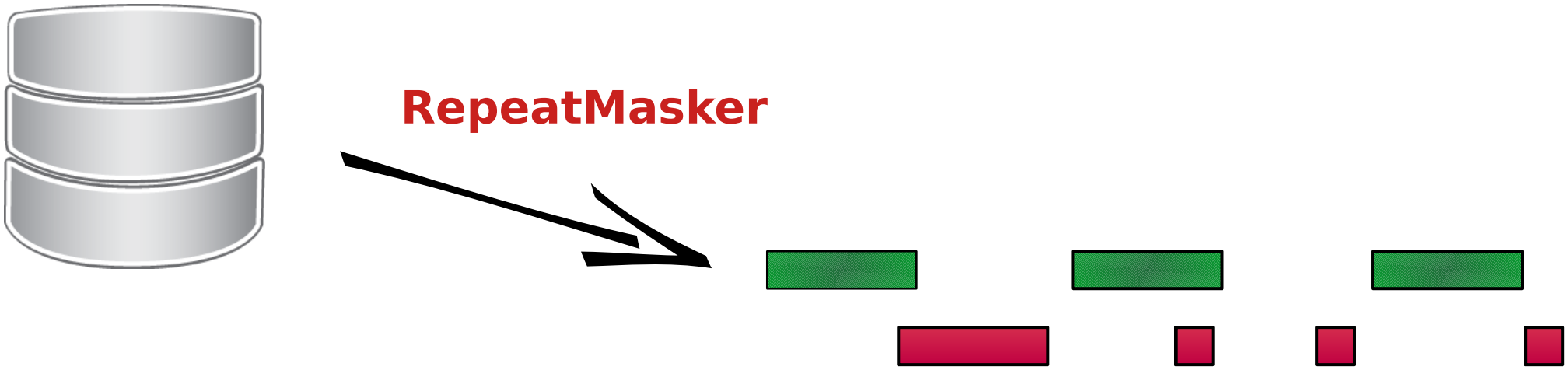
Library preparation

- Contig filtering
- Only clusters decribed in cluster table are used for library
- Fasta header - hierarchical classification

```
>CL1Contig1#Class/subclass/subclass
TAAGTAGTGTTTCCTTGTTAGAAGATACAAAGCCA
TGACTA
>CL2Contig4#Class/subclass/subclass
AGGATAAGCTTGCGGTTTAAGTTCTTATACTCAAT
```

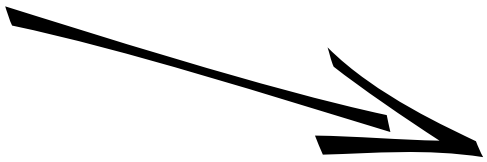
Library Based Repeat Annotation

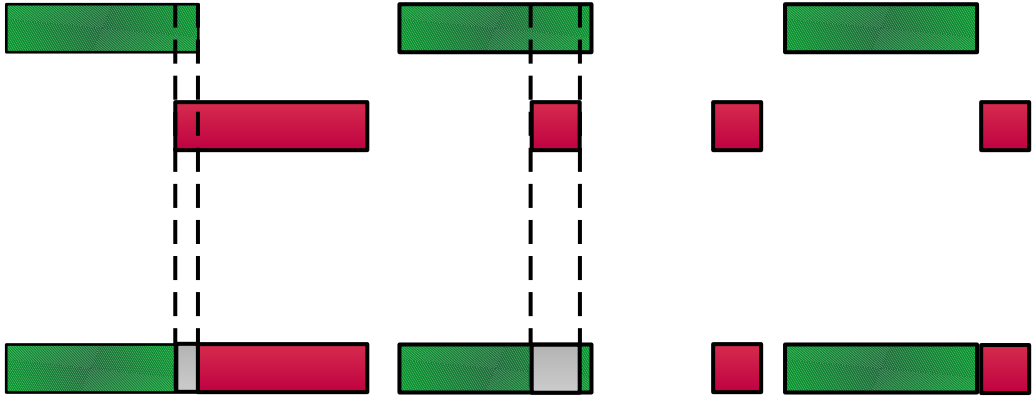
Library Based Repeat Annotation



Library Based Repeat Annotation

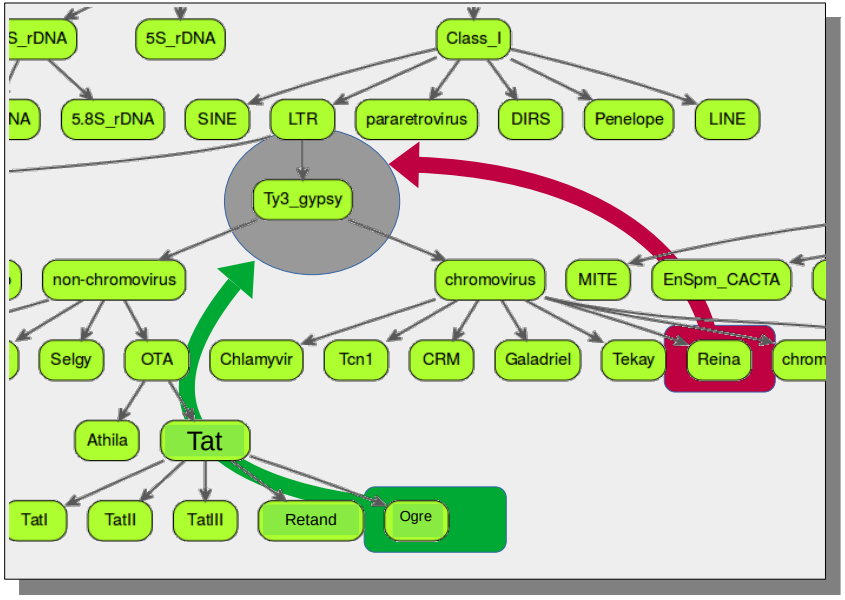
RepeatMasker

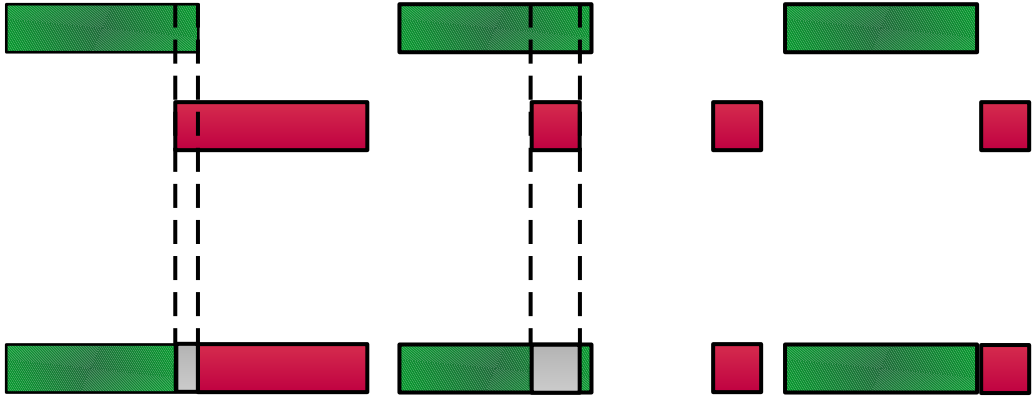




Library Based Repeat Annotation

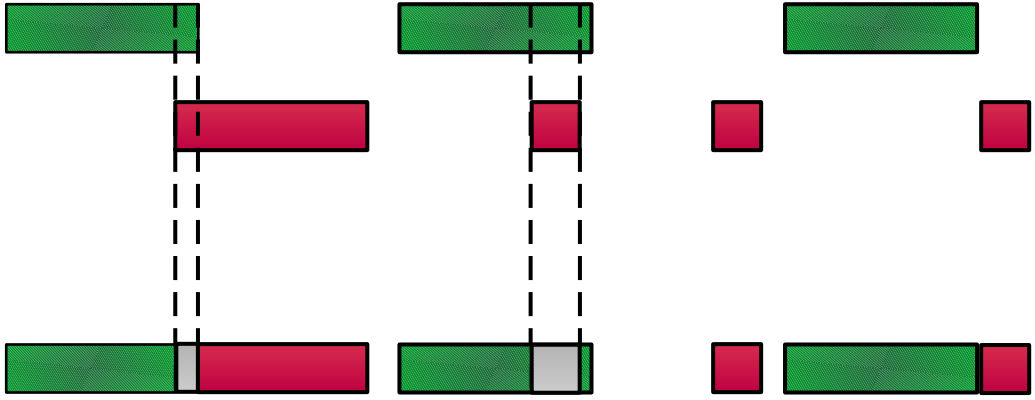
RepeatMasker





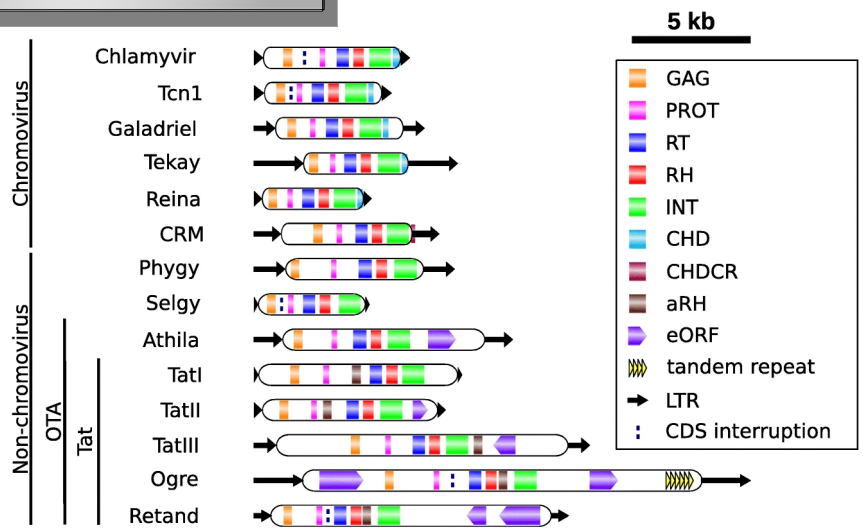
Library Based Repeat Annotation

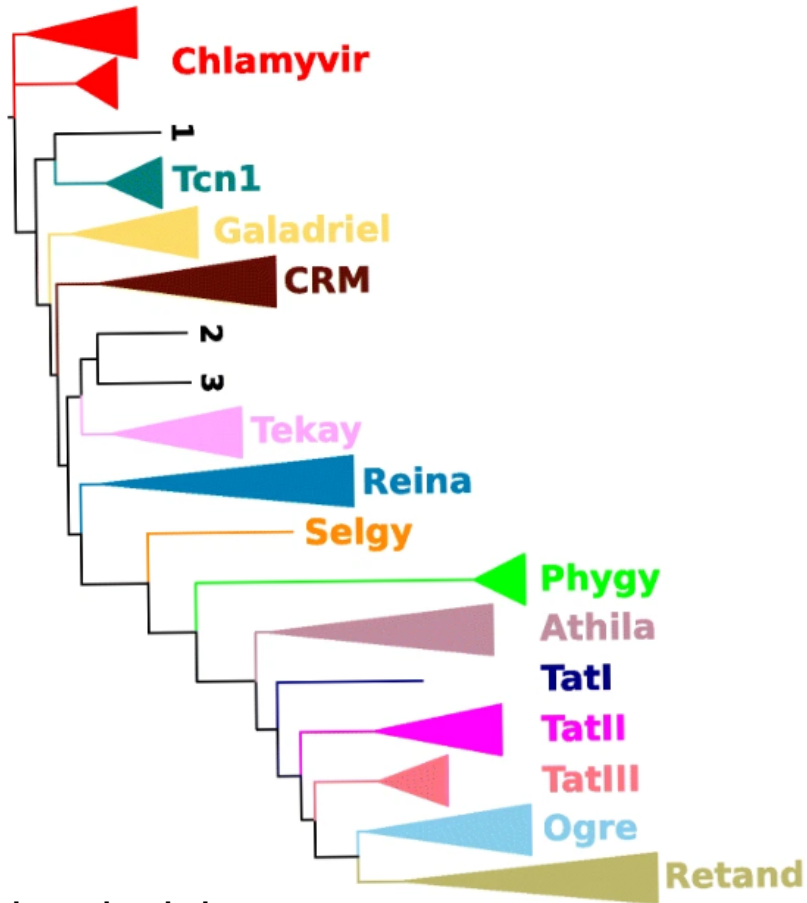
RepeatMasker



- GFF3 output
- Hierarchical user provided classification
- Information about conflicts is kept

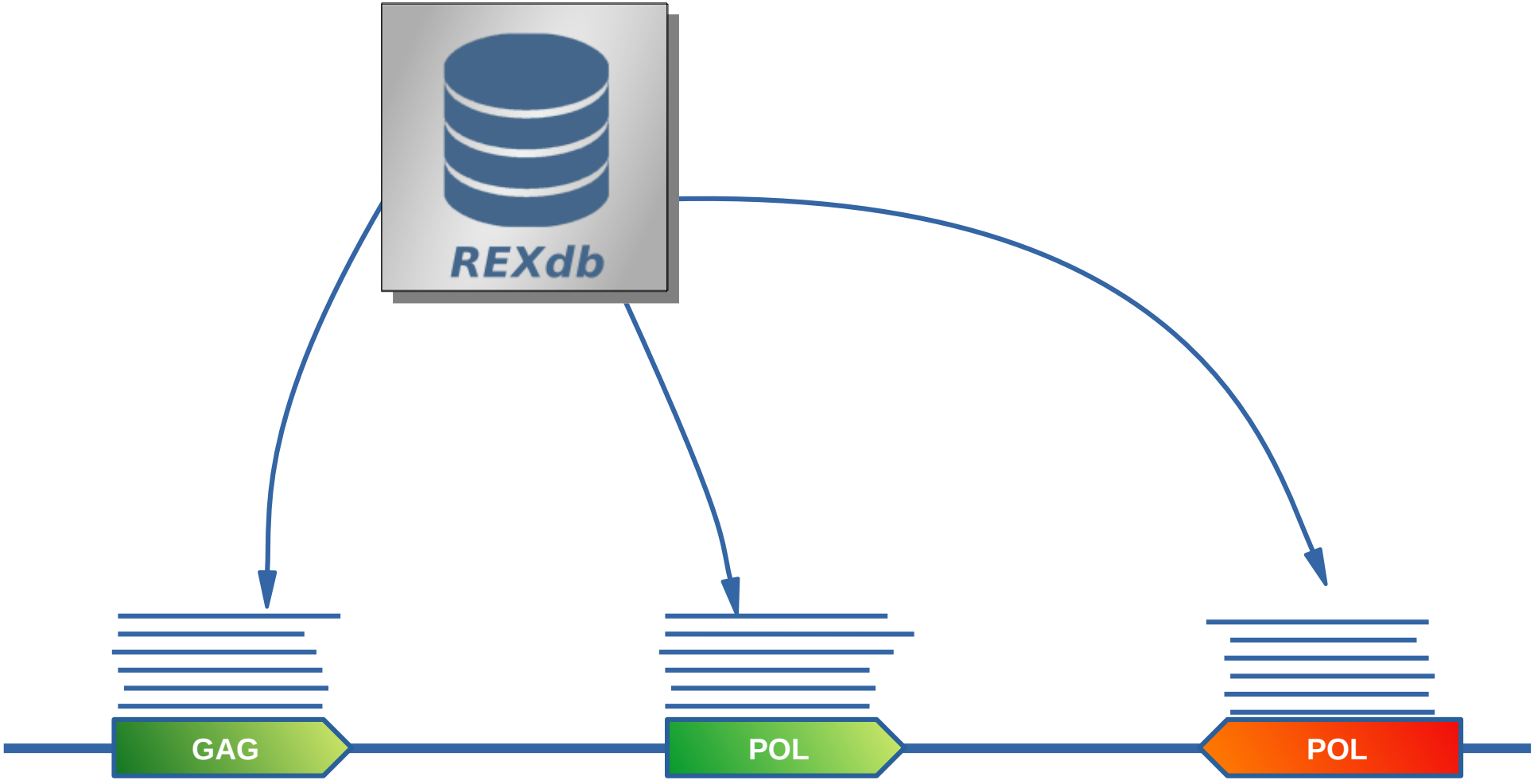
DANTE – domain based annotation of transposable elements



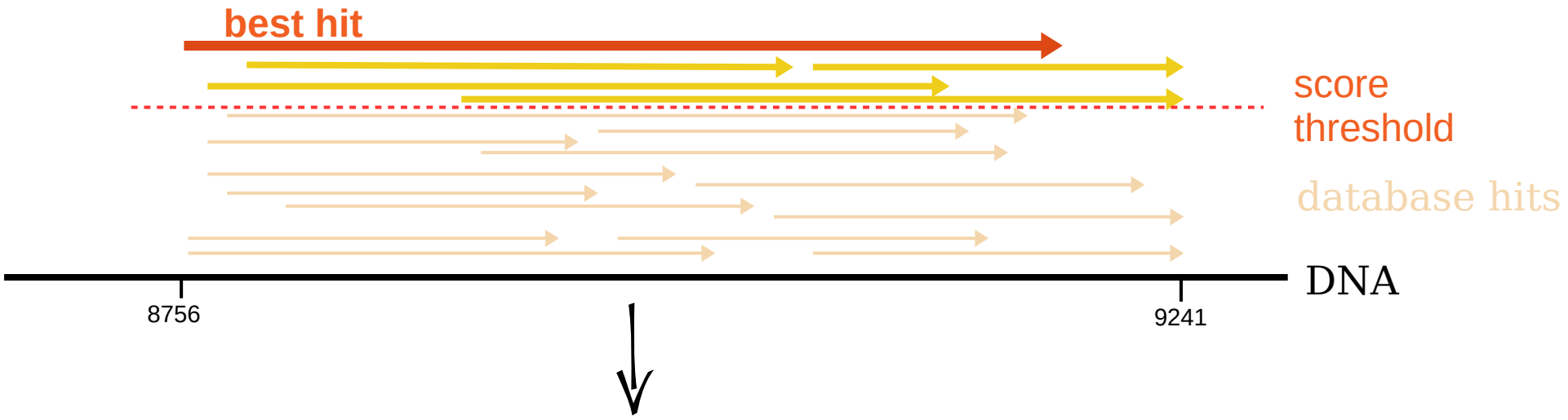


Annotation and classification based on phylogenetic principle

DANTE – domain based annotation of transposable elements



DANTE – domain based annotation of transposable elements



RT|Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Tat|TatV
RT|Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Tat|Retand
RT|Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Tat|Retand
RT|Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Tat|Ogre
RT|Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Tat|Ogre

REXdb Hits classifications

RT Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Tat

The most specific non-conflicting classification (LCA)

DANTE – domain based annotation of transposable elements

Custom library annotation

DANTE annotation

DANTE – domain based annotation of transposable elements

Custom library annotation

DANTE annotation

DANTE – domain based annotation of transposable elements

Custom library annotation

DANTE annotation

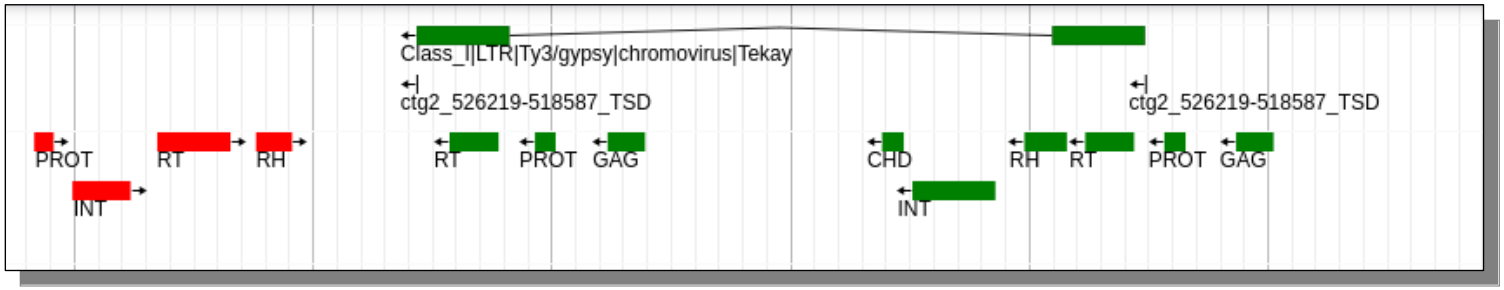
DANTE LTR – structure based annotation

Suitable tools?

- LTR_Finder
- LTRharvest
- LTRdigest
- LTRDetector
- LTR_STRUCT

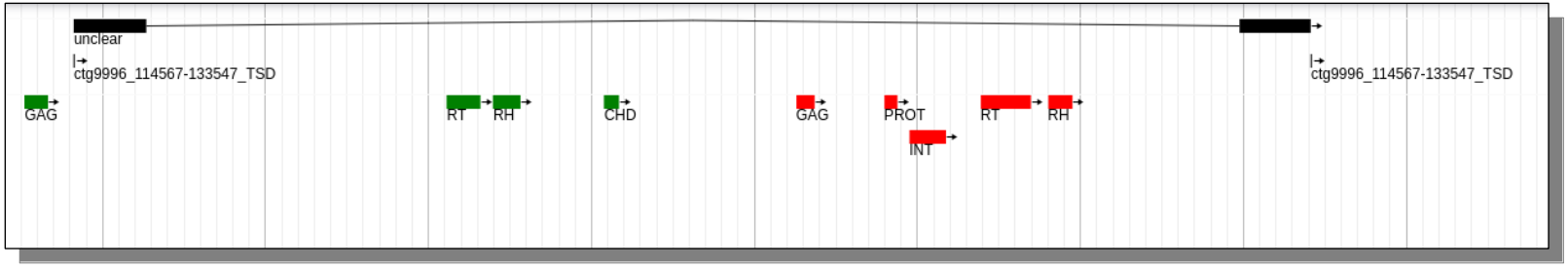
- extremely slow
- high number of false positive
- false negative
- limited protein domain database
  (Pfam, PROSITE, Uniprot)

DANTE LTR – structure based annotation



LTR_Finder

DANTE



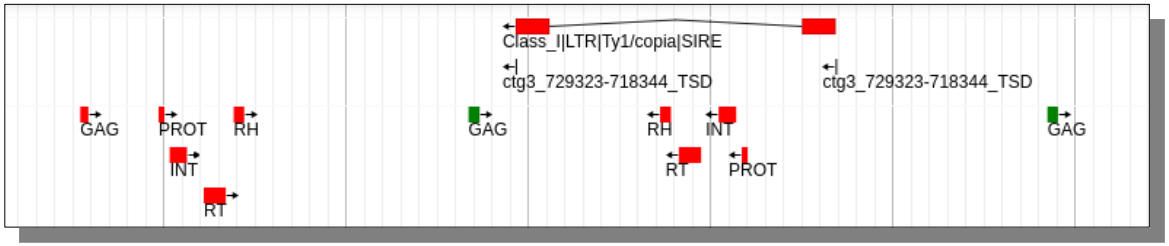
LTR_Finder

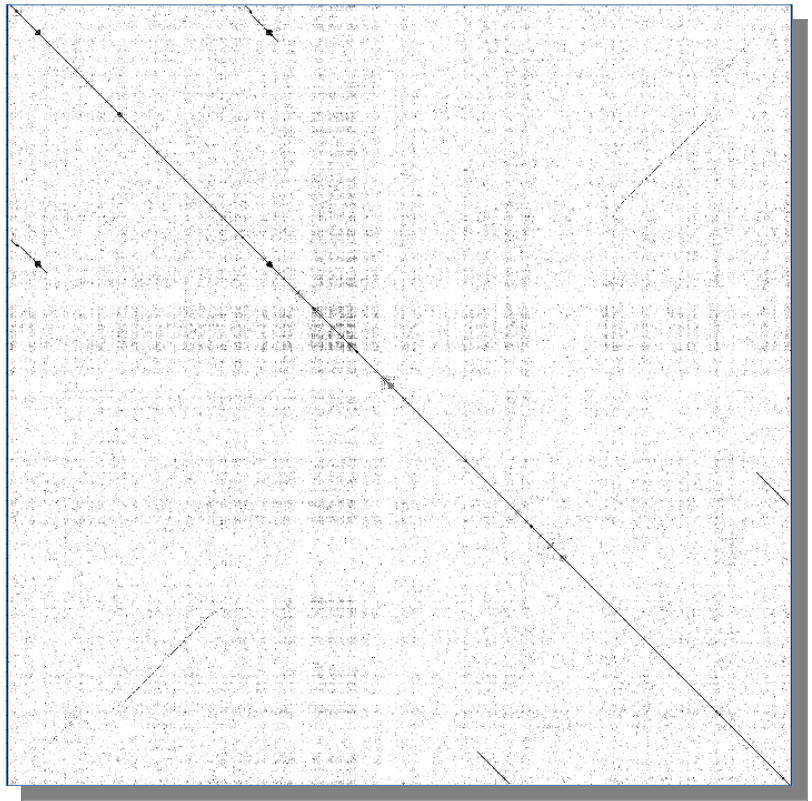
DANTE

Ty3/gypsy

Ty1/copia

DANTE LTR – structure based annotation

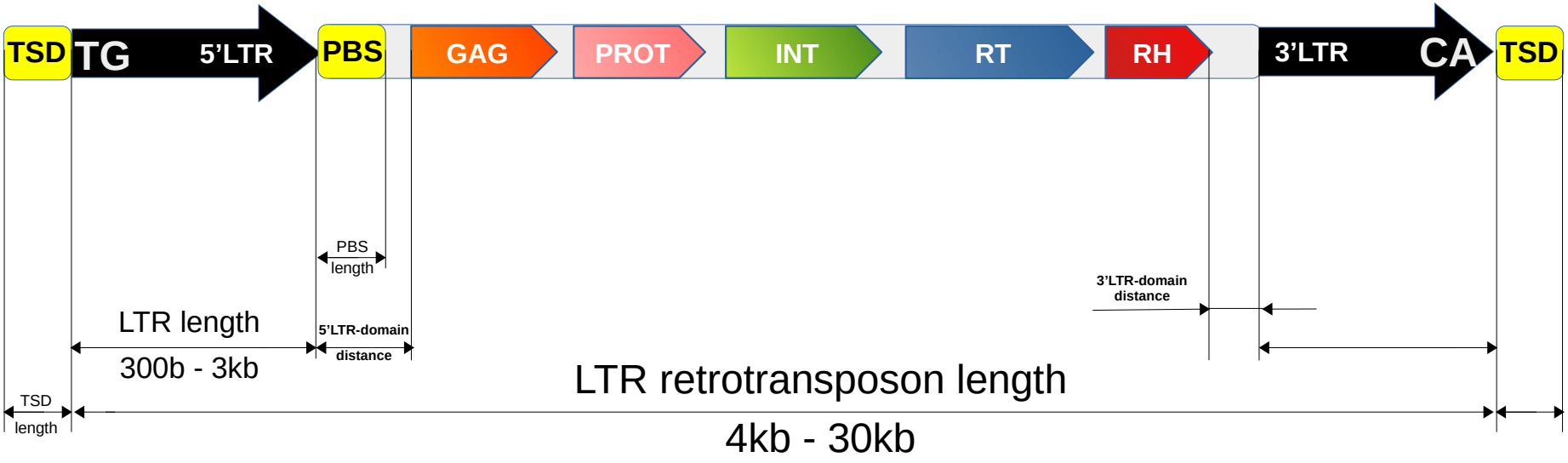




LTR_Finder

DANTE

DANTE LTR – structure based annotation



Search constraints:

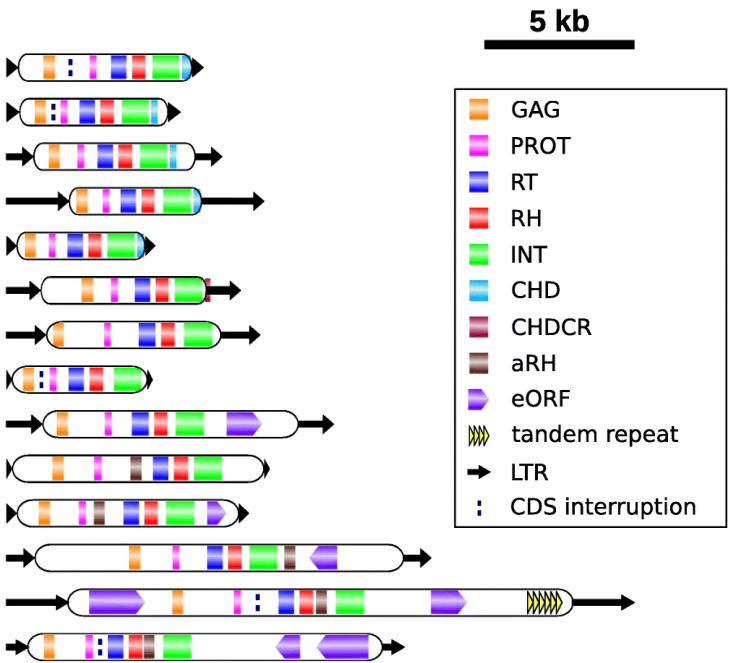
- Feature length (min, max)
- Presence of features (PBS, PPT, TSD, TG/CA, domains)

DANTE LTR – structure based annotation

- Large variability in feature lengths
- Lineage specific protein domains
- Superfamily specific protein domains order

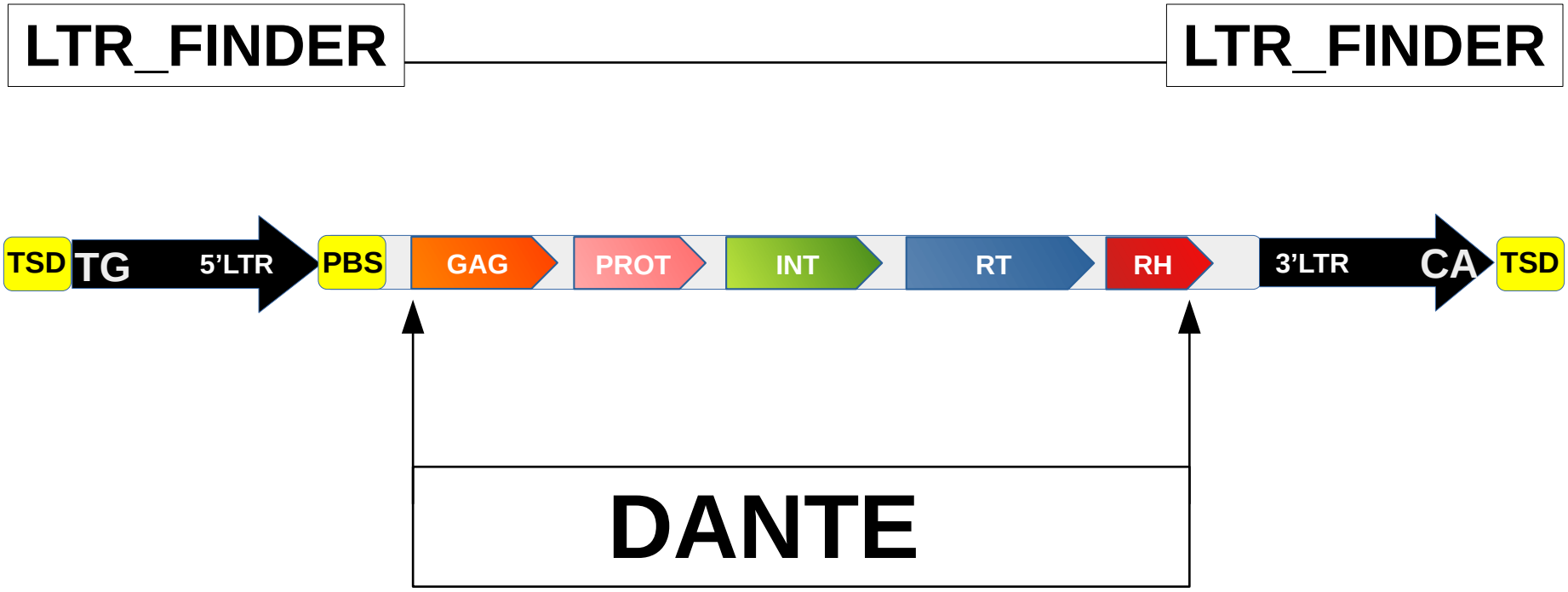
A Ty3/gypsy

		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
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



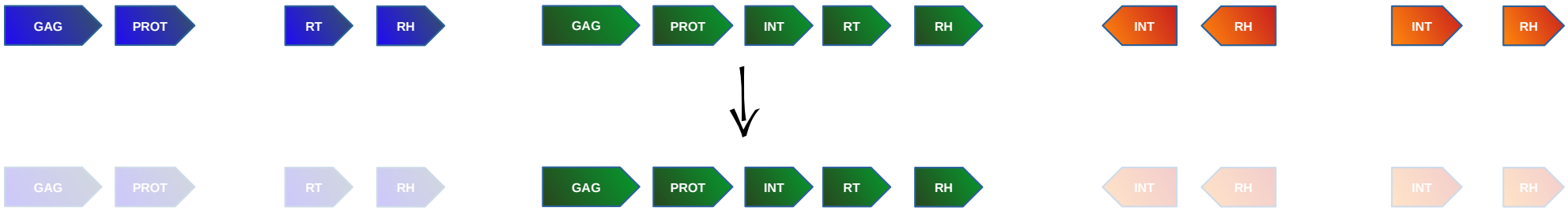
(Neumann et al, 2019)

DANTE LTR – structure based annotation

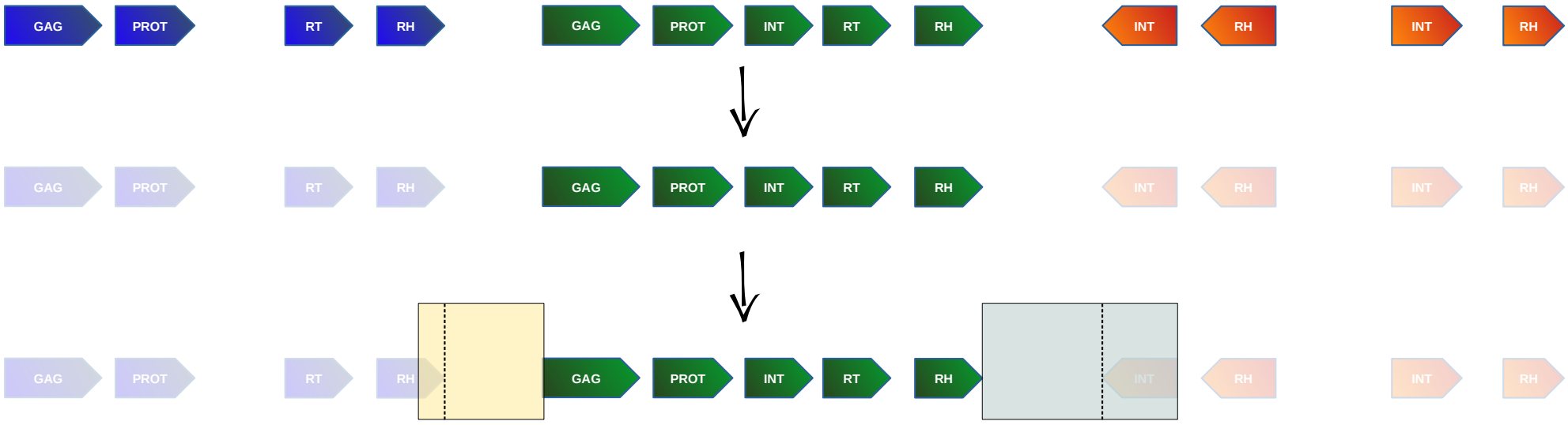


DANTE LTR – WORKFLOW

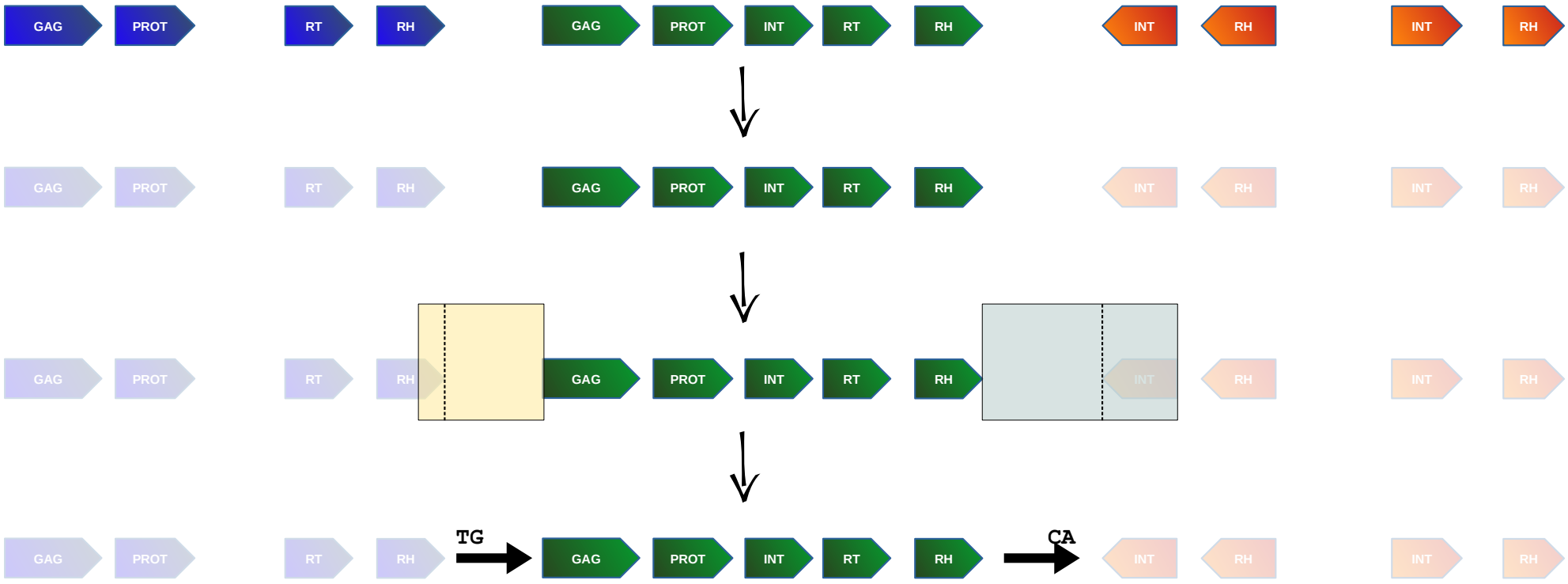
DANTE LTR – WORKFLOW



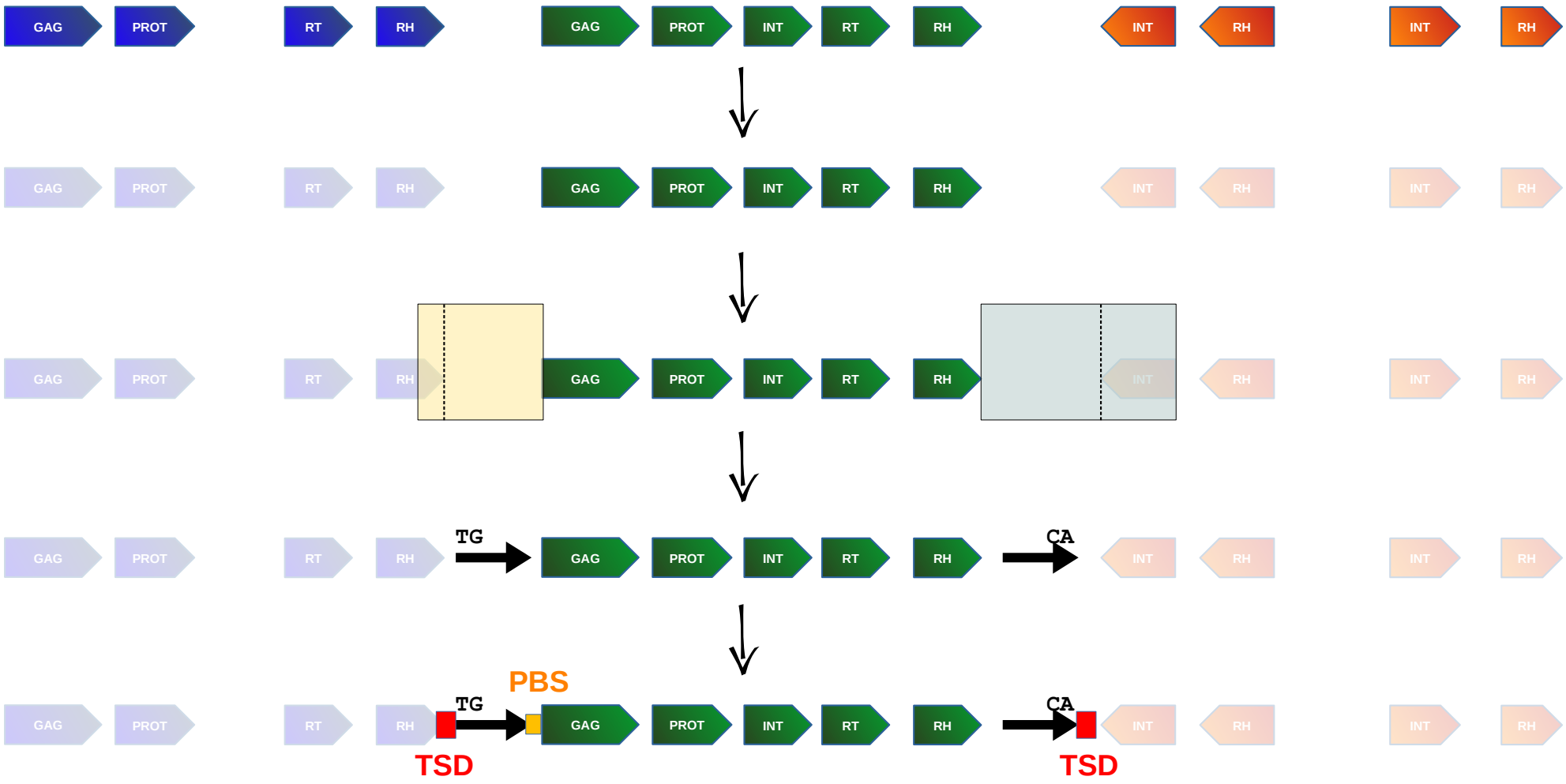
DANTE LTR – WORKFLOW



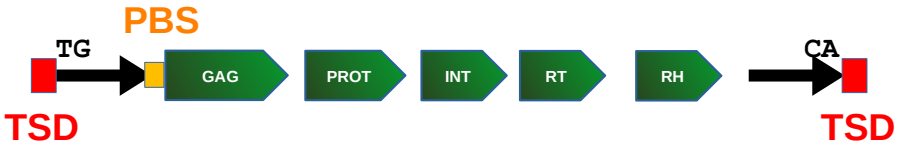
DANTE LTR – WORKFLOW



DANTE LTR – WORKFLOW



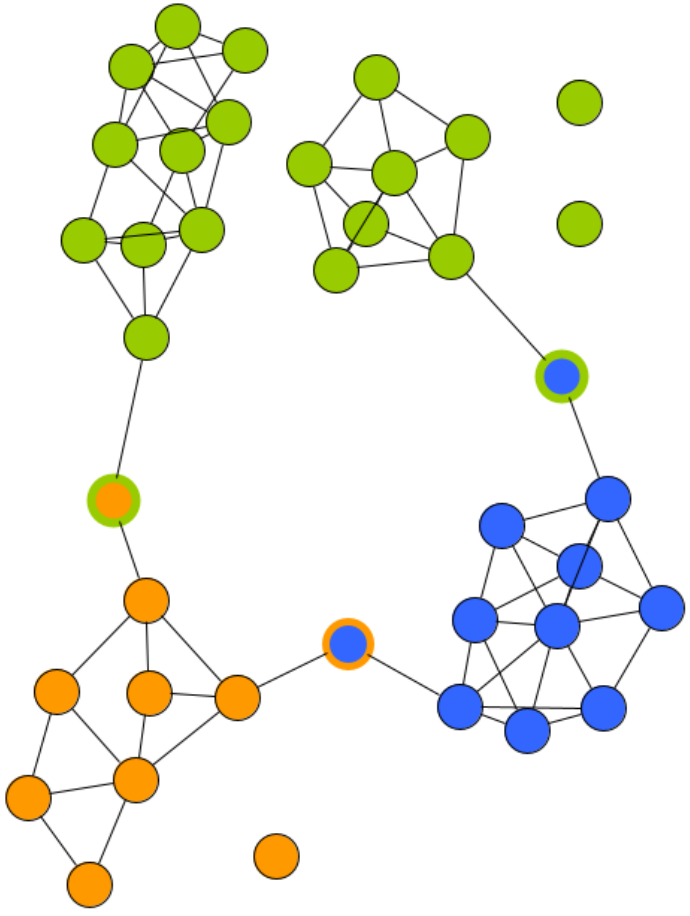
DANTE LTR – Retrotransposons ranks



Rank	Annotation
DLTP	Elements with identified protein Domains, 5'LTR, 3'LTR, TSD and PBS
DLP	Elements with identified protein **D**omains, 5'**L**TR, 3'LTR and **P**BS (TSD was not found)
DLT	Elements with identified protein **D**omains, 5'**L**TR, 3'LTR and **T**SD (PBS was not found)
DL	Elements with protein **D**omains, 5'**L**TR and 3'LTR (PBS and LDS were not found)

DANTE LTR – Filtering

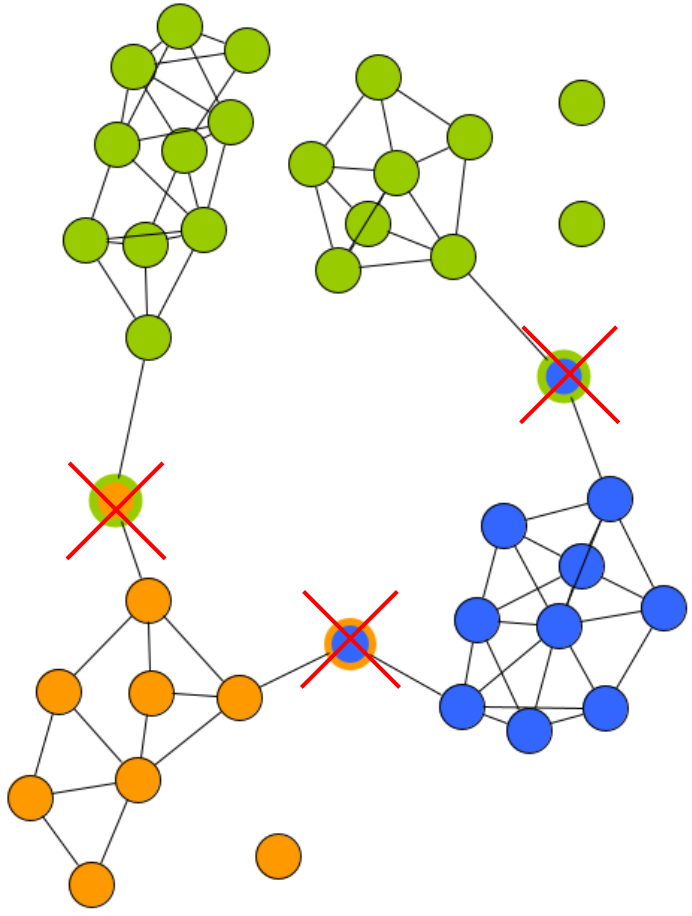
All-to-All sequence comparison



DANTE LTR – Filtering

All-to-All sequence comparison

Cross-lineage similarities

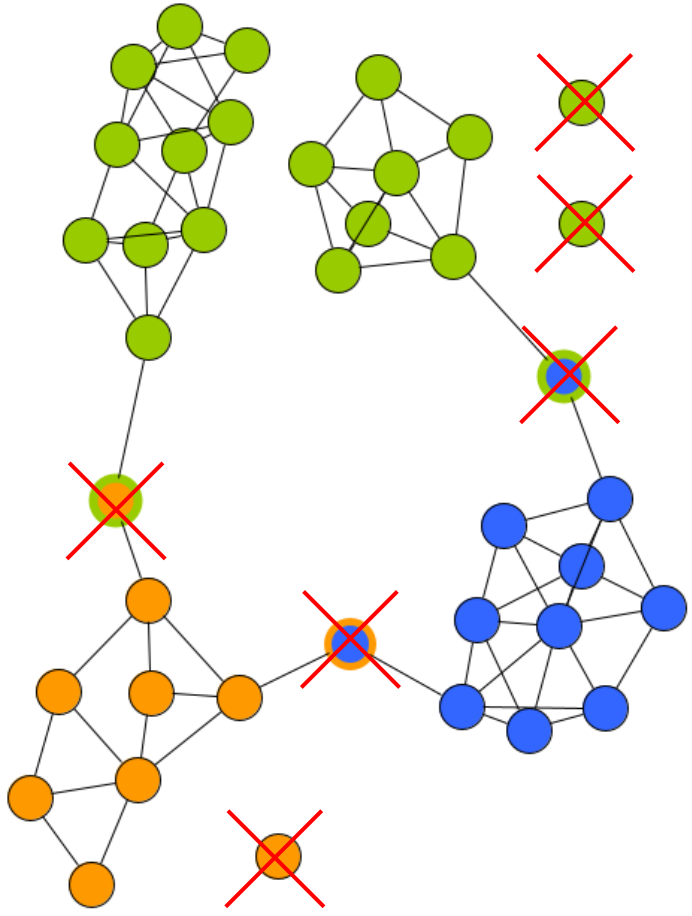


DANTE LTR – Filtering

All-to-All sequence comparison

Cross-lineage similarities

Minimal copy number (DLP, DLT)



DANTE LTR - Output

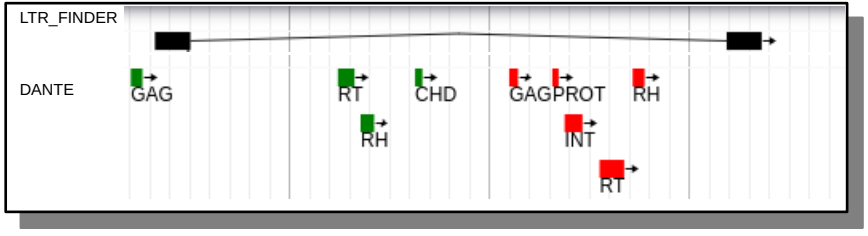
	SOURCE	TYPE	START	END				ATTRIBUTES
1	dante_ltr	transposable_element	3780765	3785720	.	+	.	ID=TE_00000001;LTR_Identity=100;LTR5_length=440;LTR3_length=440; TSD=CTTGT;Final_Classification=Class_I|LTR|Ty1/copia|Ivana;Region_Hits_Classifications=NA; trna_id=ATCAAAACCTAGCTCTGATAcca__Met-3x;Rank=DLTP
1	dante_ltr	long_terminal_repeat	3785281	3785720	.	+	.	LTR_Identity=100;Final_Classification=Class_I|LTR|Ty1/copia|Ivana; LTR=3LTR;Parent=TE_00000001;Region_Hits_Classifications=NA;Rank=DLTP
1	dante_ltr	long_terminal_repeat	3780765	3781204	.	+	.	LTR_Identity=100;Final_Classification=Class_I|LTR|Ty1/copia|Ivana;L TR=5LTR;Parent=TE_00000001;Region_Hits_Classifications=NA;Rank=DLTP
1	dante	protein_domain	3781451	3781729	498	+	.	Final_Classification=Class_I|LTR|Ty1/copia|Ivana;Parent=TE_00000001;Name=GAG;Region_Hits_Clas
1	dante	protein_domain	3782237	3782452	406	+	.	Final_Classification=Class_I|LTR|Ty1/copia|Ivana;Parent=TE_00000001;Name=PROT;Region_Hits_Cla
1	dante	protein_domain	3782639	3783238	1132	+	.	Final_Classification=Class_I|LTR|Ty1/copia|Ivana;Parent=TE_00000001;Name=INT;Region_Hits_Clas
1	dante	protein_domain	3783782	3784549	1448	+	.	Final_Classification=Class_I|LTR|Ty1/copia|Ivana;Parent=TE_00000001;Name=RT;Region_Hits_Class
1	dante	protein_domain	3784817	3785197	728	+	.	Final_Classification=Class_I|LTR|Ty1/copia|Ivana;Parent=TE_00000001;Name=RH;Region_Hits_Class
1	dante_ltr	target_site_duplication	3785721	3785725	.	.	.	Parent=TE_00000001;Region_Hits_Classifications=NA;Rank=DLTP
1	dante_ltr	target_site_duplication	3780760	3780764	.	.	.	Parent=TE_00000001;Region_Hits_Classifications=NA;Rank=DLTP
1	dante_ltr	primer_binding_site	3781208	3781220	.	+	.	Parent=TE_00000001;Region_Hits_Classifications=NA;trna_id=ATCAAAACCTAGCTCTGATAcca__Met-3x;Ran

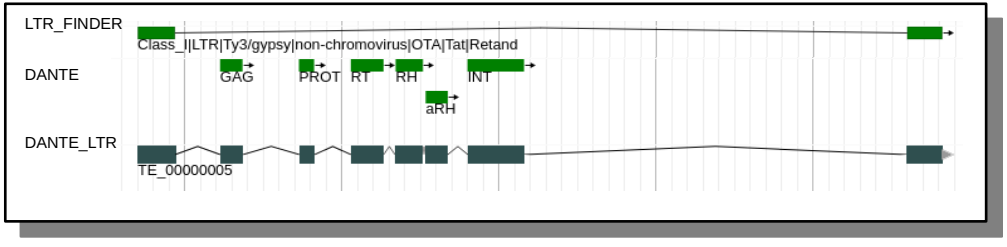
DANTE LTR vs. LTR_FINDER

		LTR_FINDER	DANTE_LTR (unfiltered)	DANTE_LTR (clean)
Total		10975	8982	4515
With protein domain(s)	Single lineage	6739	8982	4515
	Multiple lineages (conflict)	281	-	-

DANTE LTR vs. LTR_FINDER

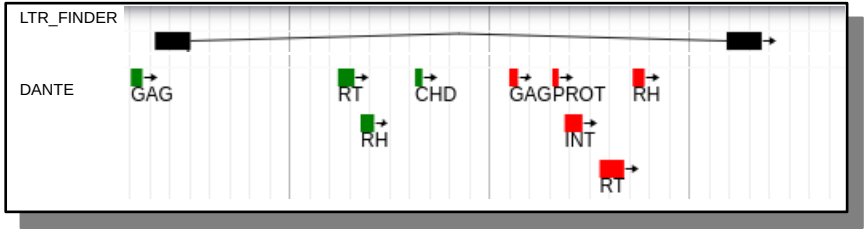
		LTR_FINDER	DANTE_LTR (unfiltered)	DANTE_LTR (clean)
Total		10975	8982	4515
With protein domain(s)	Single lineage	6739	8982	4515
	Multiple lineages (conflict)	281	-	-

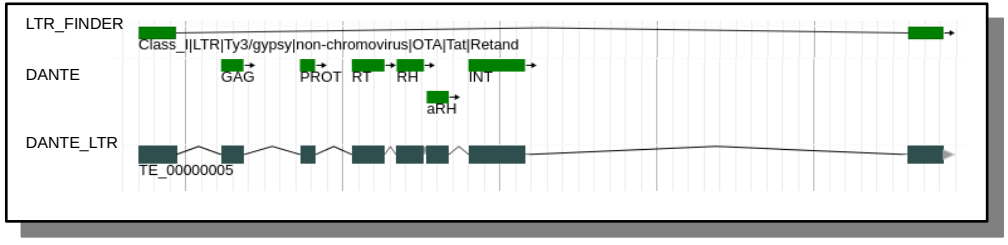


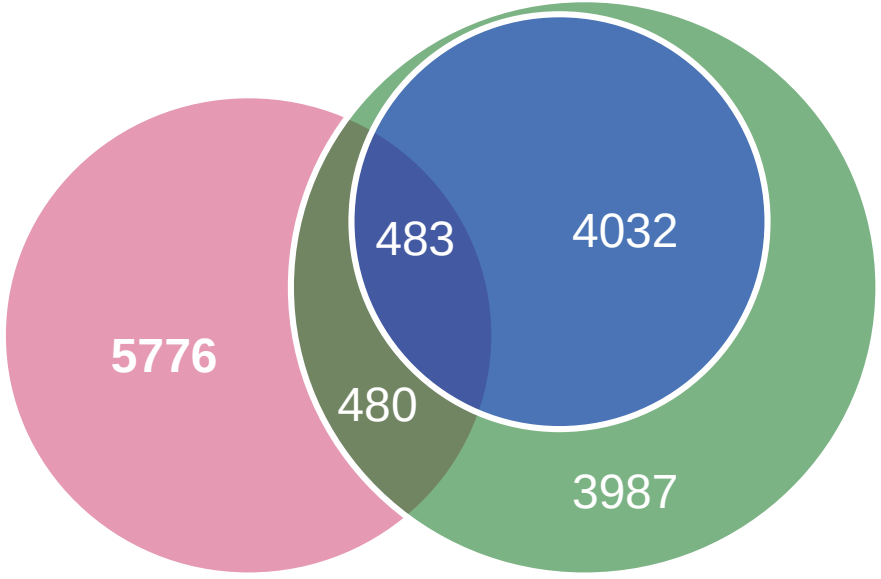


DANTE LTR vs. LTR_FINDER

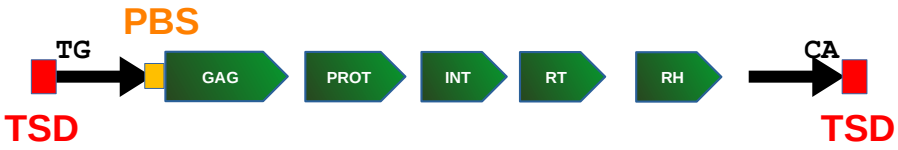
		LTR_FINDER	DANTE_LTR (unfiltered)	DANTE_LTR (clean)
Total		10975	8982	4515
With protein domain(s)	Single lineage	6739	8982	4515
	Multiple lineages (conflict)	281	-	-





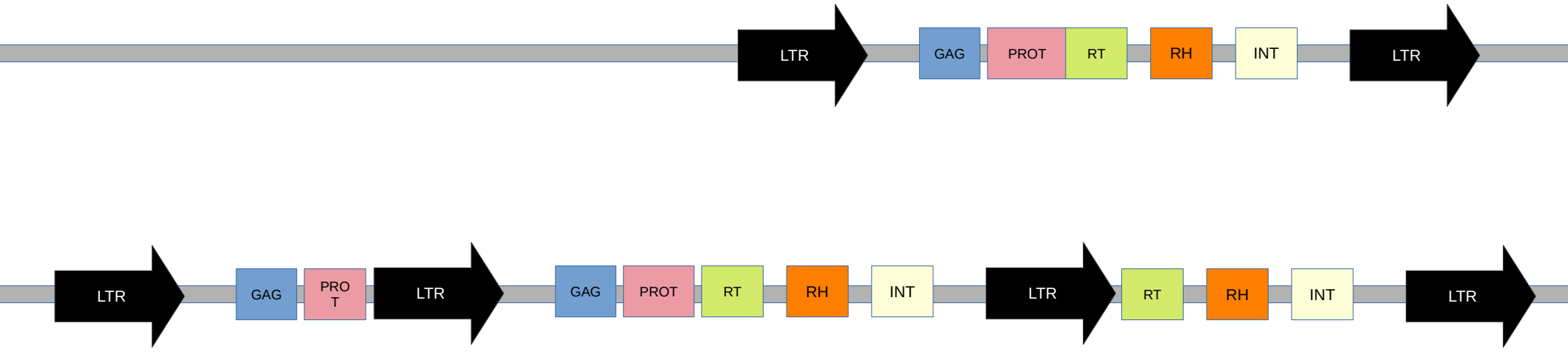


DANTE LTR – Applications



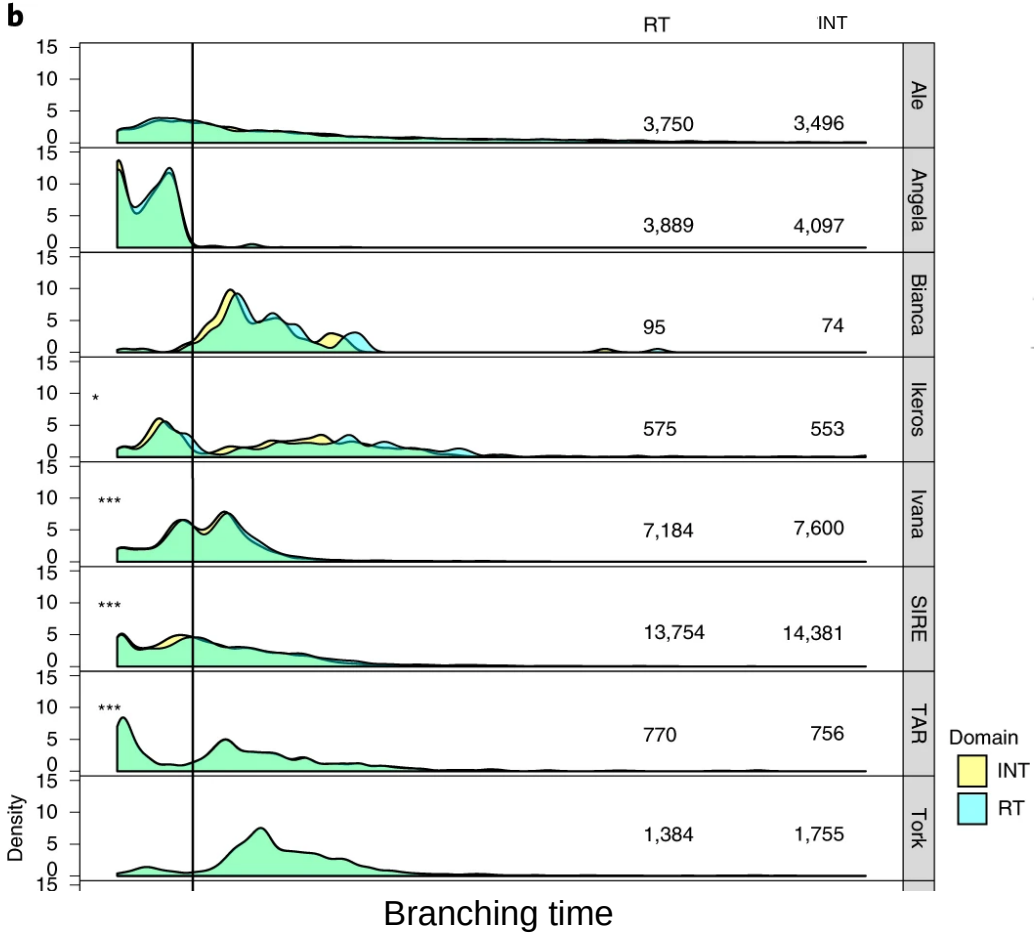
- Custom library for assembly annotation
- Dating of retrotransposons activity ???

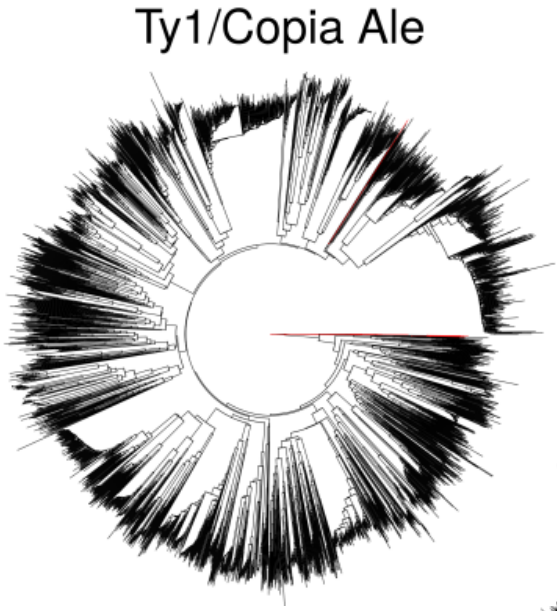
Dating of retrotransposons activity

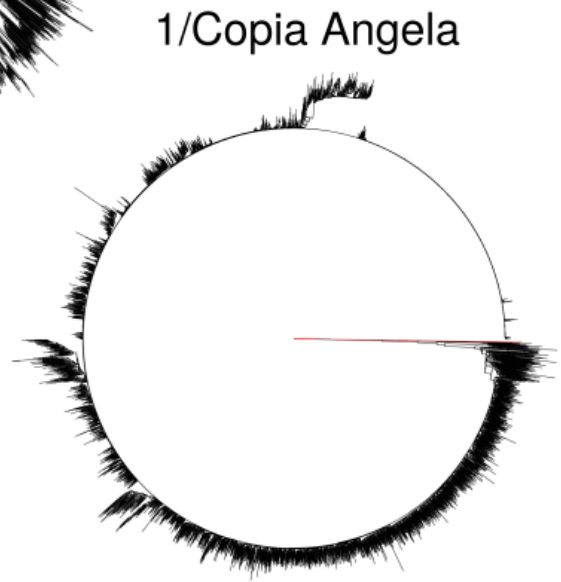


Nested TE insertion

Dating of retrotransposons activity



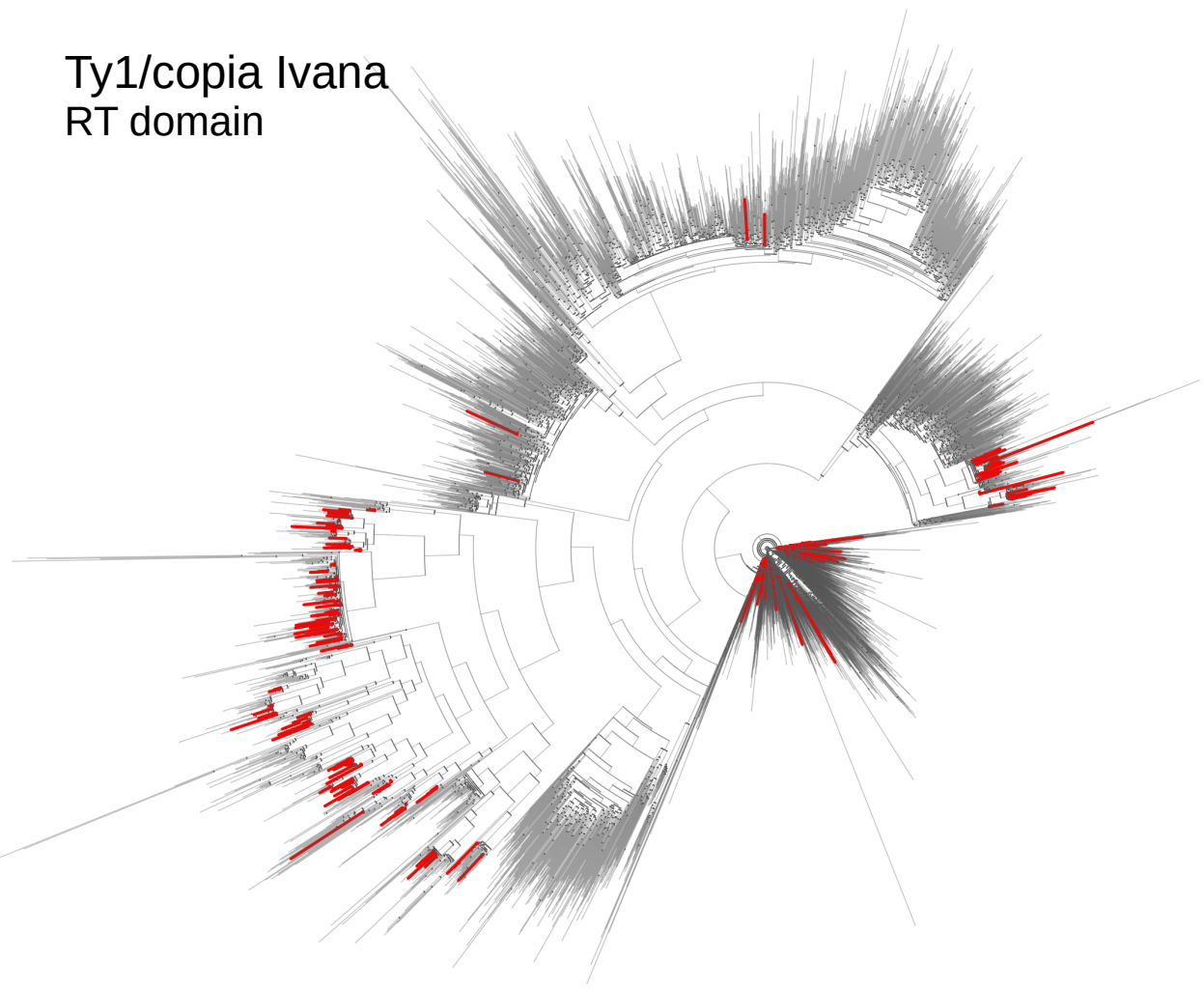


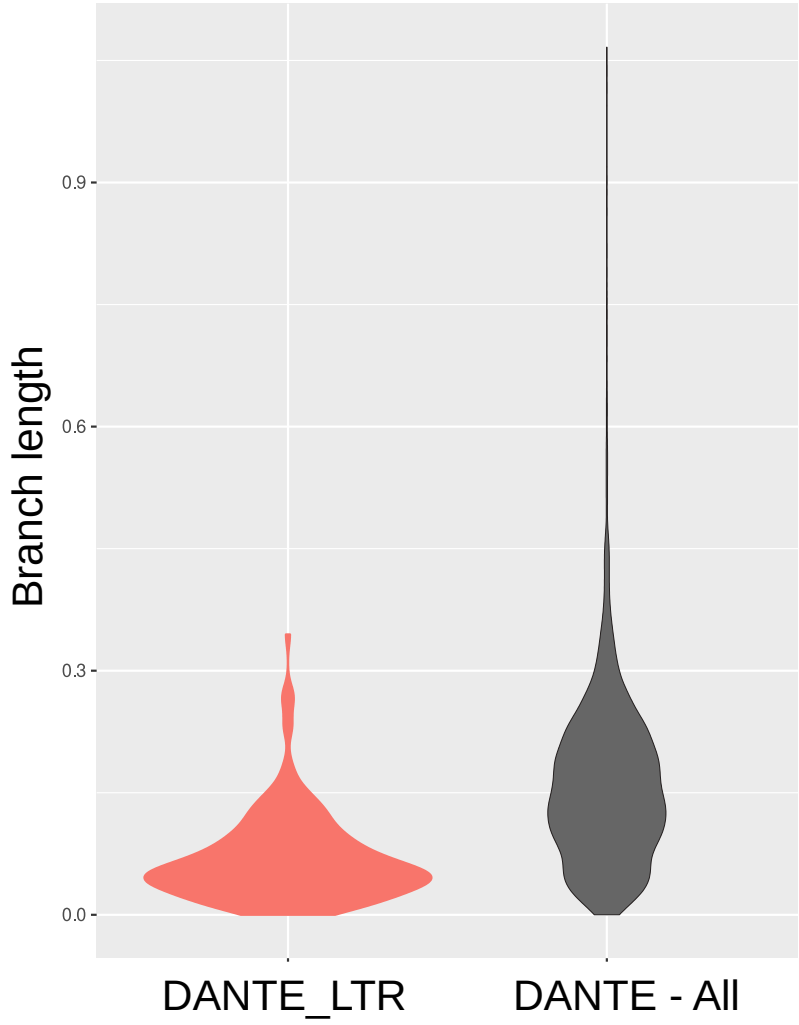


Kreplak, J., Madoui, MA., Cápal, P. et al. A reference genome for pea provides insight into legume genome evolution. Nat Genet 51, 1411–1422 (2019). https://doi.org/10.1038/s41588-019-0480-1

DANTE LTR

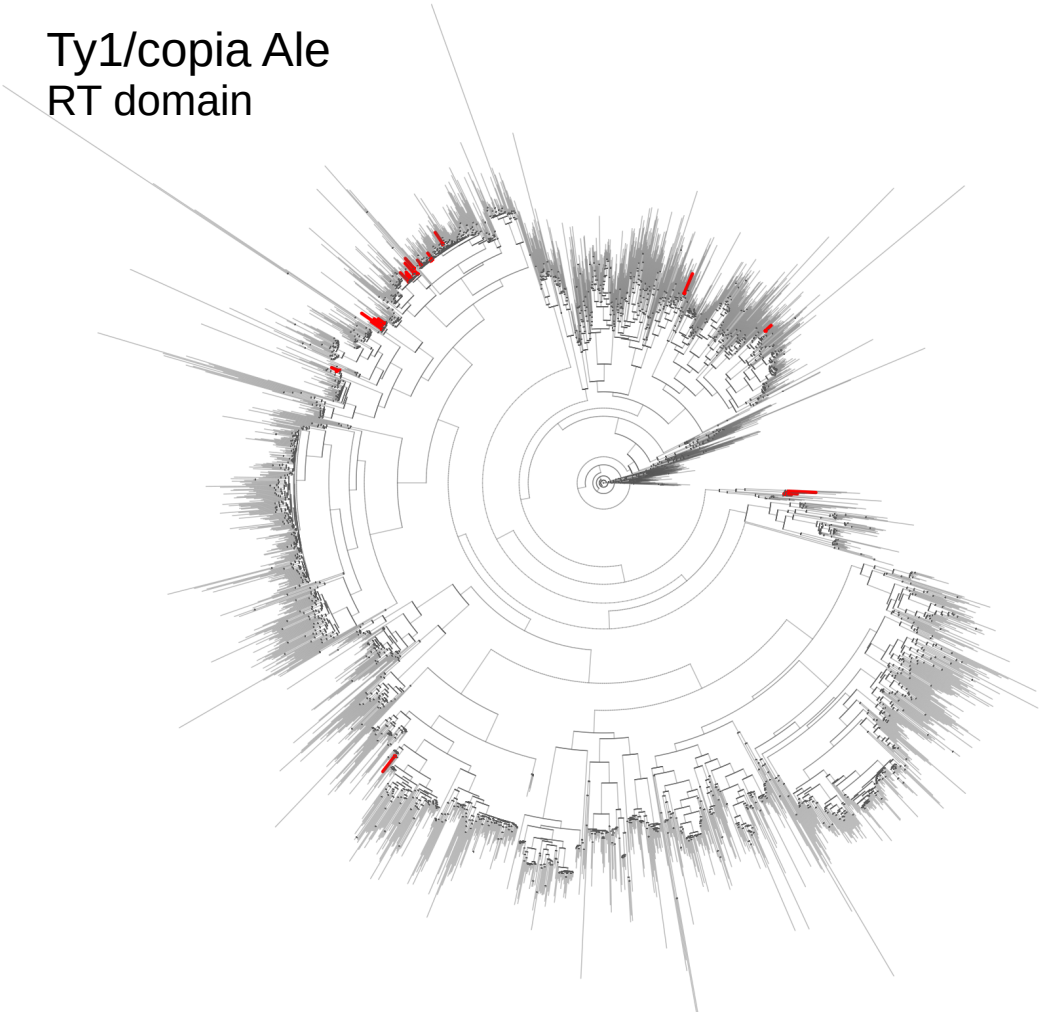
Ty1/copia Ivana
RT domain

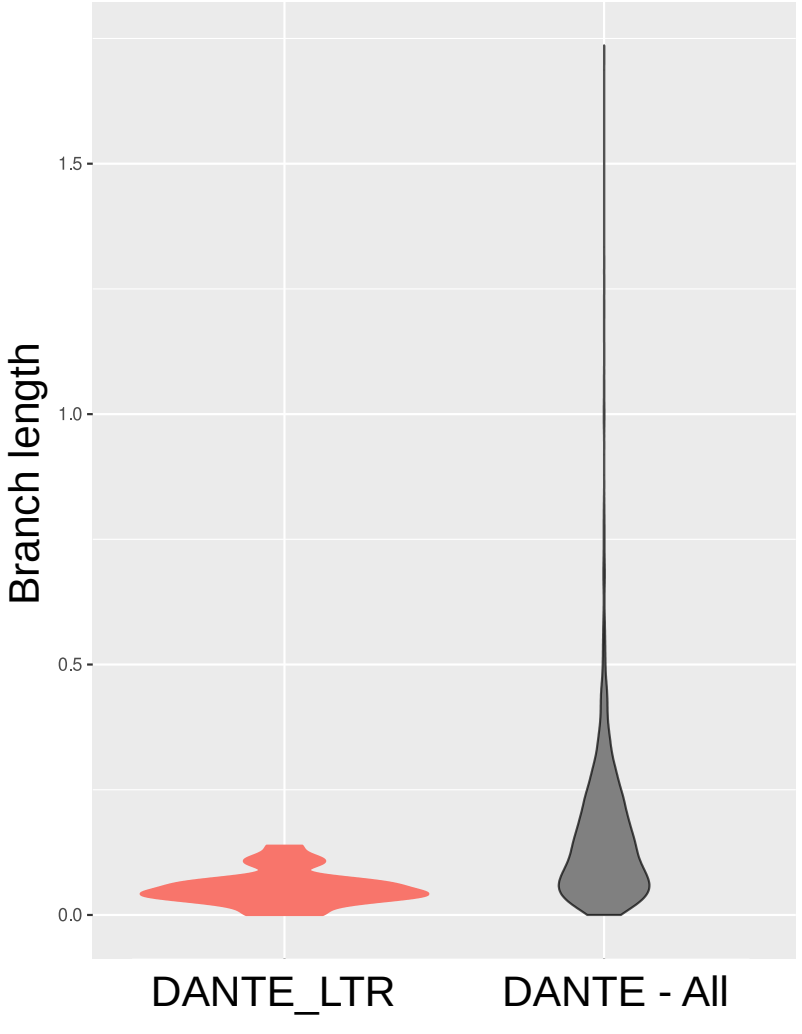




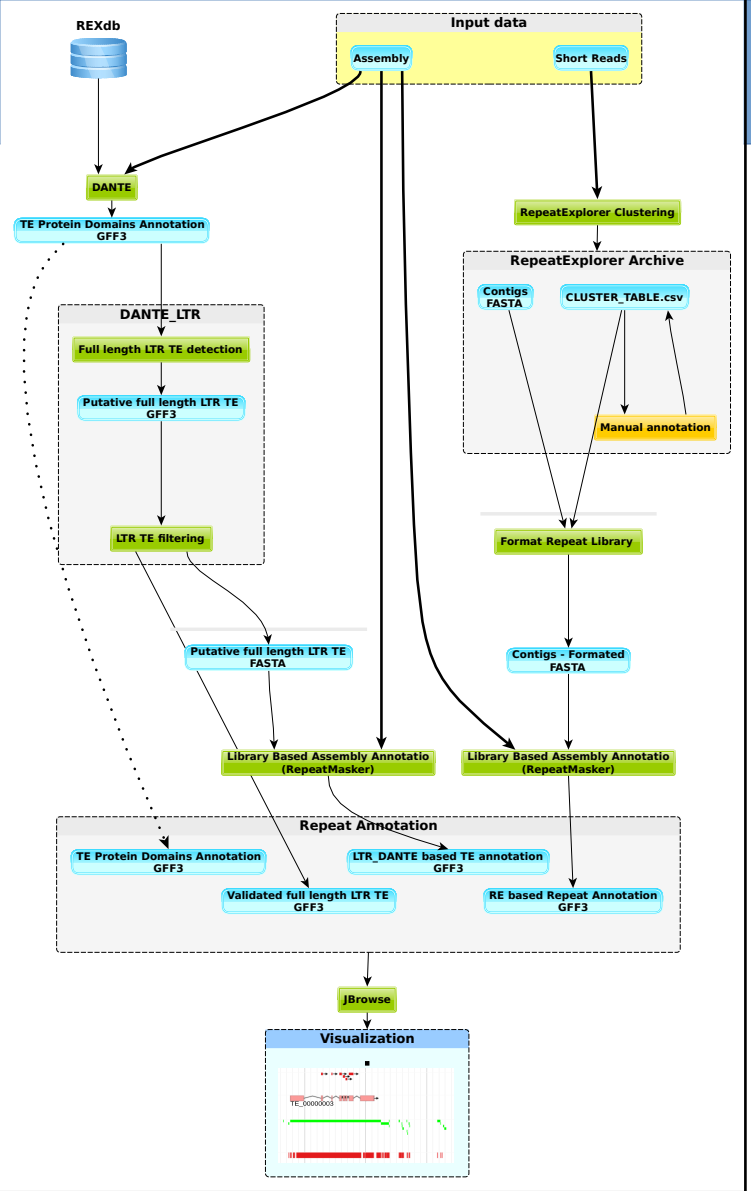
DANTE LTR

Ty1/copia Ale
RT domain





Complete workflow



- RepeatExplorer to create custom library
- Library based annotation using RepeatMasker

- DANTE
- DANTE_LTR to identify intact LTR retrotransposons
- Use intact elements as second custom library

Acknowledgment

Laboratory of Molecular Cytogenetics
Jiri Macas
Pavel Neumann
Nina Hostakova
Petr Novak

Biology Centre
Czech Academy of Sciences

Masaryk University – CERIT-SC
Zdenek Salvet
Ivana Krenkova
Martin Demko

cerit scientific cloud

elixir CZECH REPUBLIC