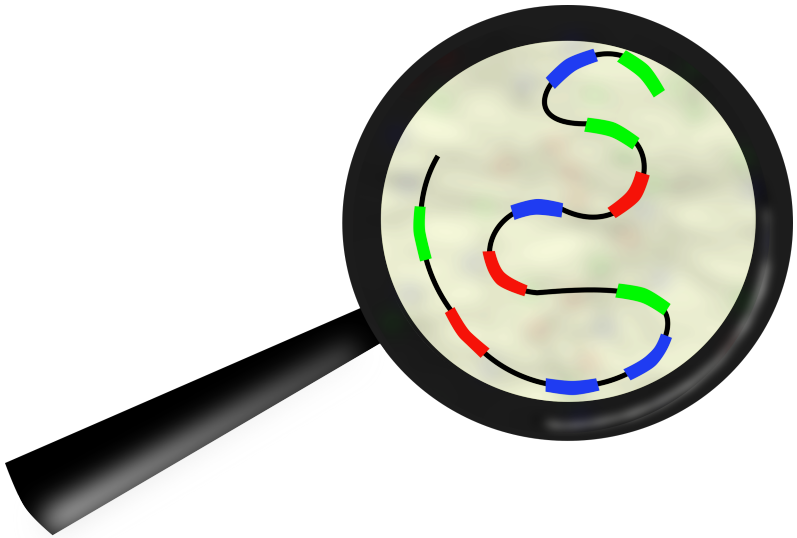
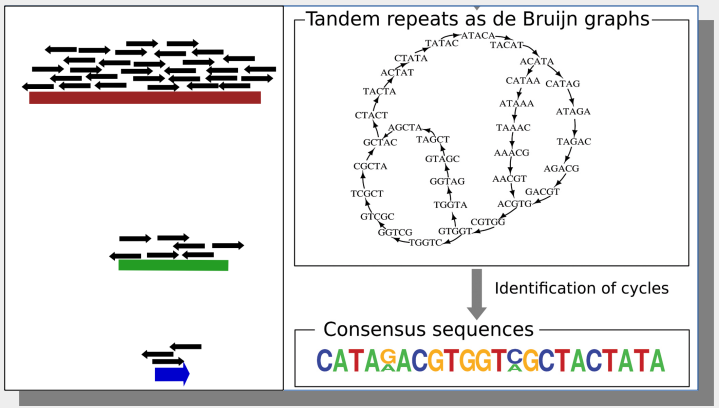
RepeatExplorer tools for genome annotation

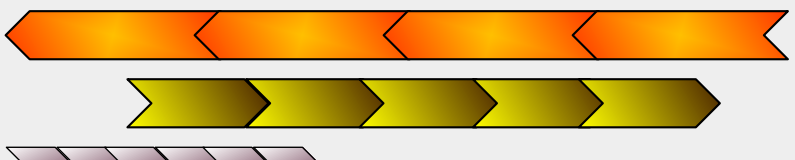


RepeatExplorer server tools for genome annotation

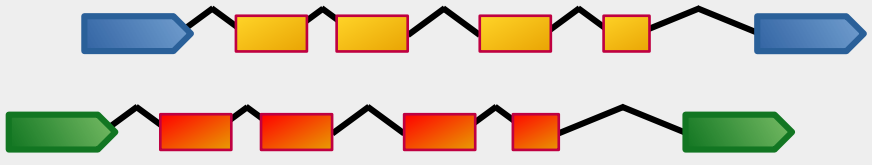


Library based annotation

DANTE



TideCluster

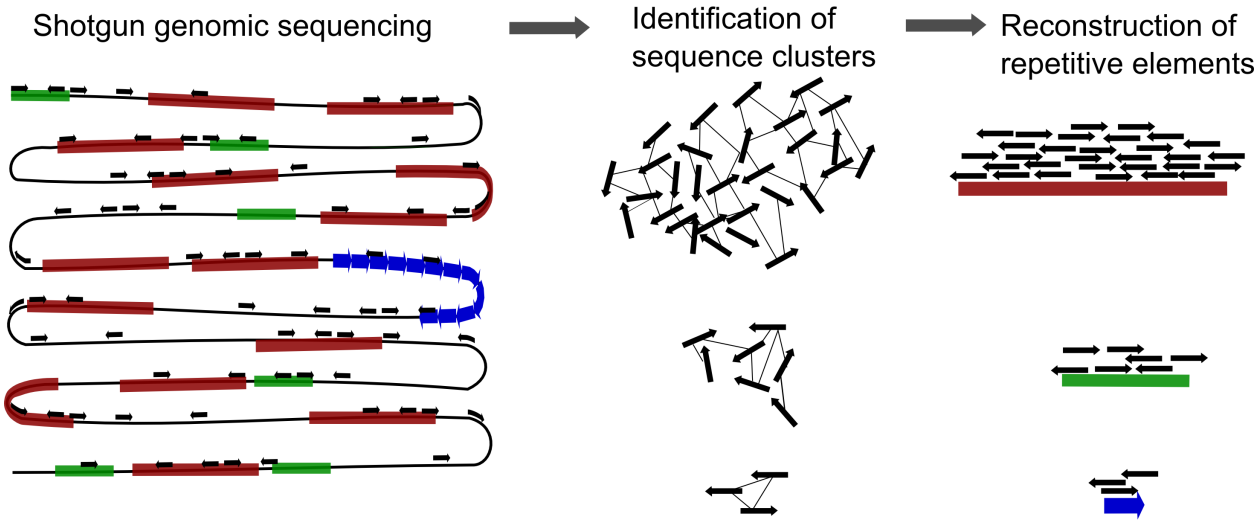


DANTE LTR

Library Based Repeat Annotation

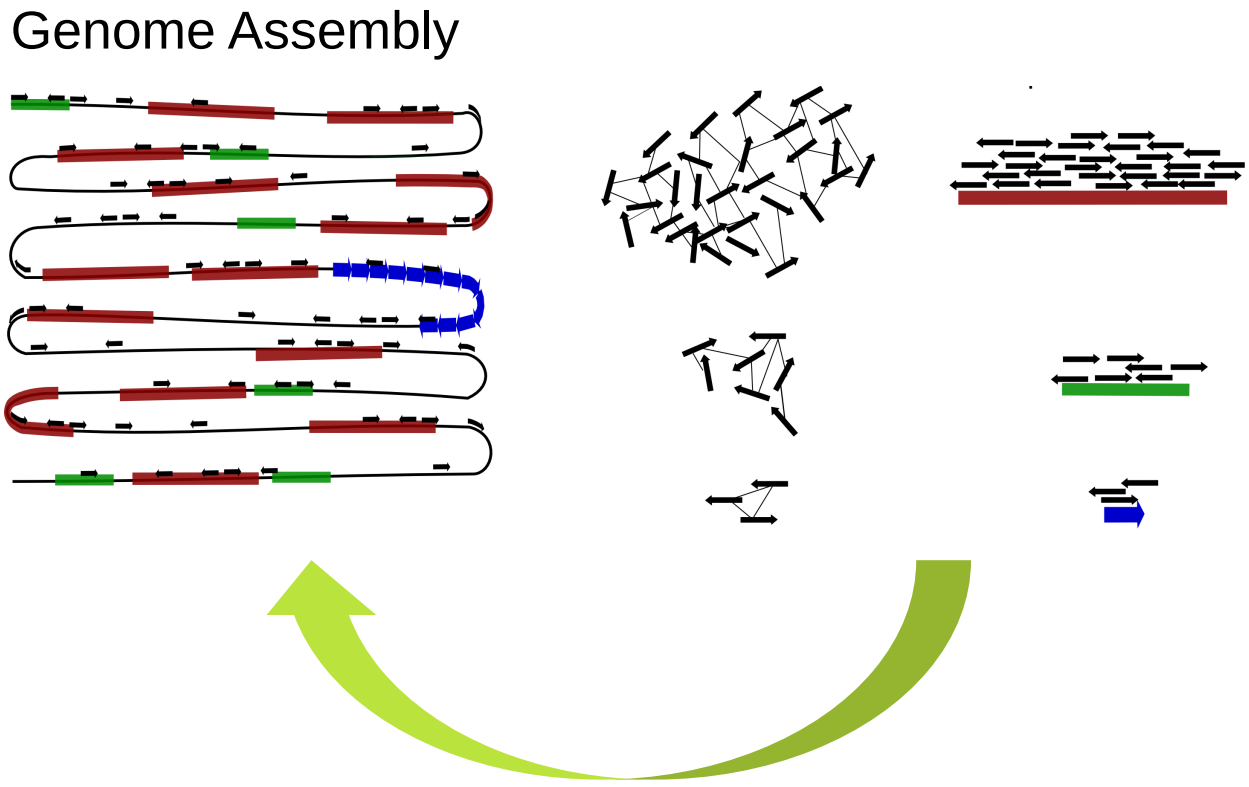
- Repbase
- DFAM
- **Custom library**

RepeatExplorer tools for genome annotation



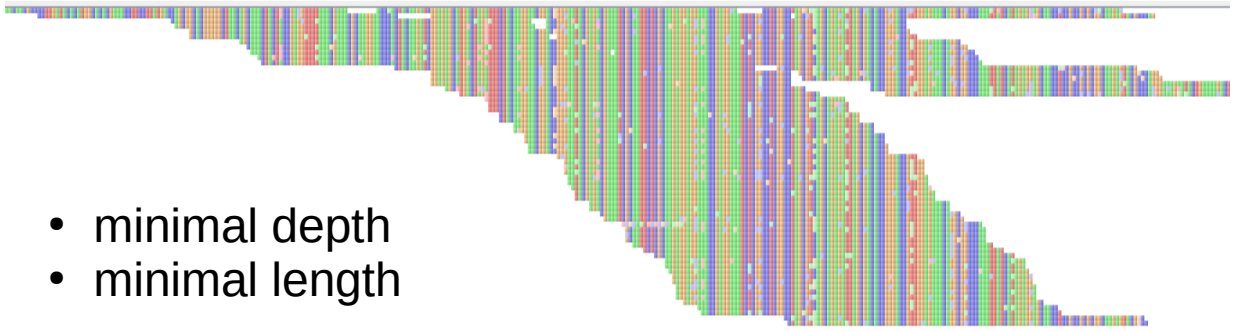
Characterization of repeat from low-pass shogun sequencing

RepeatExplorer tools for genome annotation



Library Based Repeat Annotation

Contigs from clustering results



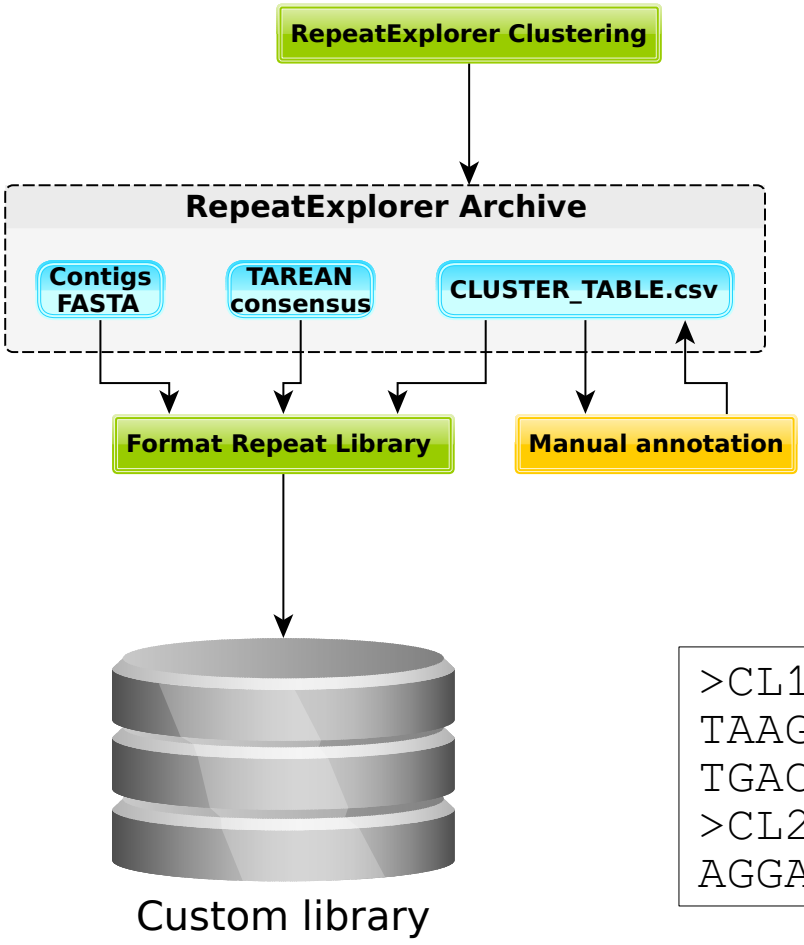
- minimal depth
- minimal length

- **Custom library**

TAREAN consensus

- as dimers

Library Based Repeat Annotation



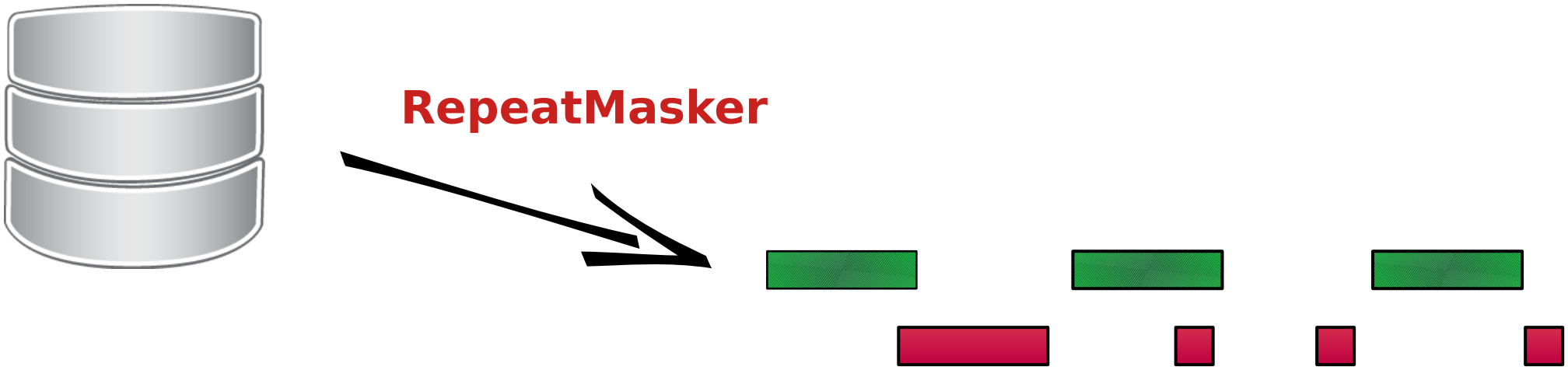
Library preparation

- 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
```

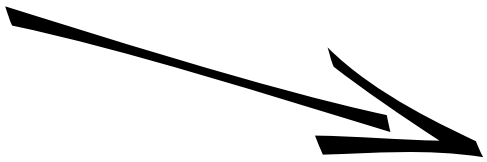
Library Based Repeat Annotation

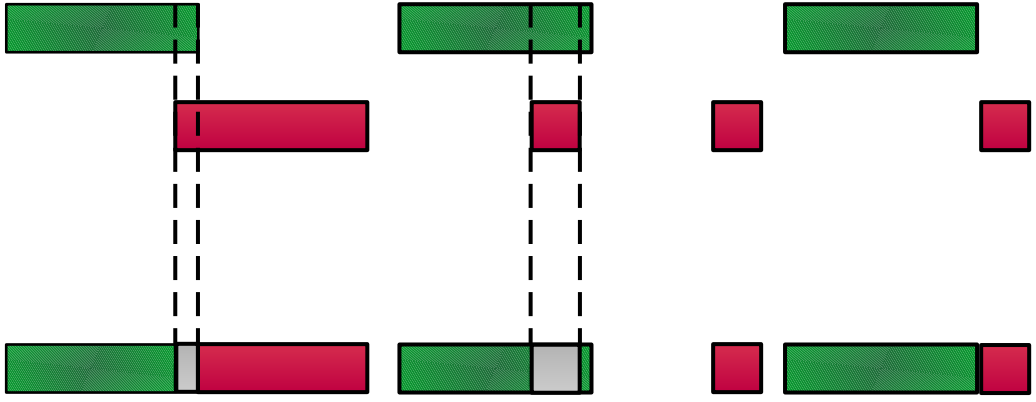
Library Based Repeat Annotation



Library Based Repeat Annotation

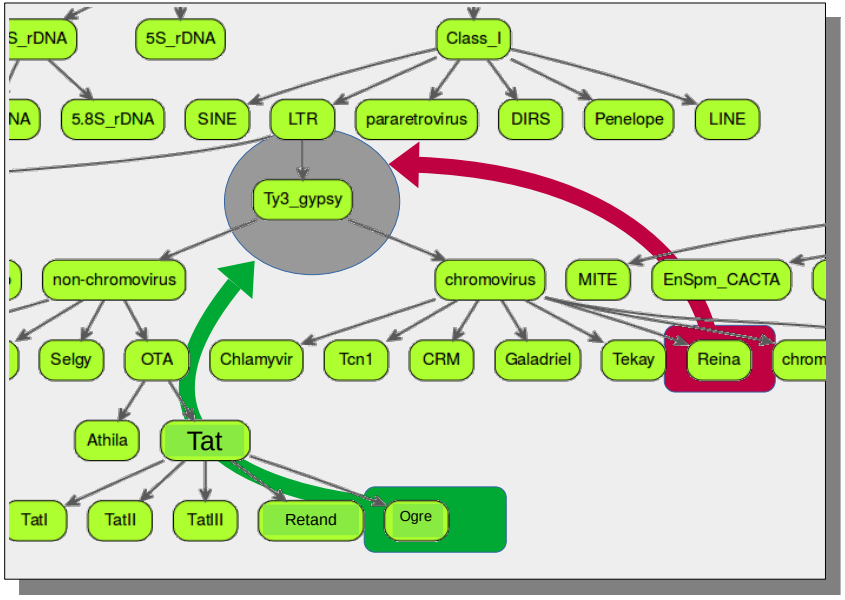
RepeatMasker

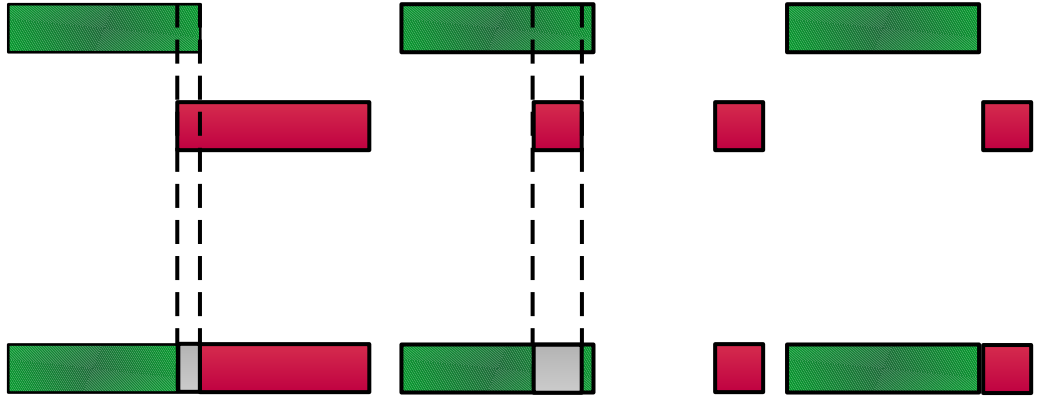




Library Based Repeat Annotation

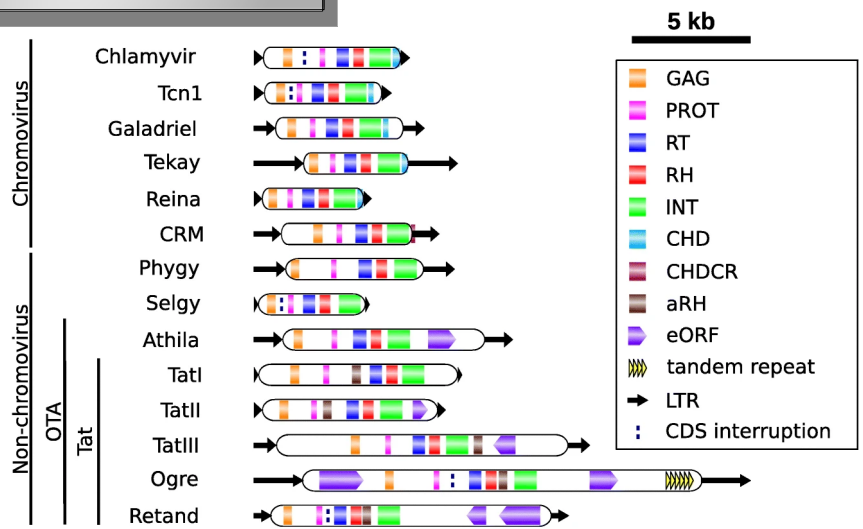
RepeatMasker

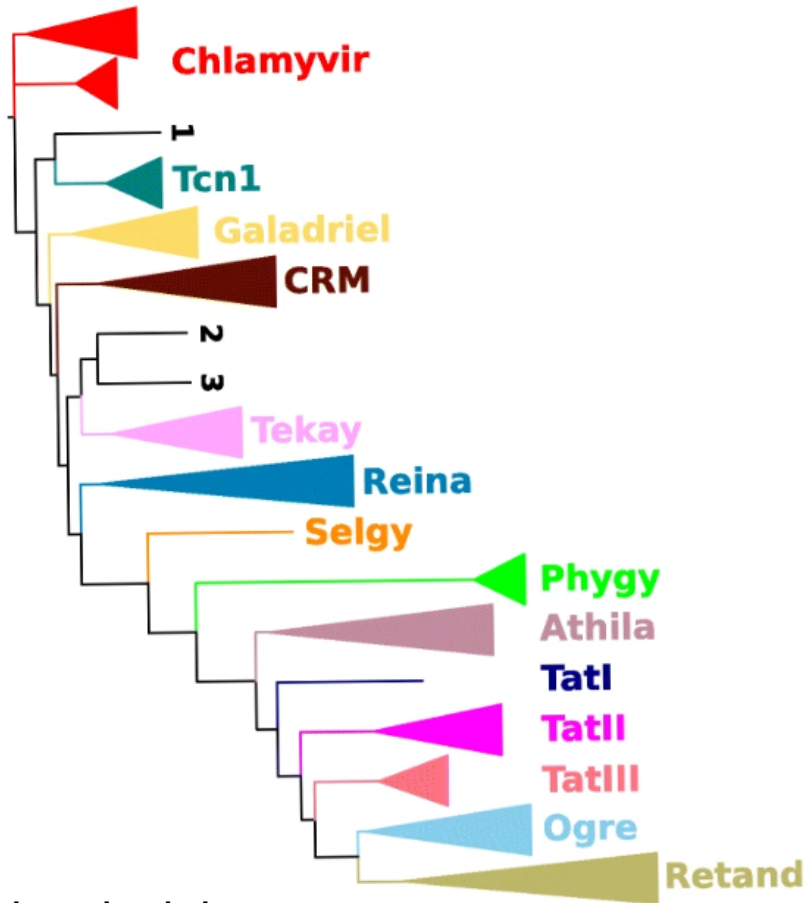




- GFF3 output
- Hierarchical user provided classification
- Conflicts in annotation → LCA

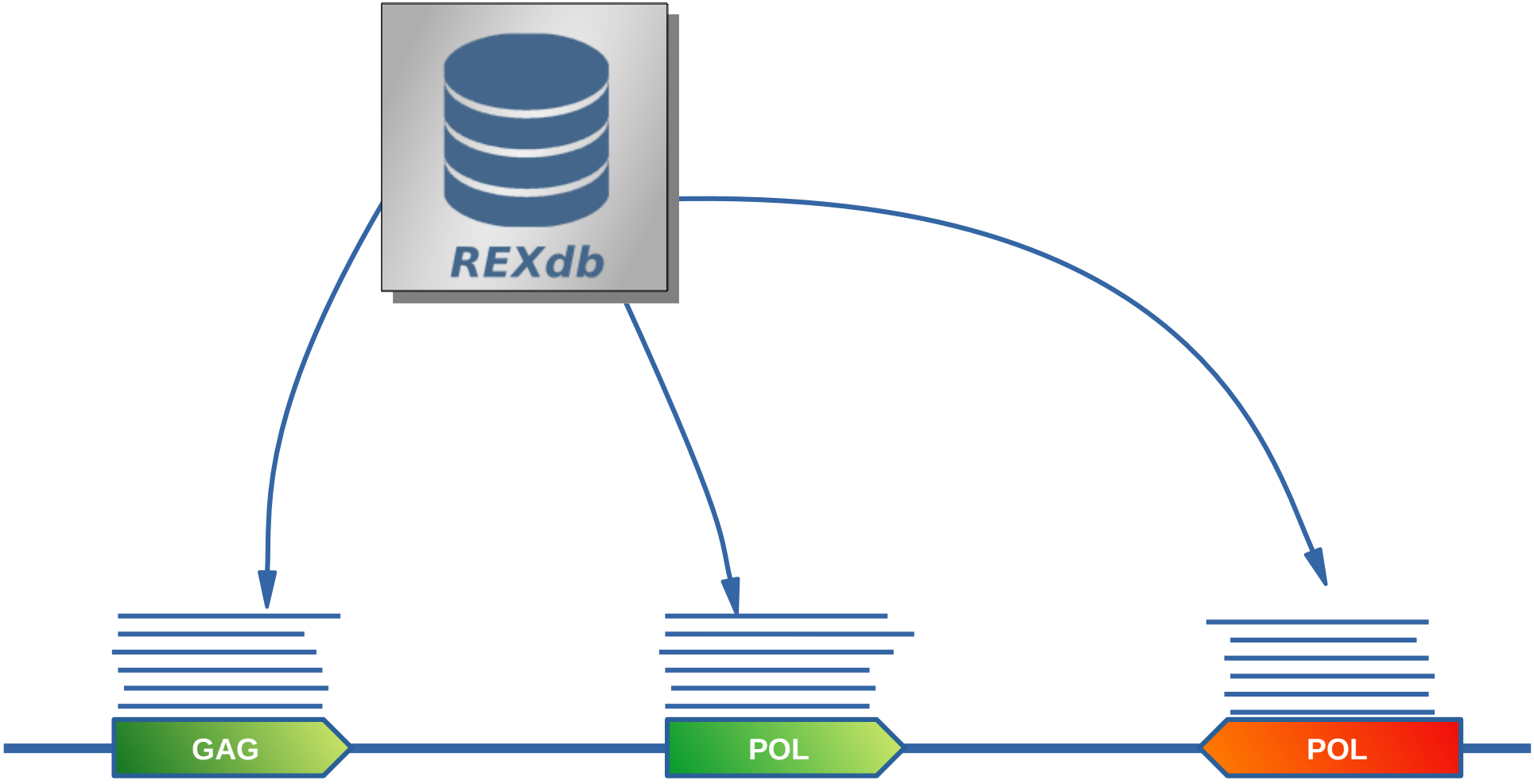
DANTE – domain based annotation of transposable elements



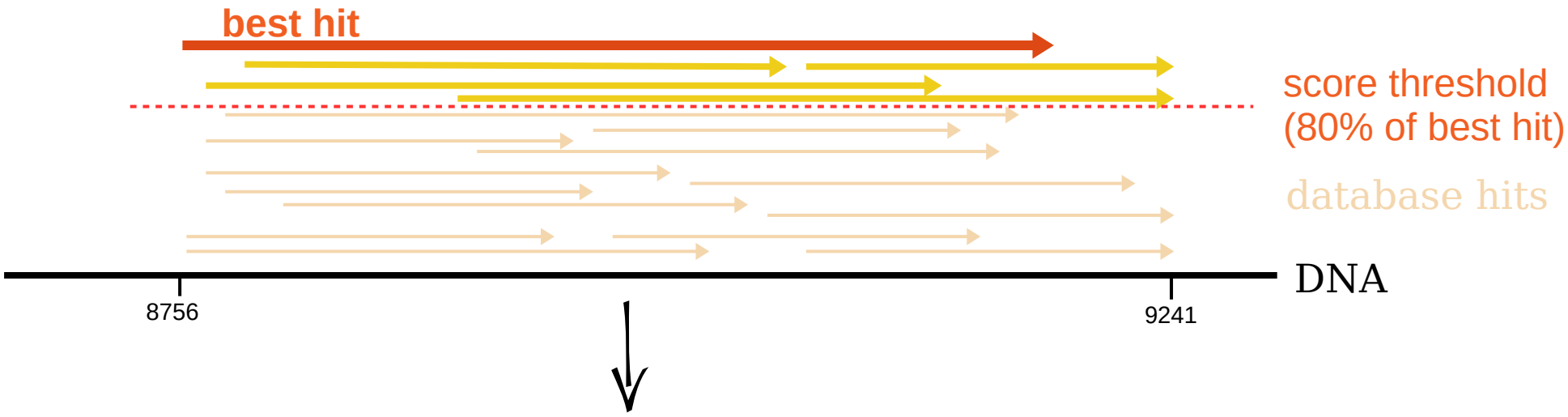


Annotation and classification based on phylogenetic principle

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

Structure Based Annotation of LTR retrotransposons

- **LTR_STRUCT (2003)**
- **LTR_FINDER (2007)**
- **LTRharvest (2008)**
- **LTRdigest (2009)**
- **LTR_detector (2019)**

Principle of detection:

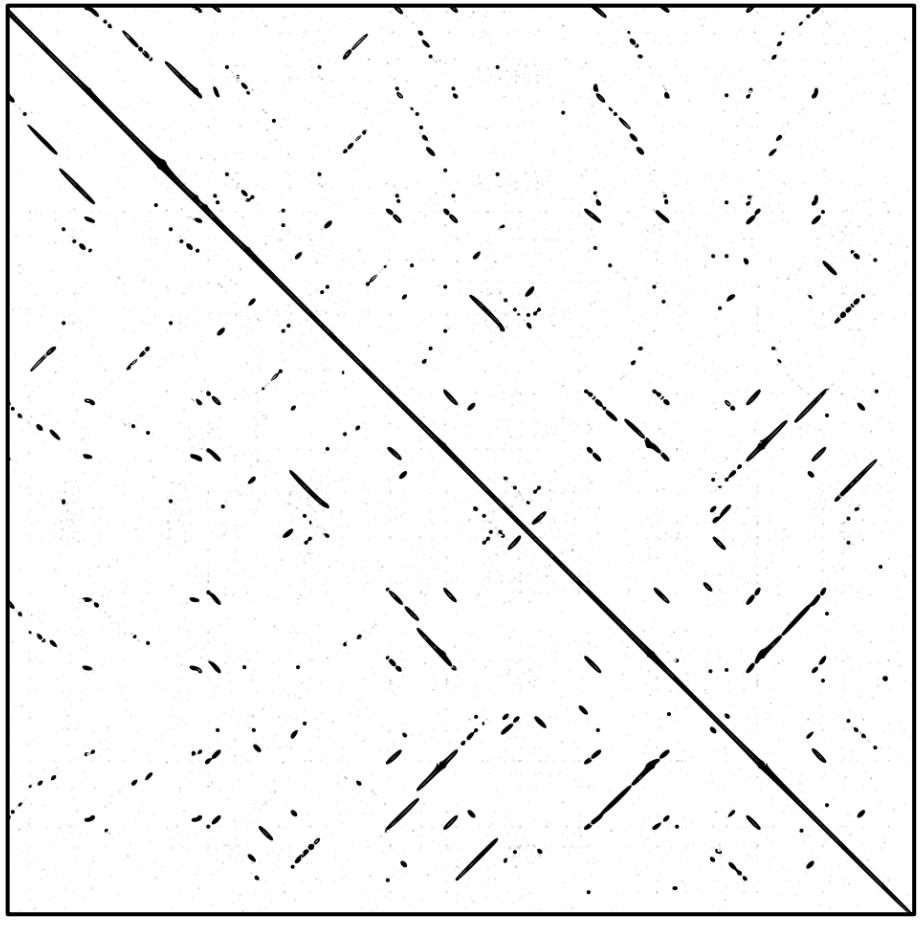
1) LTR detection (SW, k-mer, suffix array)
2) TSD, PBS, PPT detection
3) ORF or protein domains – limited database (uniprot

Meta tools:

- **LTRpred (2020)**
  - LTRharvest
  - LTRdigest
- **LTR_retriever (2018)**
  - LTR_STRUCT
  - LTR_FINDER
  - LTRharvest
  - MGESacn3
  - LTR_detector
- **Inpactor2 (2022)**
  - CNN
  - LTR_FINDER
  - Lineage base classification
  - InpactorDB

Structure Based Annotation of LTR retrotransposons

- LTR_STRUCT (2003)
- LTR_FINDER (2007)
- LTRharvest (2008)
- LTRdigest (2009)
- LTR_detector (2019)

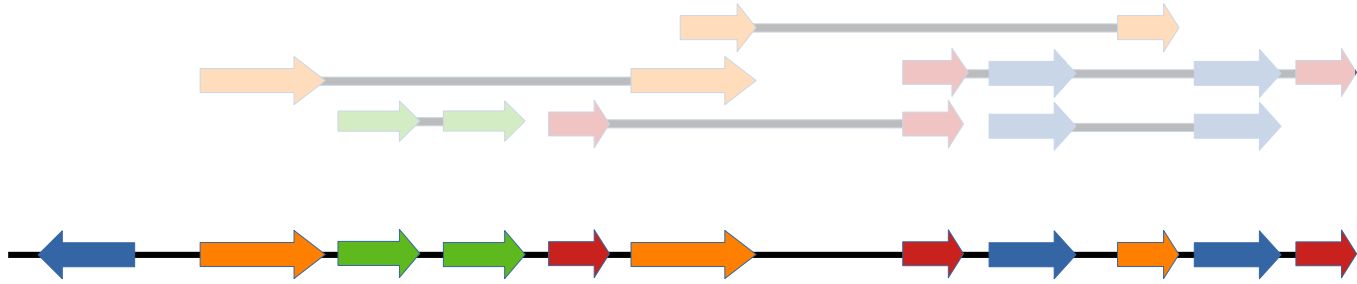


Principle of detection:

1) LTR detection (SW, k-mer, suffix array)
2) TSD, PBS, PPT detection
3) ORF or protein domains – limited database

Structure Based Annotation of LTR retrotransposons

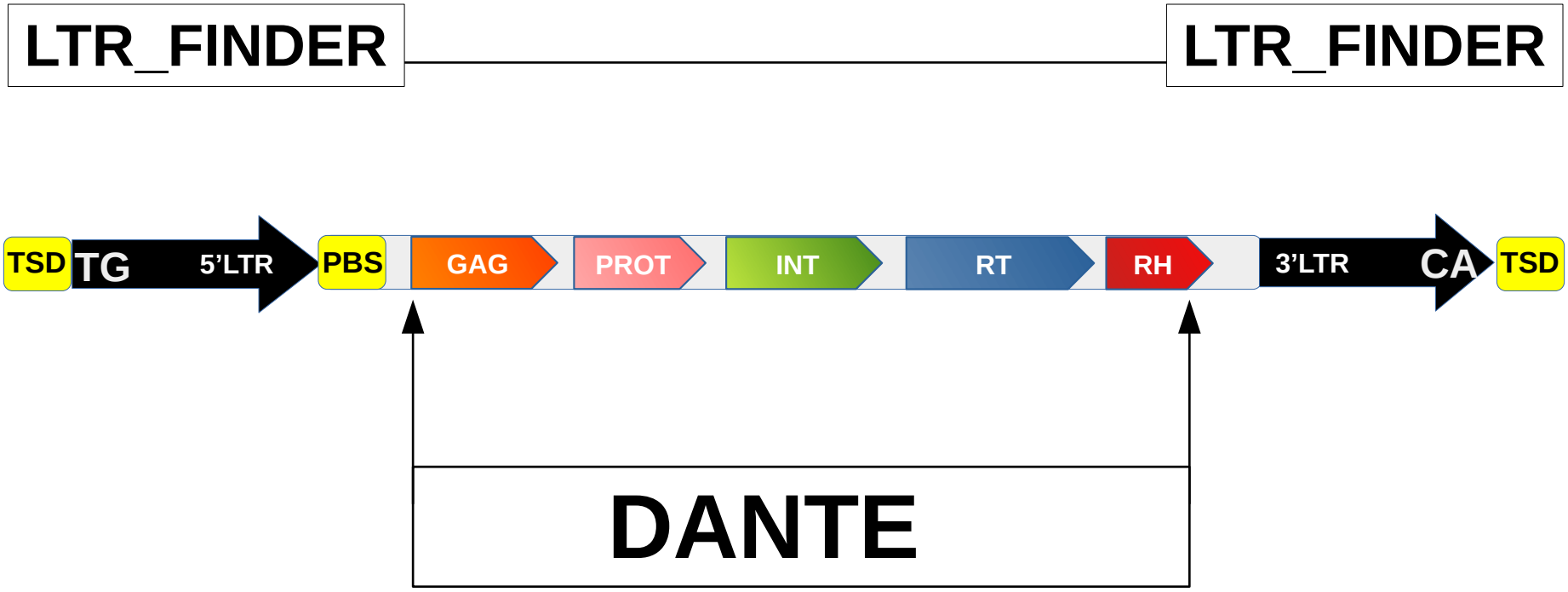
- LTR_STRUCT (2003)
- LTR_FINDER (2007)
- LTRharvest (2008)
- LTRdigest (2009)
- LTR_detector (2019)



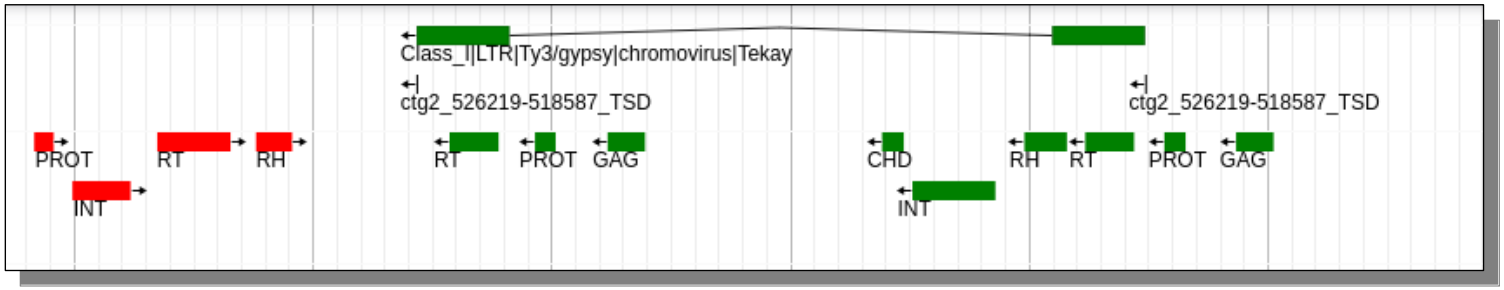
Principle of detection:

1) LTR detection (SW, k-mer, suffix array)
2) TSD, PBS, PPT detection
3) ORF or protein domains – limited database

DANTE LTR – structure based annotation

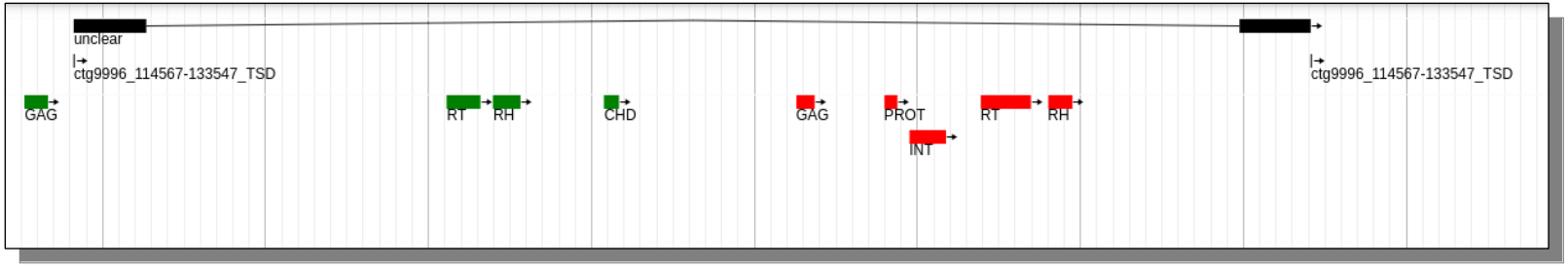


DANTE + LTR_FINDER



LTR_FINDER

DANTE



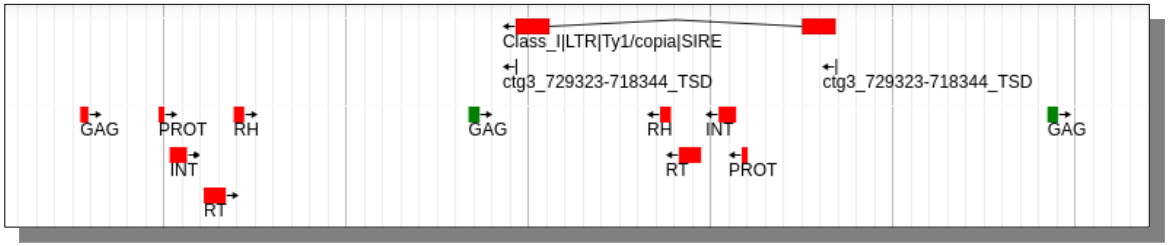
LTR_FINDER

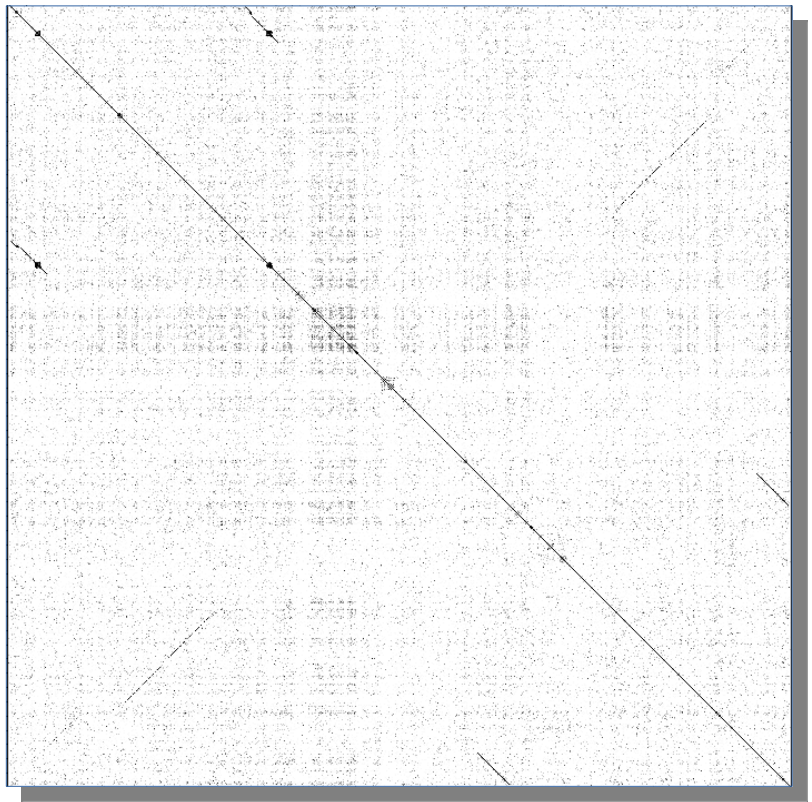
DANTE

- Ty3/gypsy
- Ty1/copia

Can we combine LTR_FINDER + DANTE to get better LTR RT annotation ?

DANTE + LTR FINDER





LTR_FINDER

DANTE

Limiting LTR_FINDER analysis to regions with conserved domains

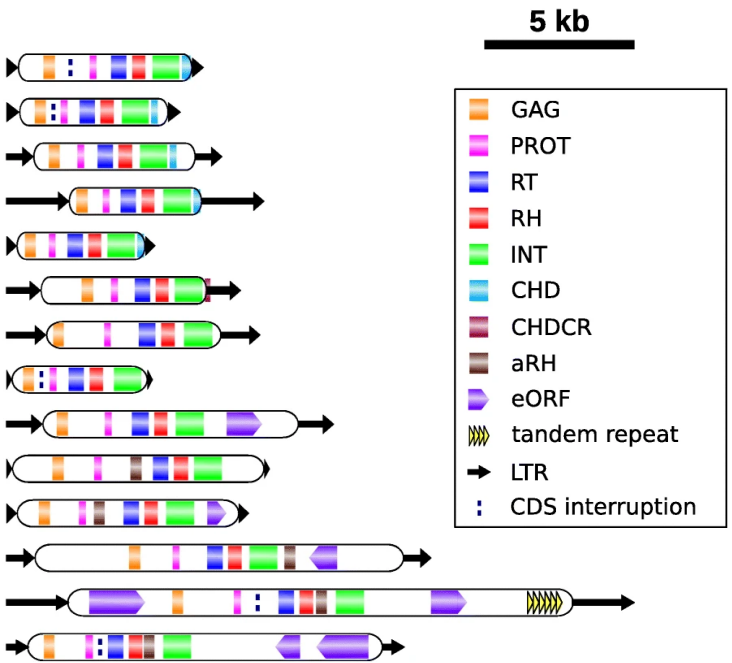
High proportion of false negatives

DANTE LTR – structure based annotation

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

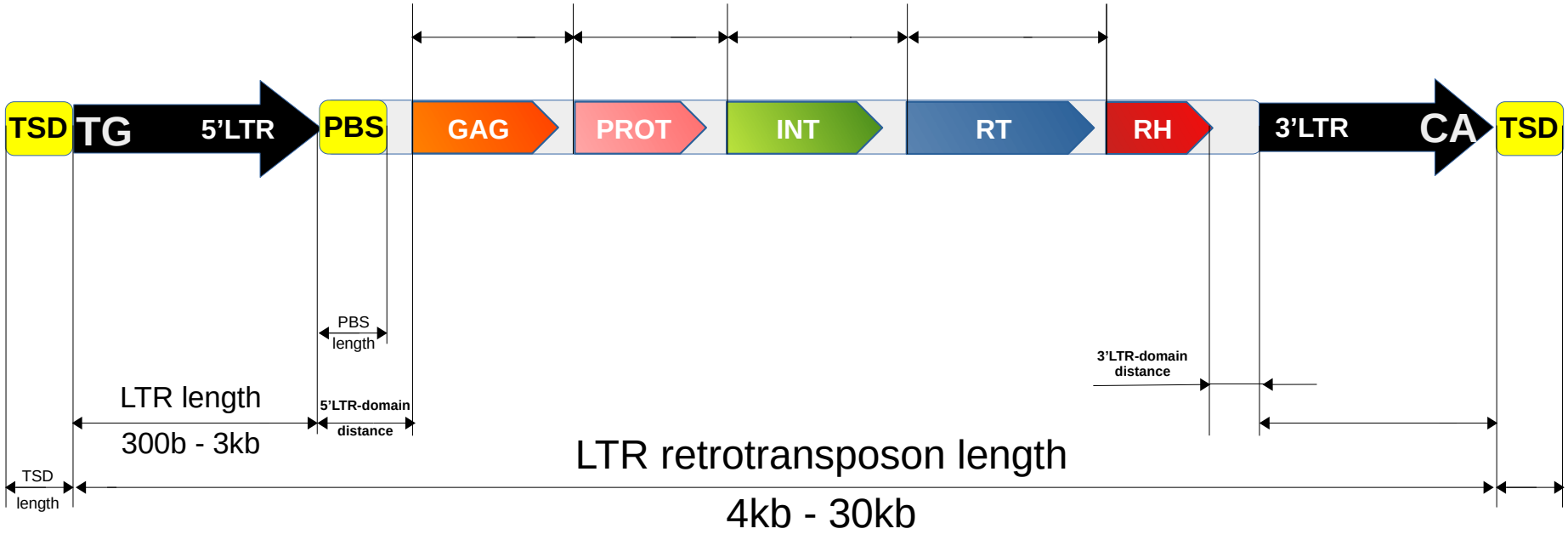
A Ty3/gypsy

		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

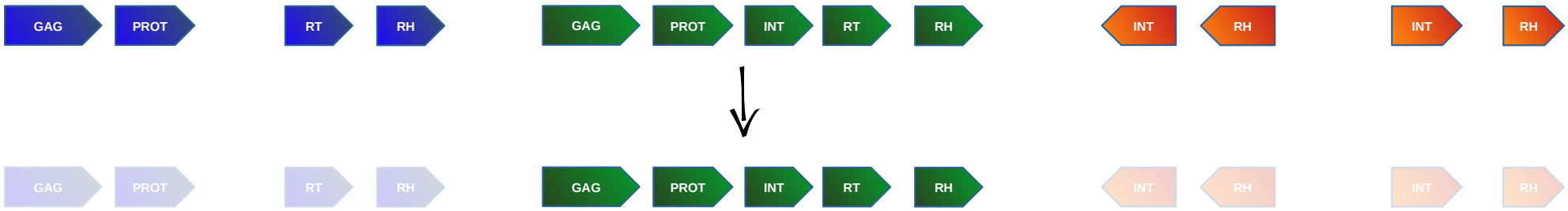


Search constraints:

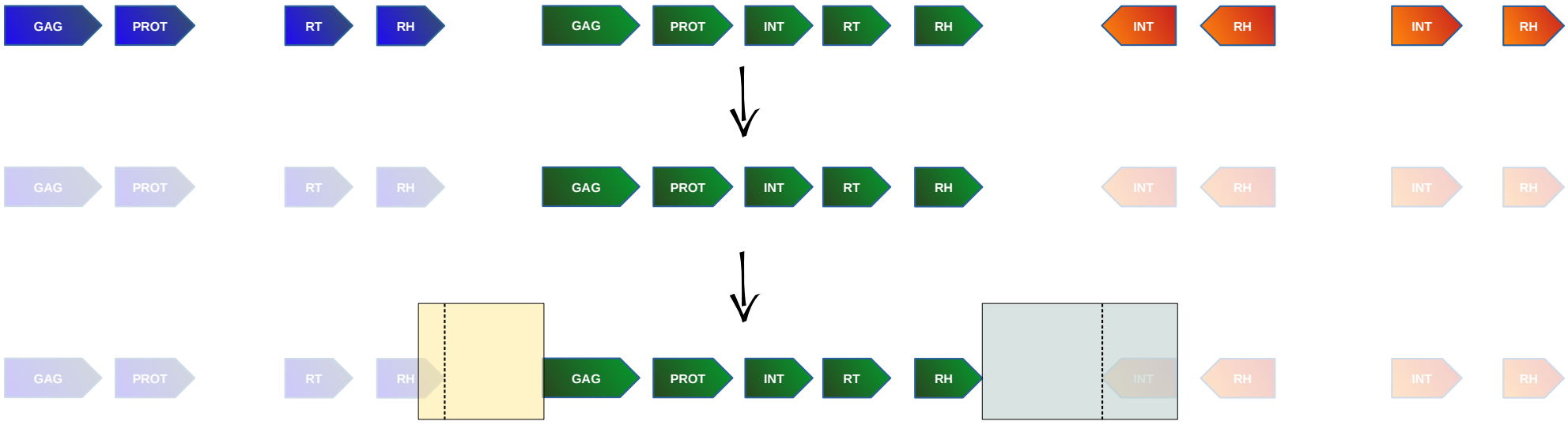
- Feature length (min, max)
- Presence of features (PBS, TSD, TG/CA, domains)
- Feature distances

DANTE LTR – WORKFLOW

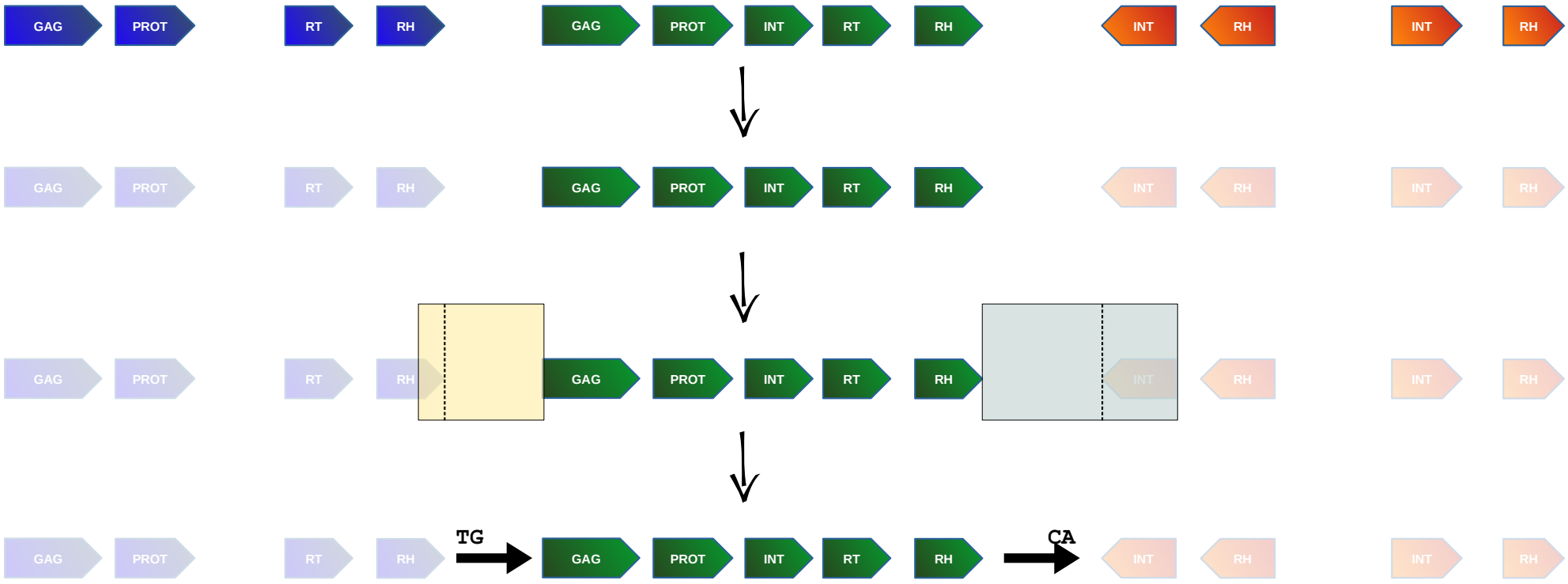
DANTE LTR – WORKFLOW



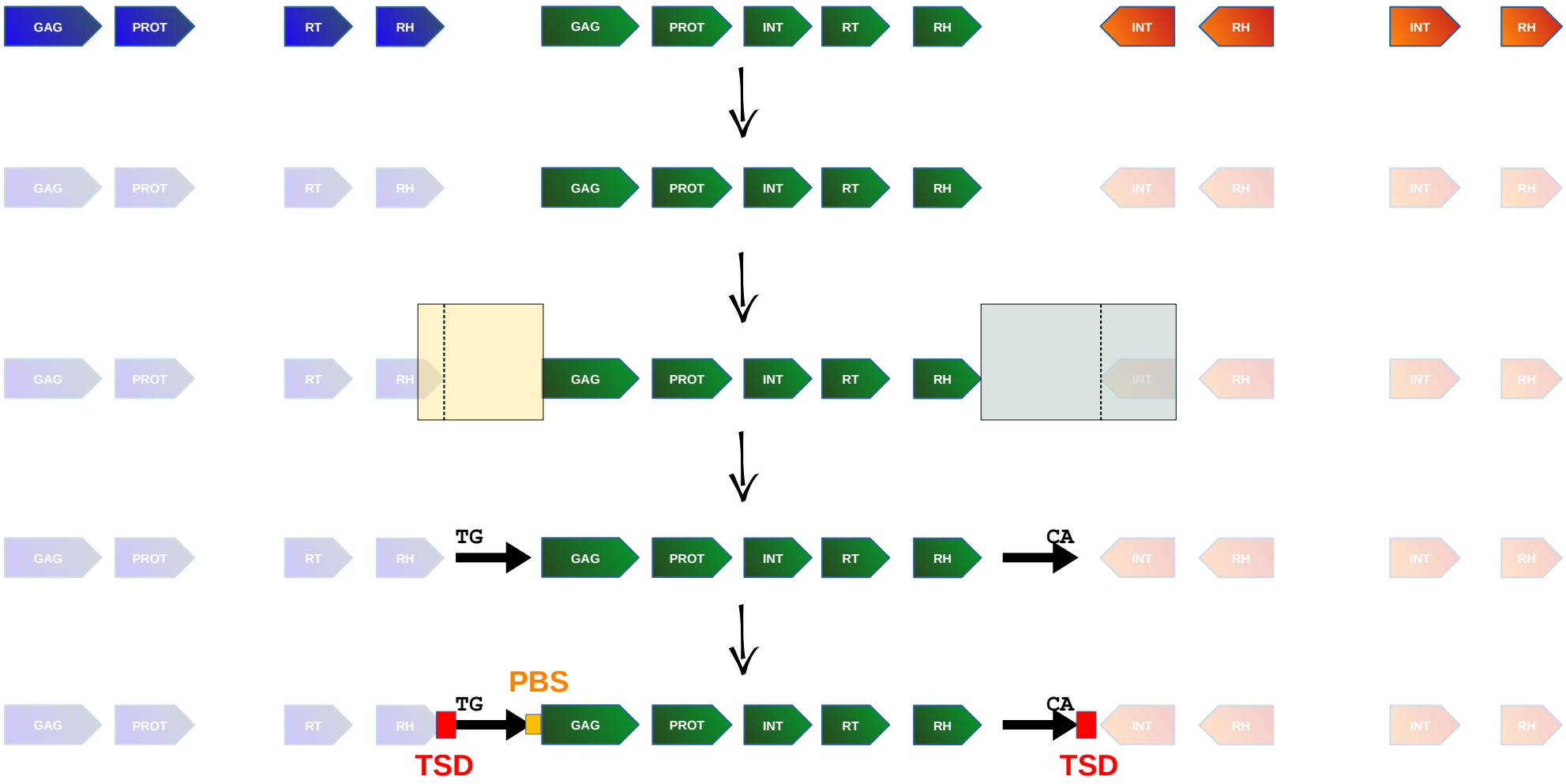
DANTE LTR – WORKFLOW



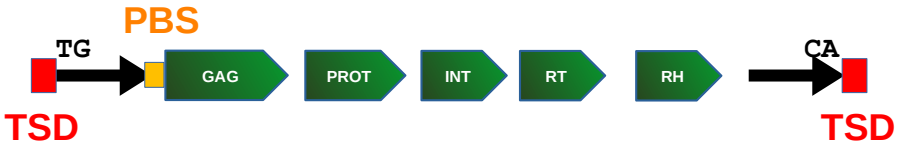
DANTE LTR – WORKFLOW



DANTE LTR – WORKFLOW



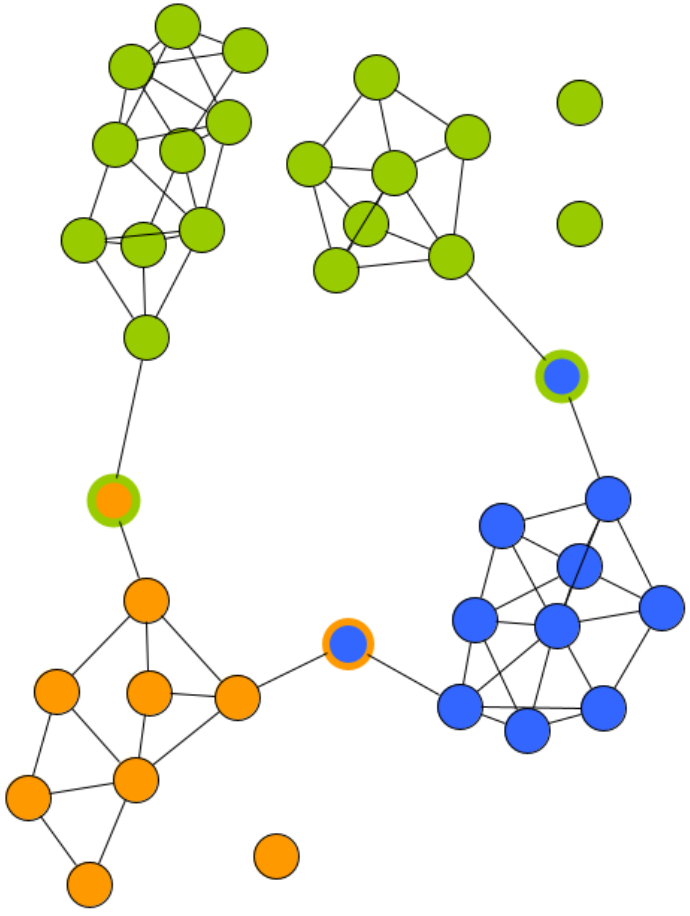
DANTE LTR – Retrotransposons ranks



Rank	Annotation
DLTP	Elements with identified protein domains, LTRs, TSD and PBS
DLP	Elements with identified protein domains, LTRs and **P**BS (TSD not found)
DLT	Elements with identified protein domains, LTR, LTRs and TSD (PBS not found)
DL	Elements with protein domains, LTRs (PBS and LDS were not found)
D	Elements with protein domains from the same lineage (LTR not found)

DANTE LTR – Filtering

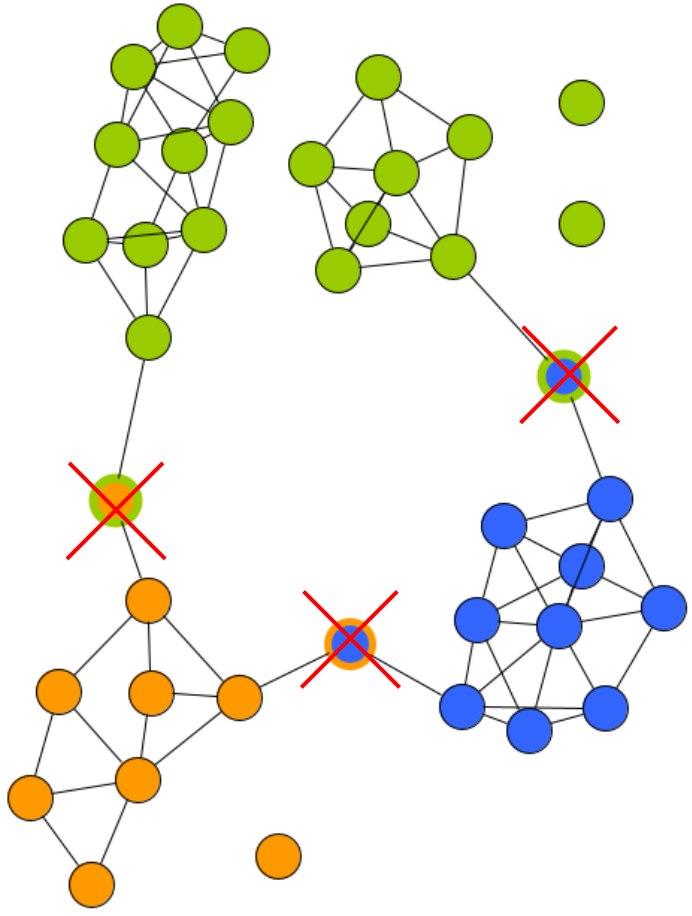
All-to-All sequence comparison



DANTE LTR – Filtering

All-to-All sequence comparison

Cross-lineage similarities

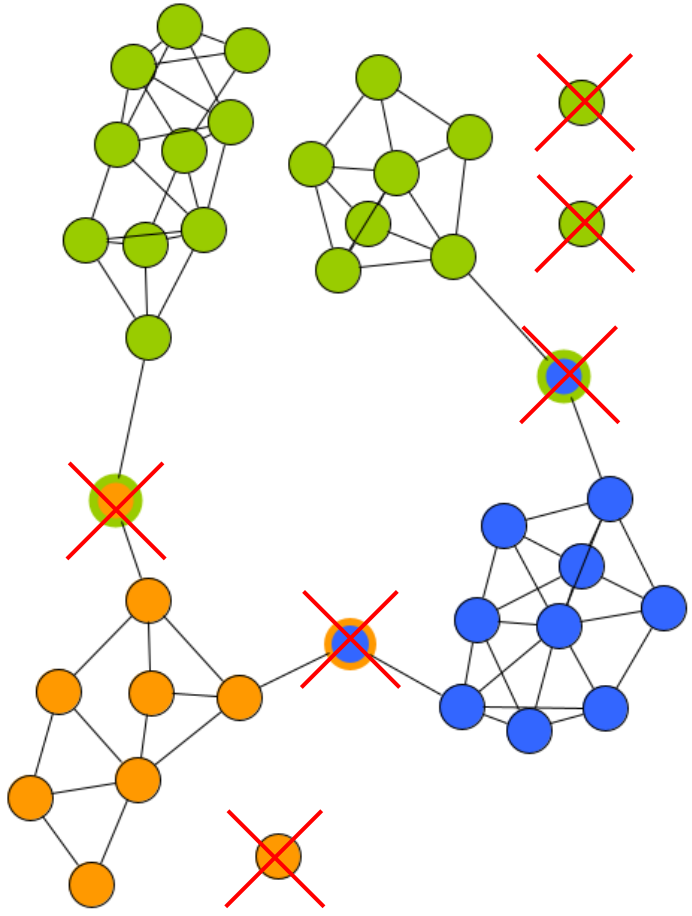


DANTE LTR – Filtering

All-to-All sequence comparison

Cross-lineage similarities

Minimal copy number (DLP, DLT)

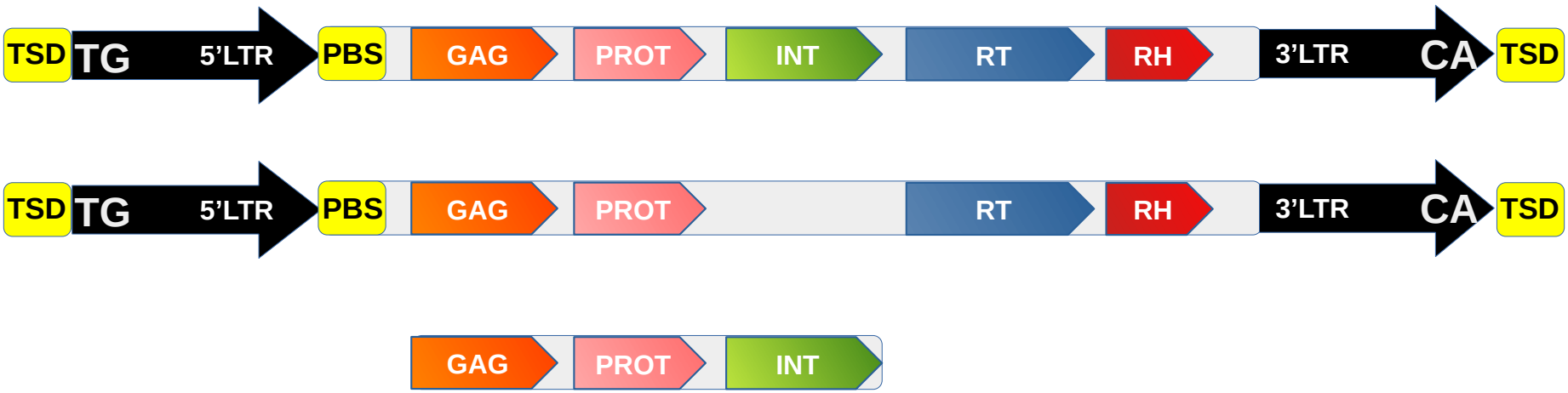


DANTE LTR - Output

	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 output

- **Rank=DLTP** (DL, DLT, DLP)
- **Ndomains=5;**
- **ID=TE_00000034_CEN6_ver_220406;**

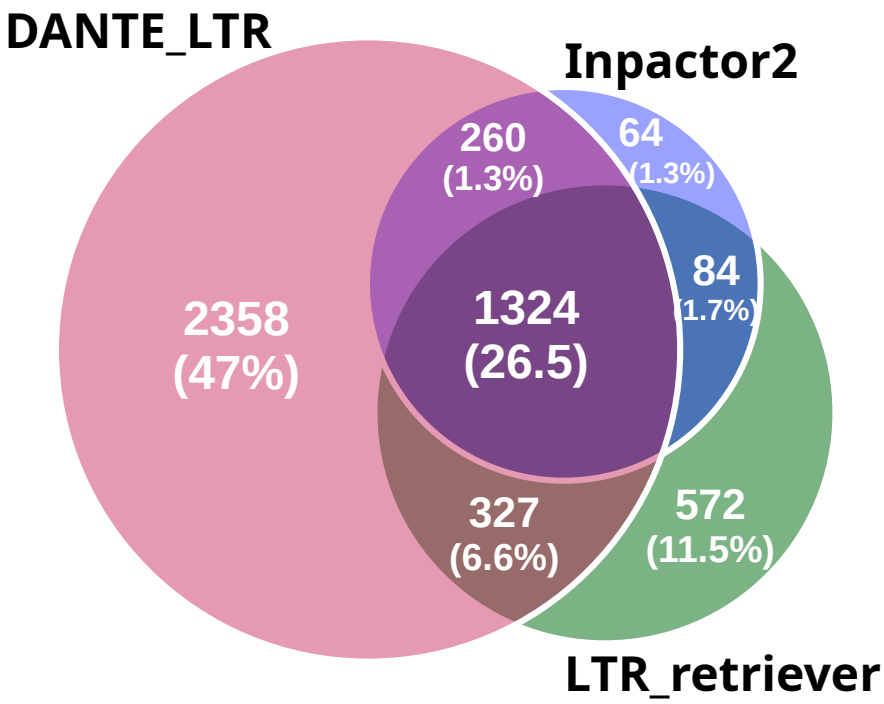


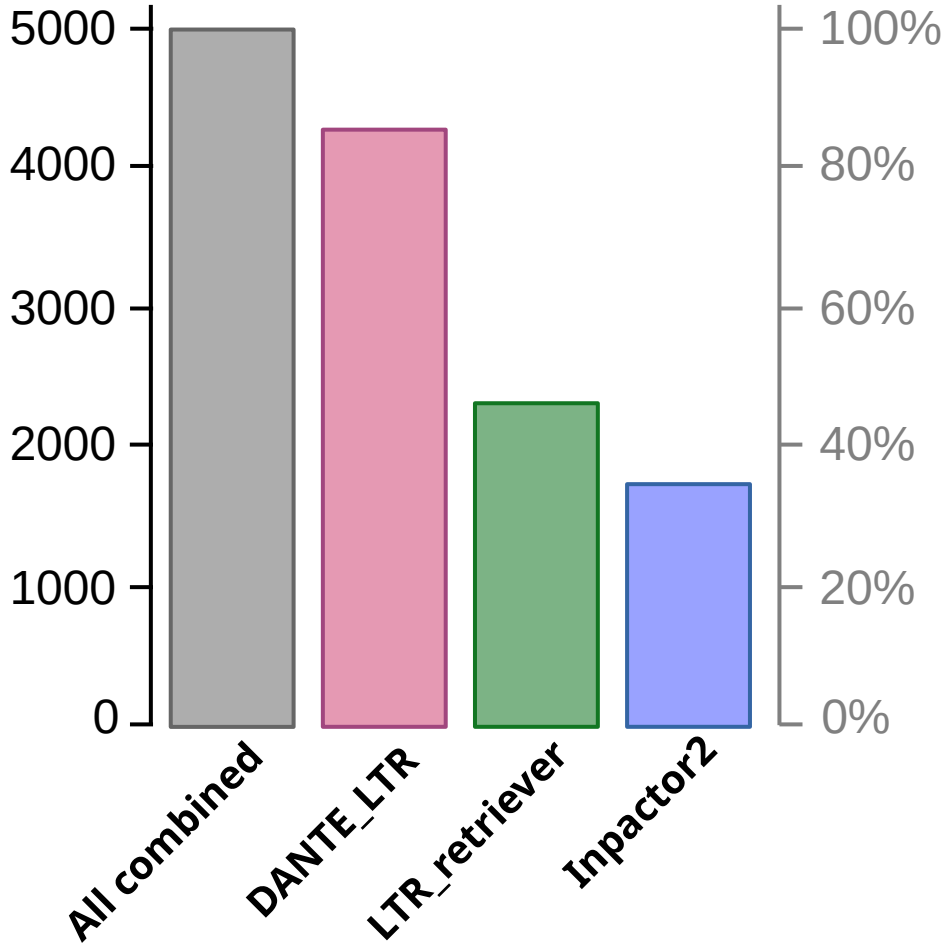
TE_partial_0003_CEN6_ver_220406

DANTE_LTR output

- **Rank=DLTP** (DL, DLT, DLP)
- **Ndomains=5;**
- **ID=TE_00000034_CEN6_ver_220406;** TE_partial_0003_CEN6_ver_220406
- **LTR_Identity=93.186;**
- **LTR5_length=223;**
- **LTR3_length=224;**
- **TSD=ATGAT/ATTAT;**
- **Final_Classification=Class_I|LTR|Ty1/copia|Ivana;**
- **Name=Class_I|LTR|Ty1/copia|Ivana;**
- **tRNA_id=TTCGAATCCTGCCCTGGATAcca__Met-1x;**
- **PBS_evalue=0.06;**

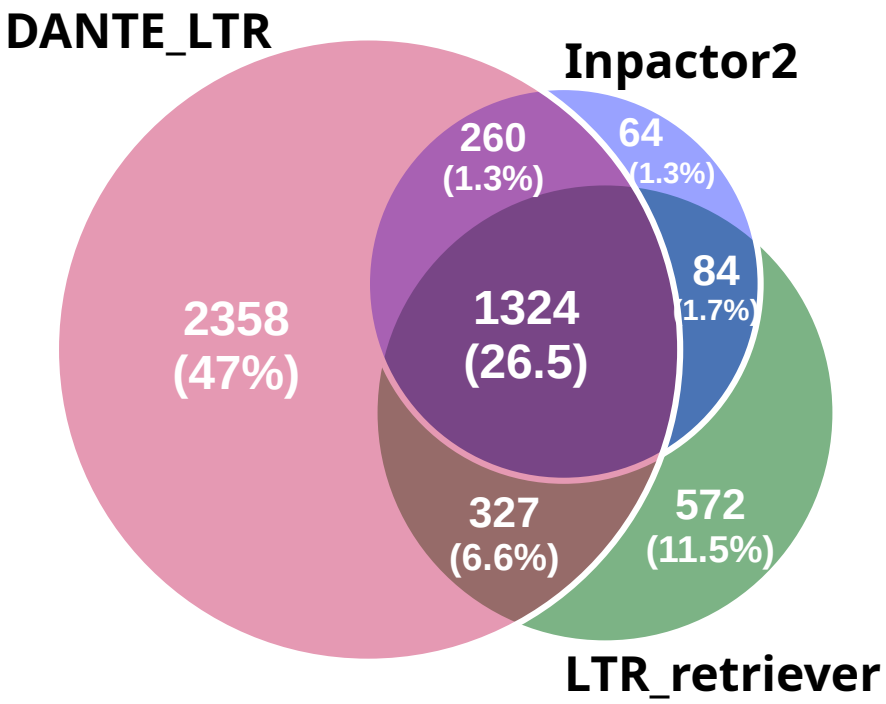
Comparison of Tools of LTR RT Detection

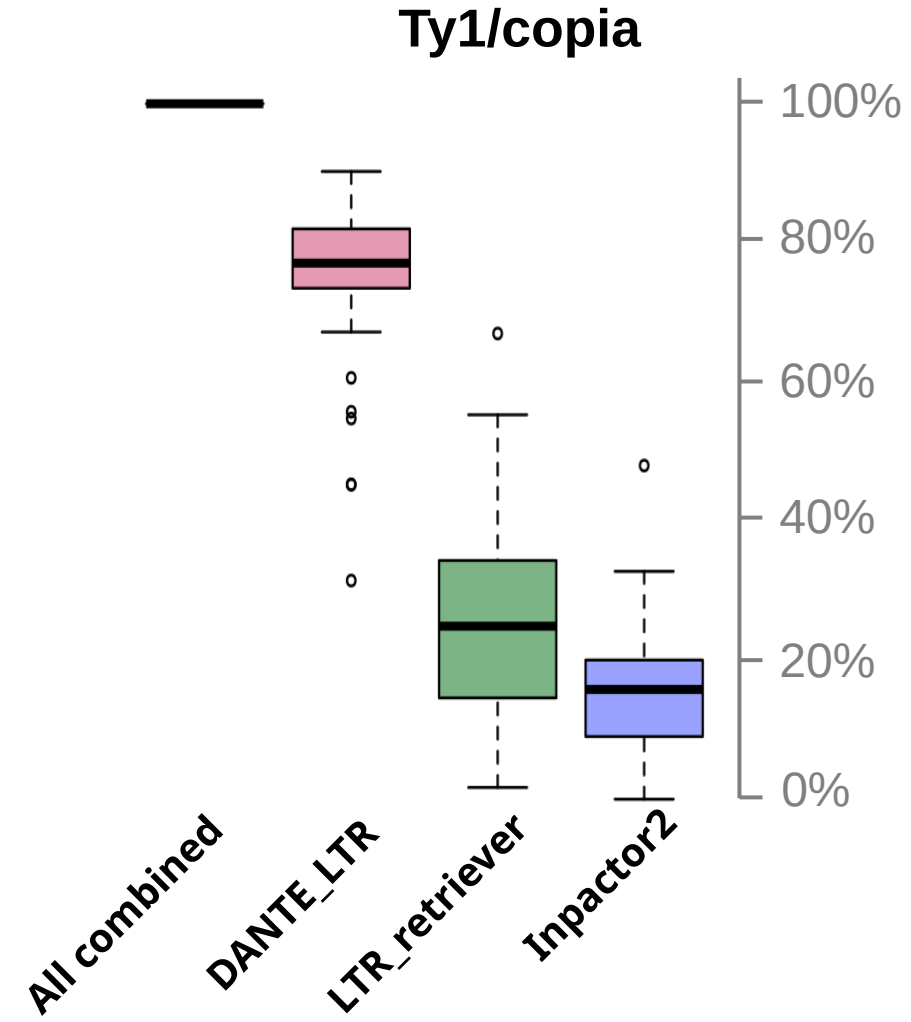




Glycine max – Ty1/copia detection

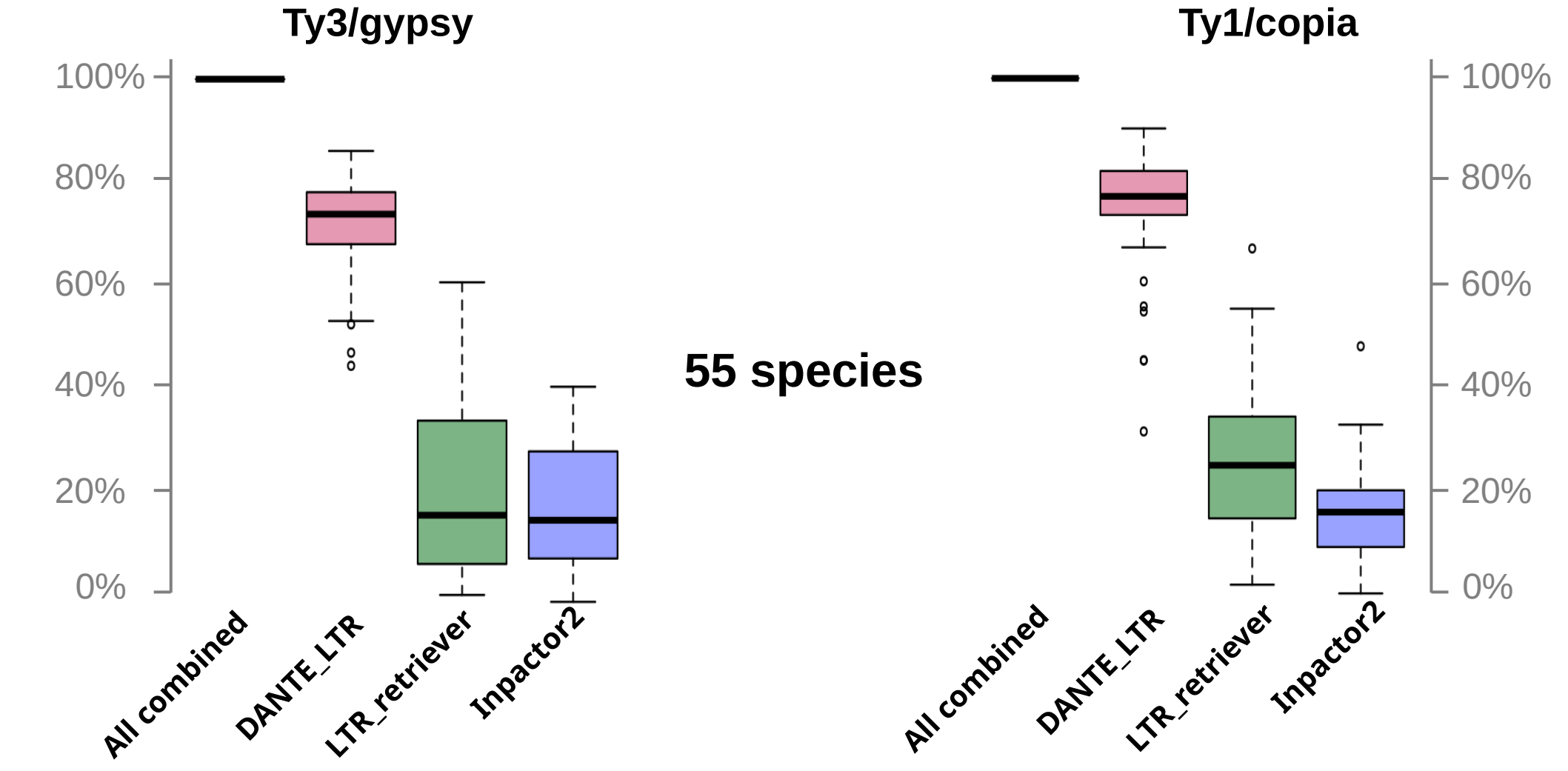
Comparison of Tools of LTR RT Detection



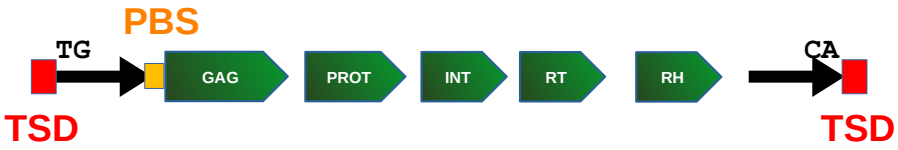


Glycine max – Ty1/copia detection

Comparison of Tools of LTR RT Detection

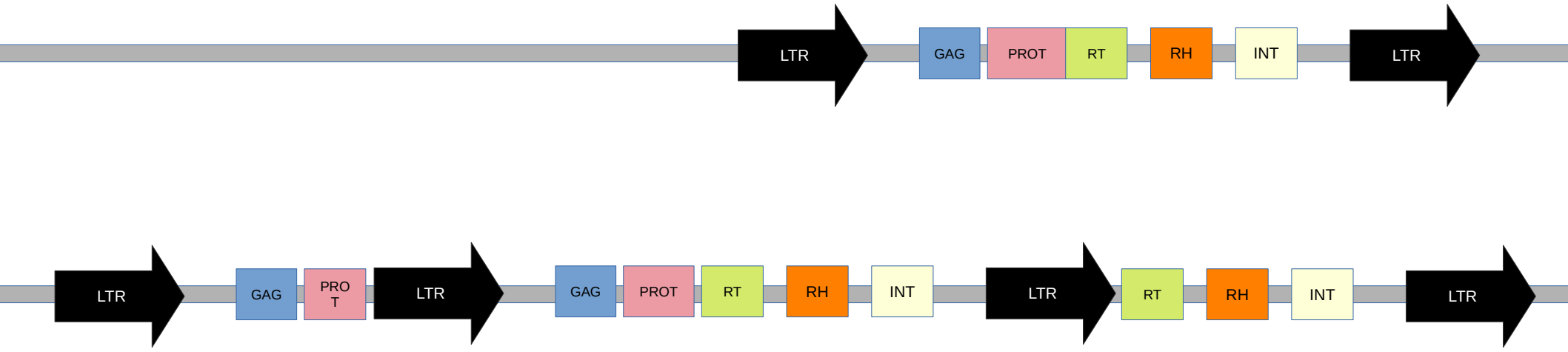


DANTE LTR – Applications



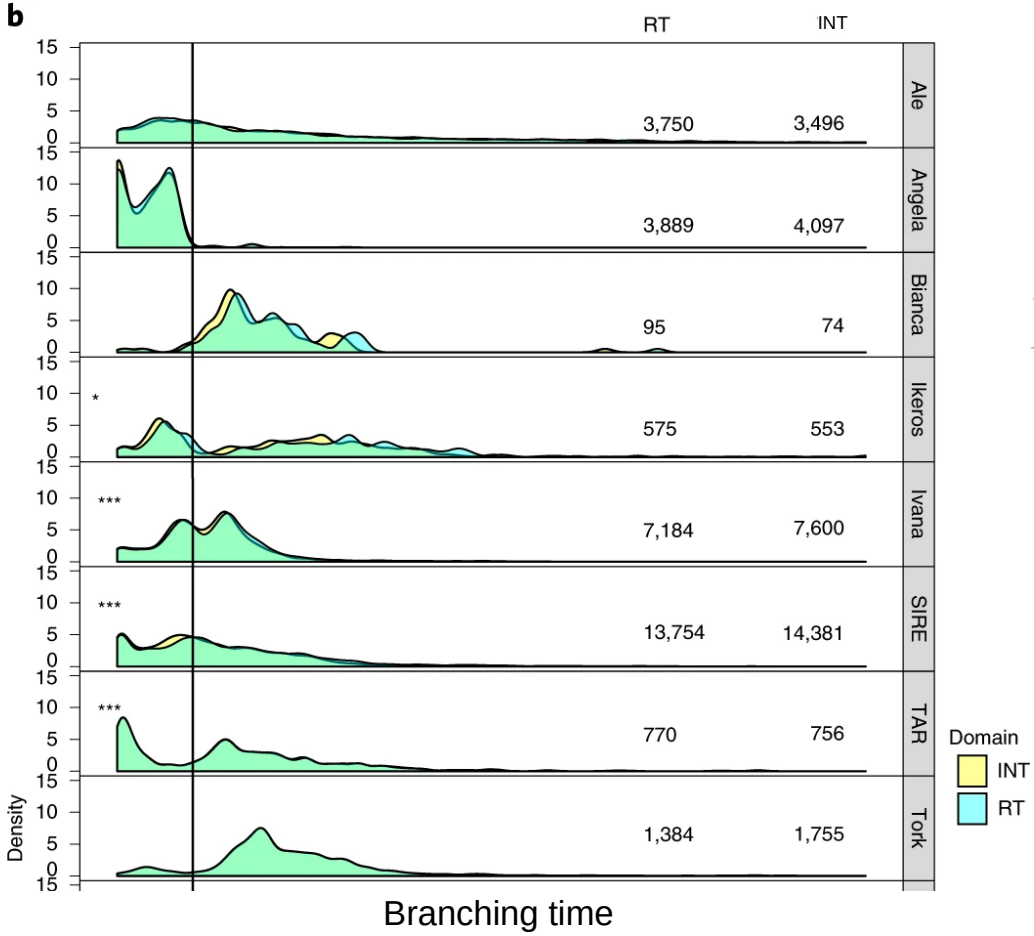
- Custom library for assembly annotation
- Dating of retrotransposons activity ???

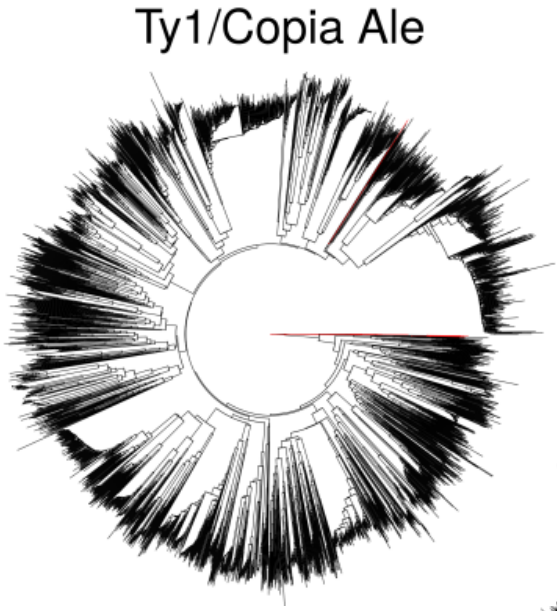
Dating of retrotransposons activity

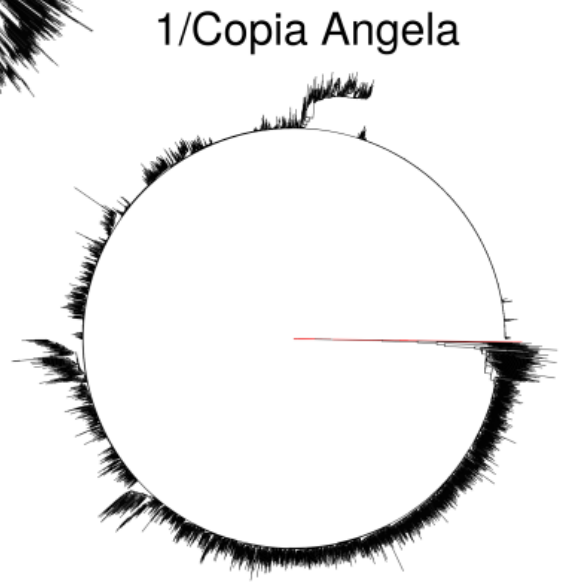


Nested TE insertion

Dating of retrotransposons activity



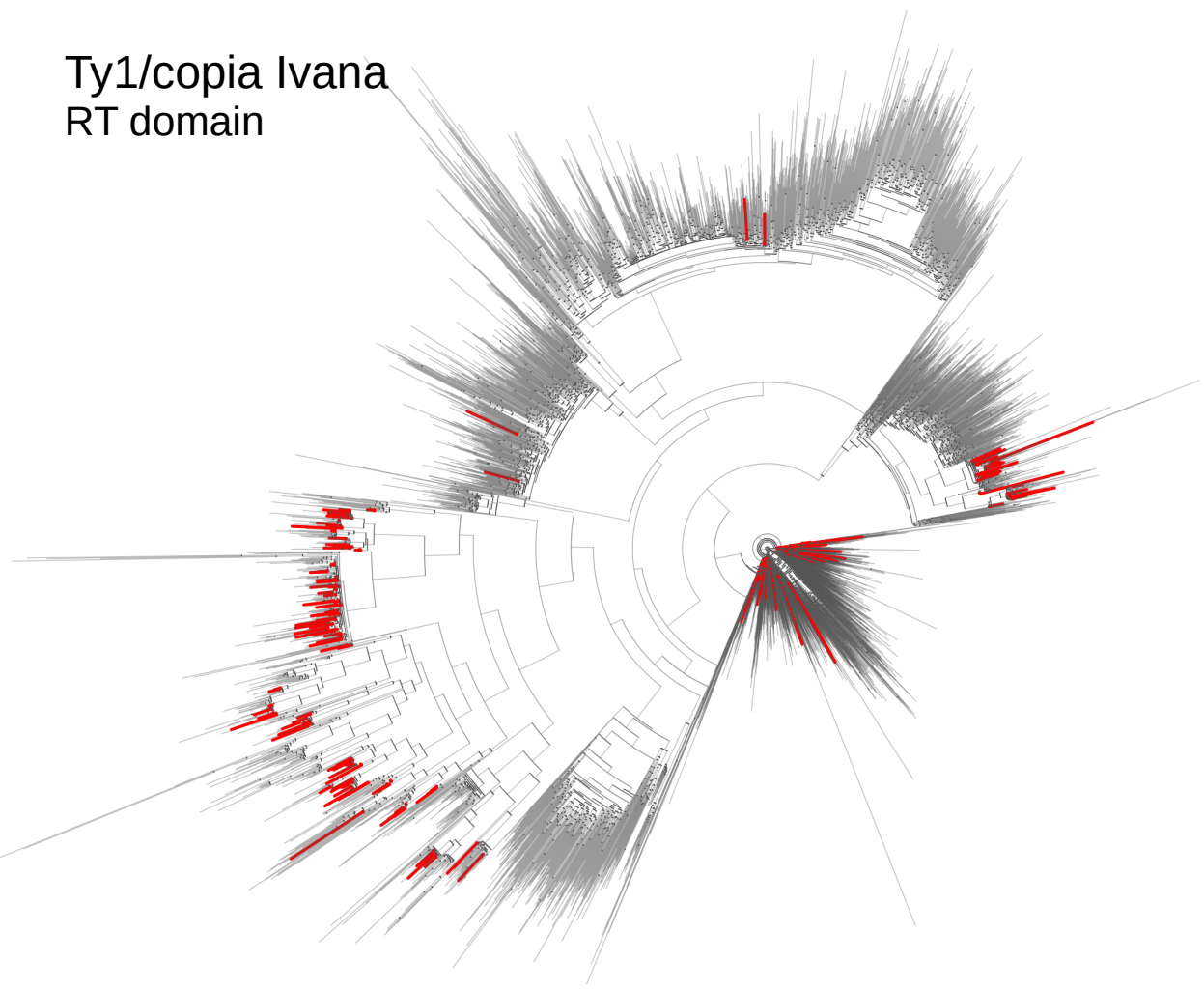


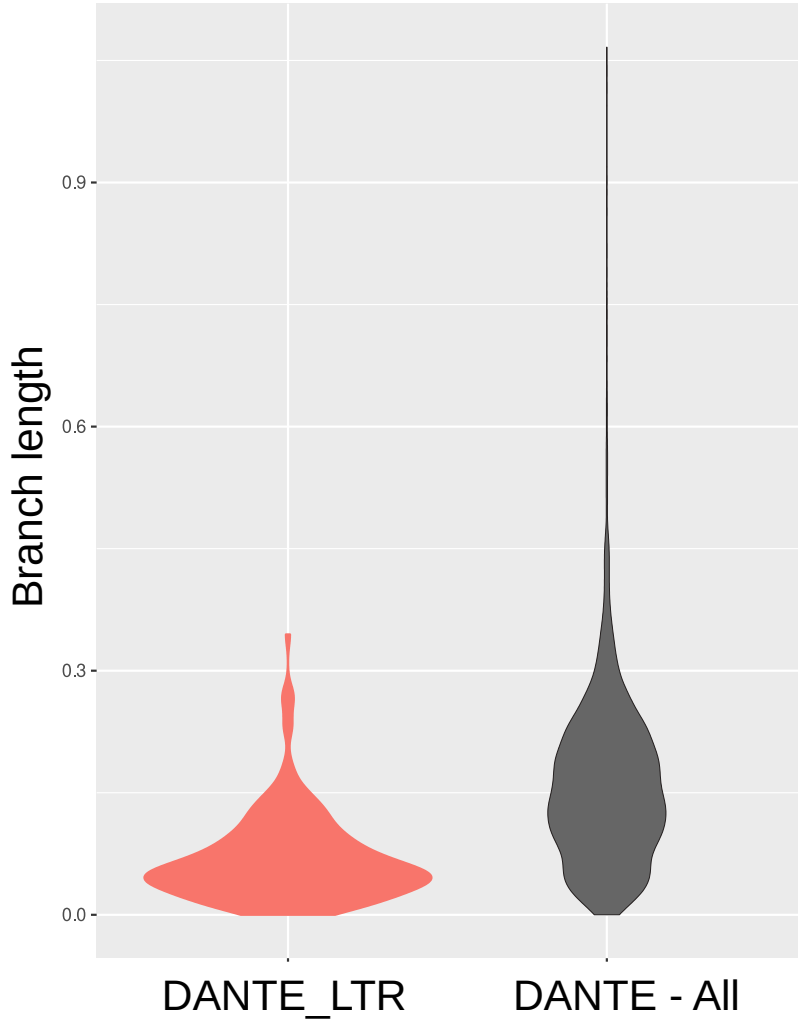


Kreplak, J., Madoui, MA., Cápal, P. et al. A reference genome for pea provides insight into legume genome evolution. Nat Genet 51, 1411–1422 (2019). https://doi.org/10.1038/s41588-019-0480-1

DANTE LTR

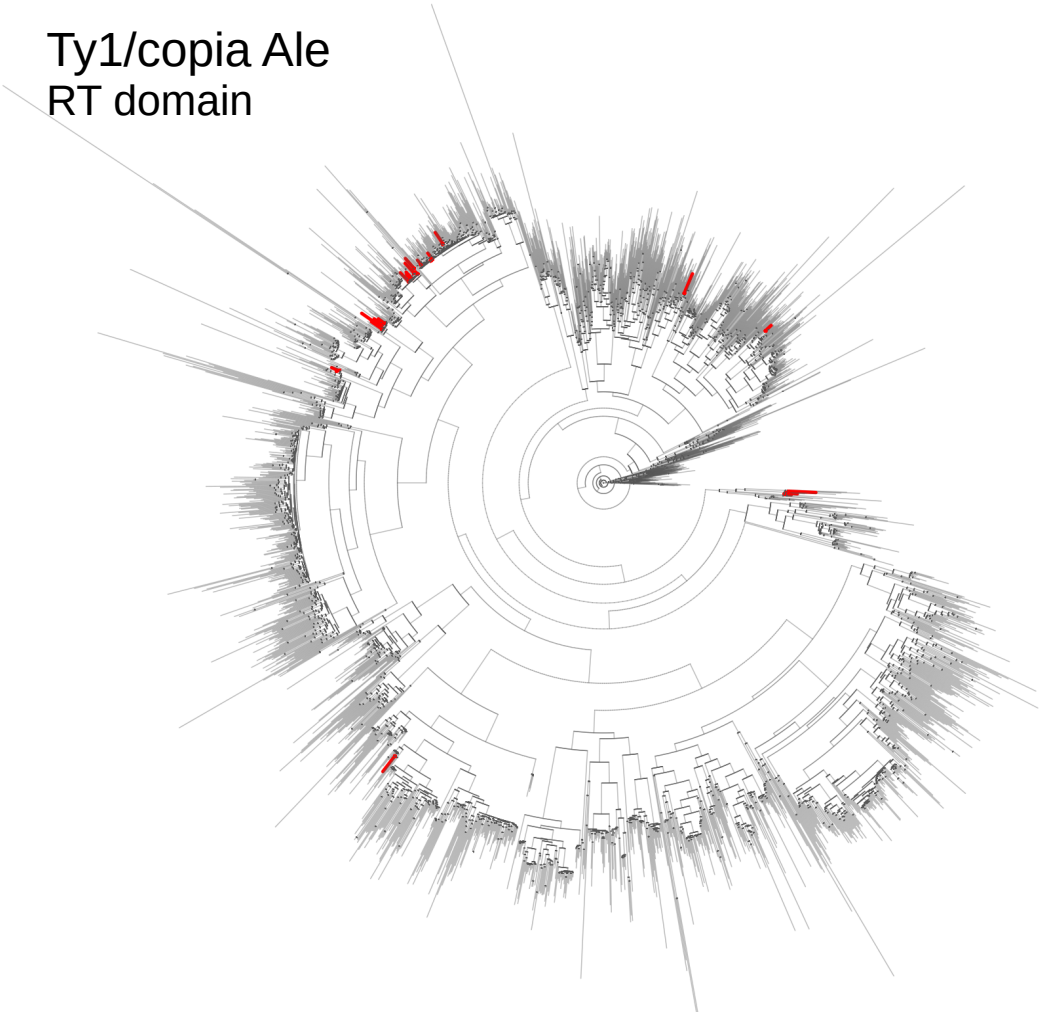
Ty1/copia Ivana
RT domain

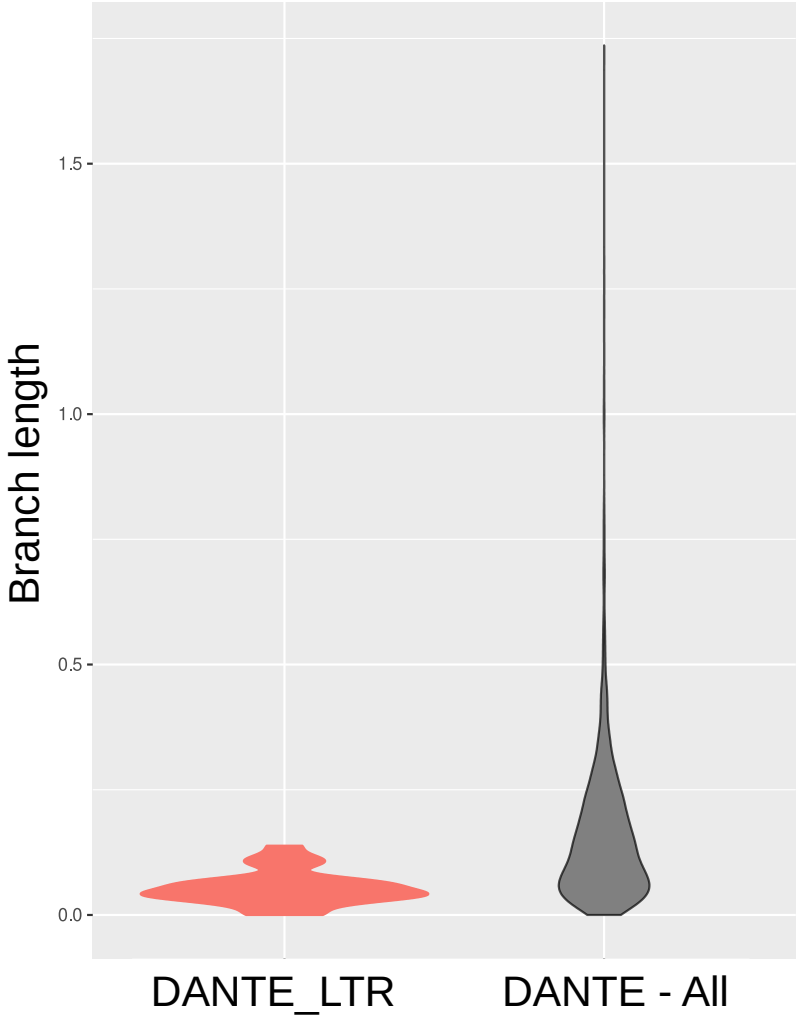




DANTE LTR

Ty1/copia Ale
RT domain





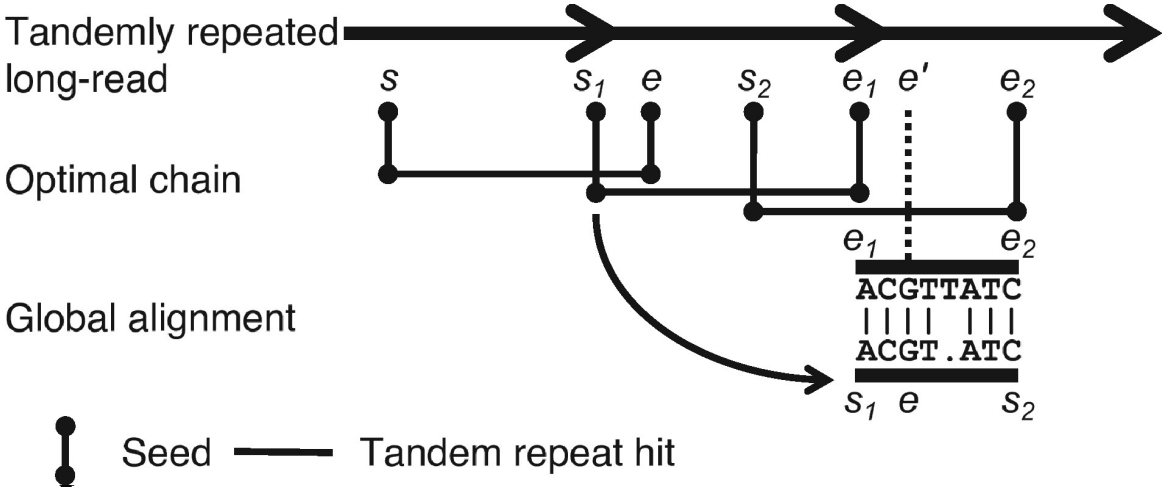
Tandem Repeat Annotation

Bioinformatics, 35, 2019, i200–i207
doi: 10.1093/bioinformatics/btz376
ISMB/ECCB 2019

OXFORD

TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain

Yan Gao[1,2], Bo Liu[1,*], Yadong Wang[1,*] and Yi Xing[2,3,*]



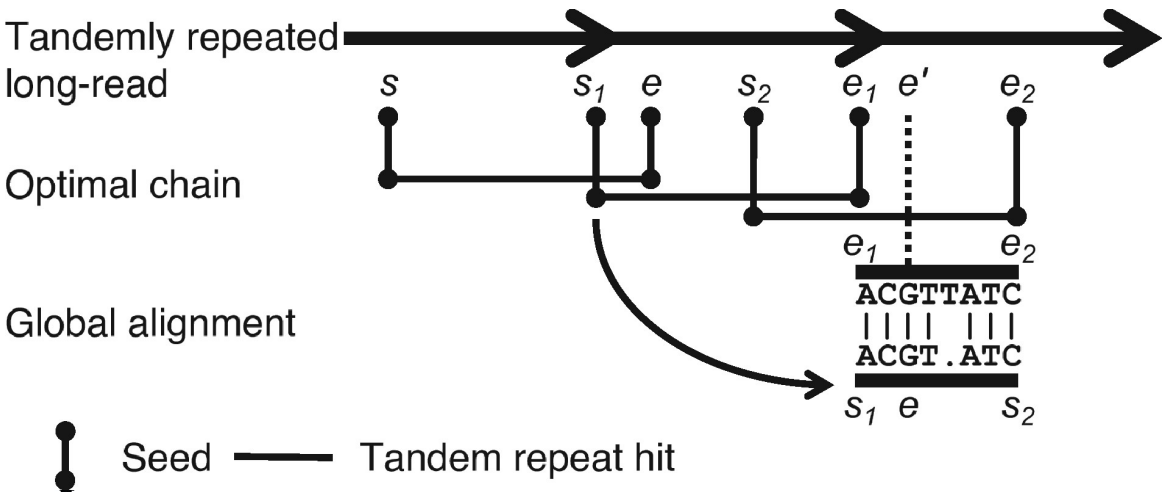
Tandem Repeat Annotation

Bioinformatics, 35, 2019, i200–i207
doi: 10.1093/bioinformatics/btz376
ISMB/ECCB 2019

OXFORD

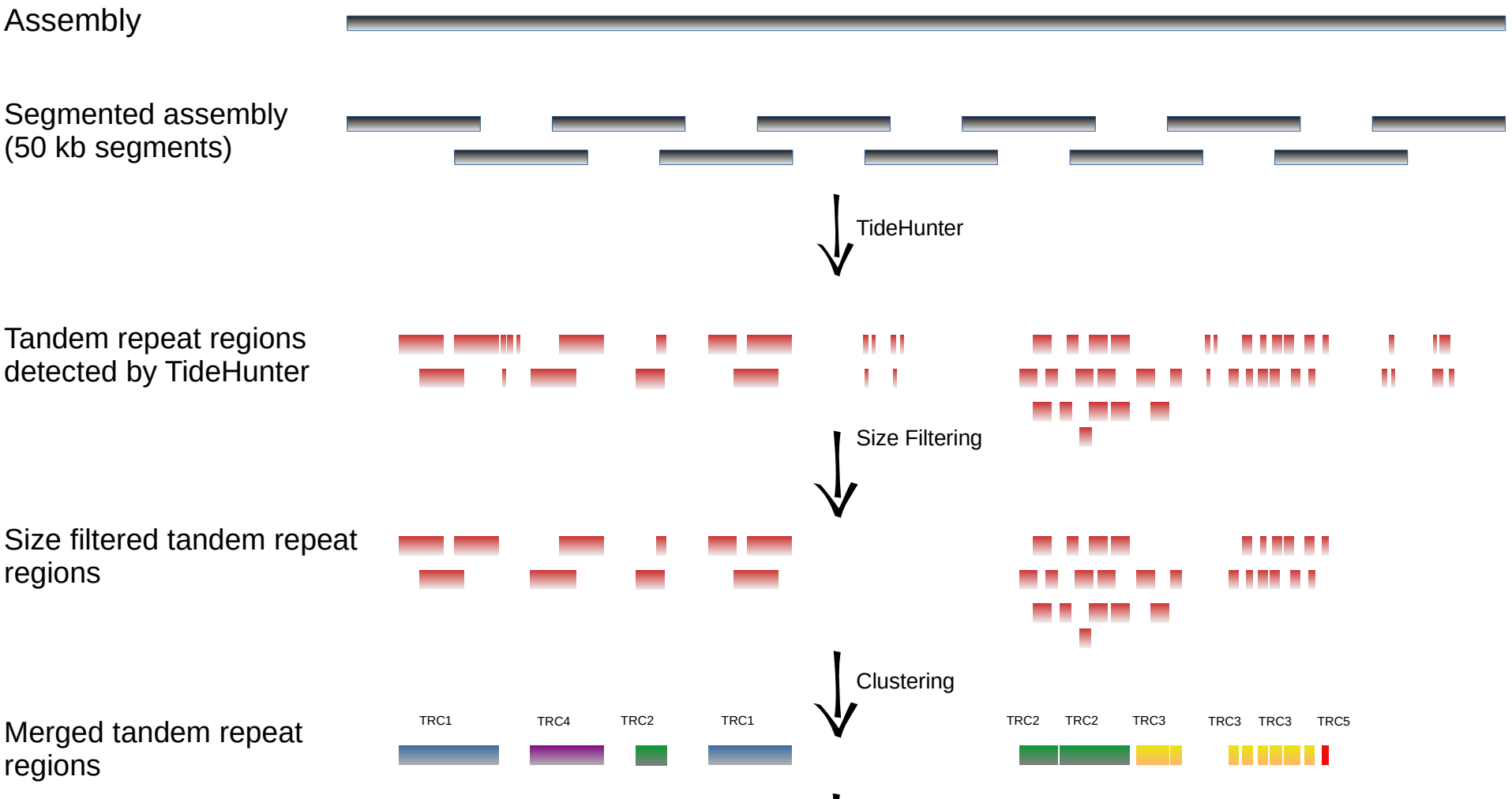
TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain

Yan Gao[1,2], Bo Liu[1,*], Yadong Wang[1,*] and Yi Xing[2,3,*]



- Sensitive – allows high divergence between consecutive repeats
- No limits on maximum repeat size
- Designed for long reads
- Slow on assemblies

TideCluster



TideCluster

Two-Step Clustering

1) Linear time clustering using mmseq2

- k-mer based
- Fast but create high number of small clusters

2) All-to-all BLAST

- Connected component

Simple sequence repeats are removed before clustering

TideCluster Workflow

Merged tandem repeat regions

Similarity based annotation (optional step)

Annotated tandem repeat regions

Satellite/FabTR-106 (99.9%25)

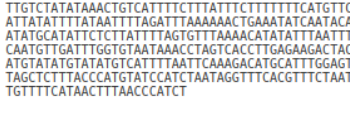
Satellite/FabTR-106 (99.9%25)

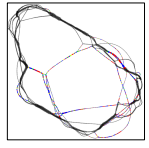
Satellite/FabTR-48 (87.8%25)

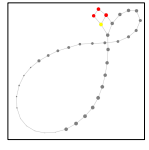
TAREAN report

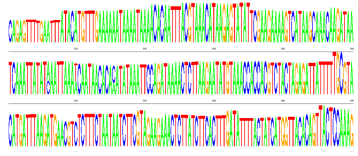
TAREAN SSRs analysis

Satellite/FabTR-106 (99.9%25)









TideCluster Workflow

Merged tandem repeat regions

Similarity based annotation

Annotated tandem repeat regions

Satellite/FabTR-106 (99.9%25)

Satellite/FabTR-106 (99.9%25)

Satellite/FabTR-48 (87.8%25)

TAREAN
SSRs analysis

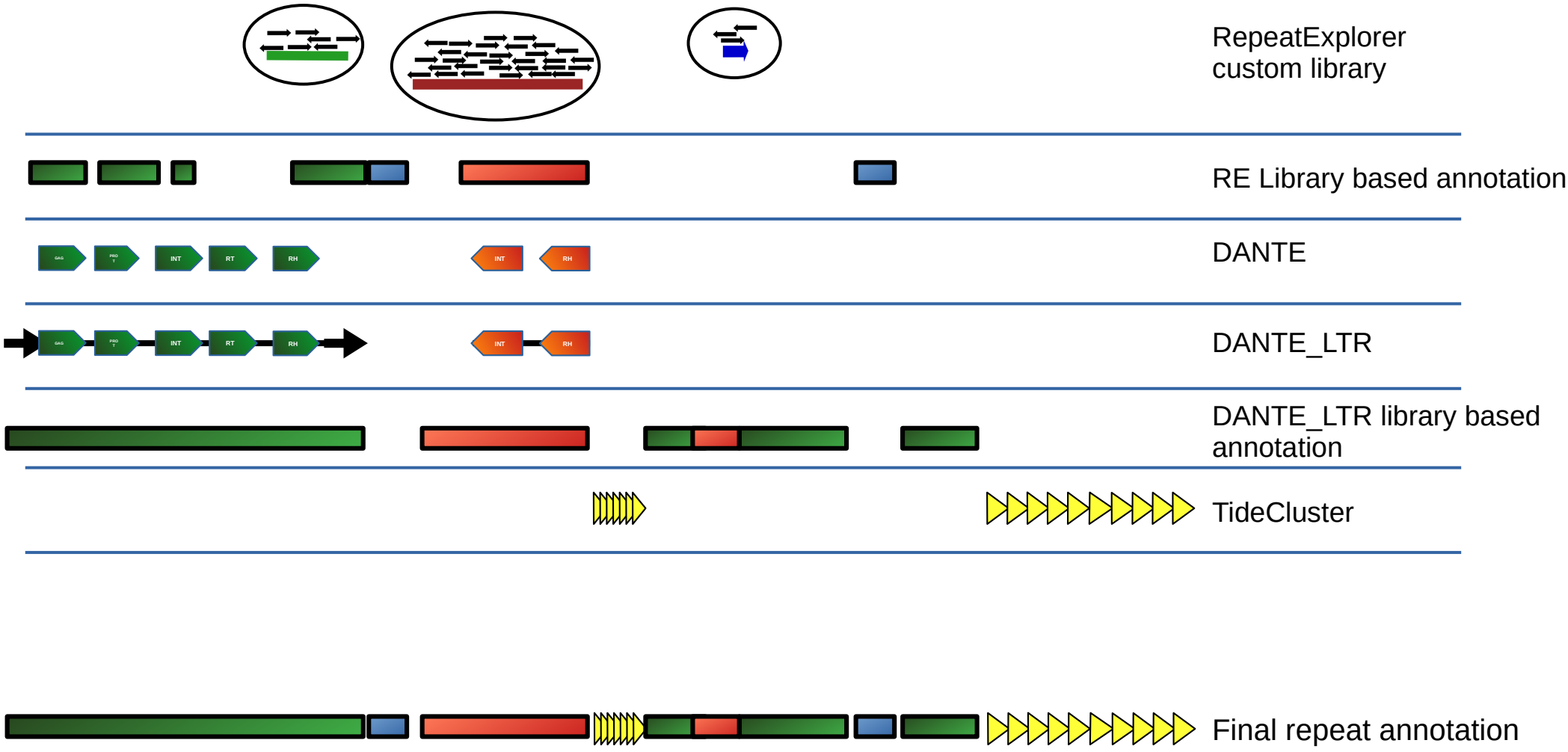
Simple sequence repeats report

TRC	Total size	SSRs	Number of arrays
TRC_22	42138	AC (85.5%)	3
TRC_31	27710	AGAT (100.0%)	2

TideCluster Parameter Settings

- **TideHunter parameters**
  - Minimum period size of tandem repeat
  - Maximum period size of tandem repeat
  - Maximum allowed divergence rate between two consecutive repeats
- **Minimum length of tandem repeat array to be included in clustering step. Shorter arrays are discarded, default 5000**
- **Minimum combined length of tandem repeat arrays within a single cluster, required for inclusion in TAREAN analysis, default 50000**

Complete Assembly Annotation Workflow



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