DANTE

PROFREP

ASSEMBLY ANNOTATION TOOLS FOR REPEATS

Nina Hoštáková

June 2018

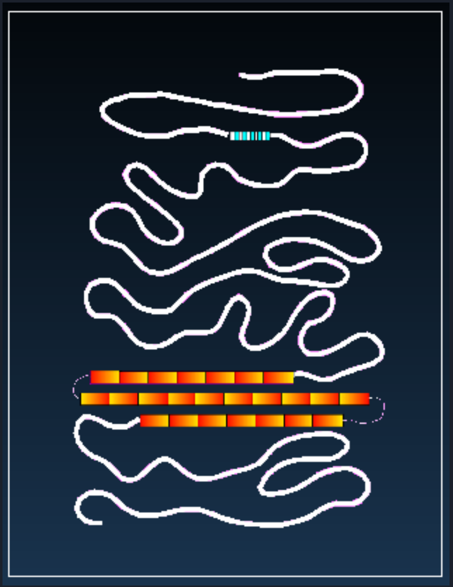
Repetitive DNA

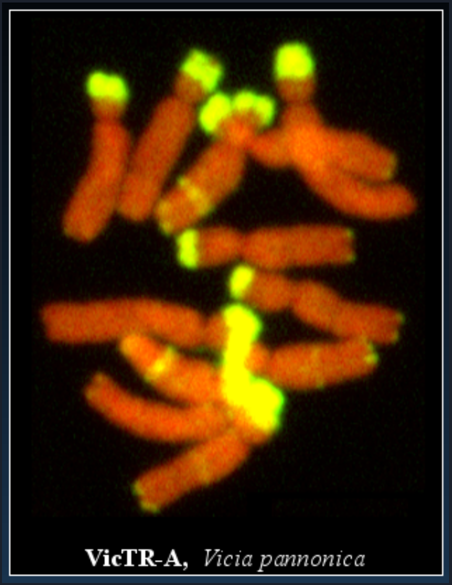
- Sequences present in tens up to millions of copies
- Substantial part of the genomes (> 90% in some plants)
- It creates groups of genetic elements sharing mechanism of formation and multiplication in the genome

PROFREP

DANTE

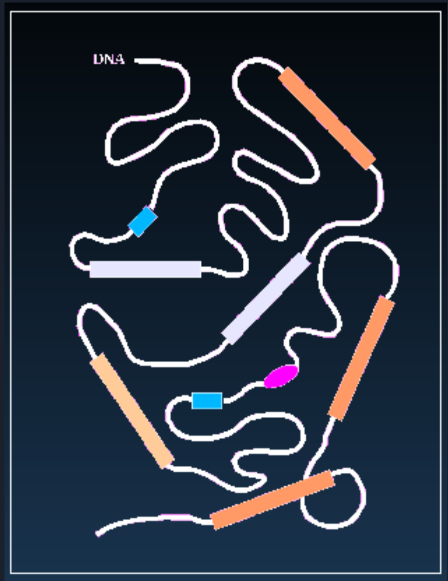
Tandem repeats

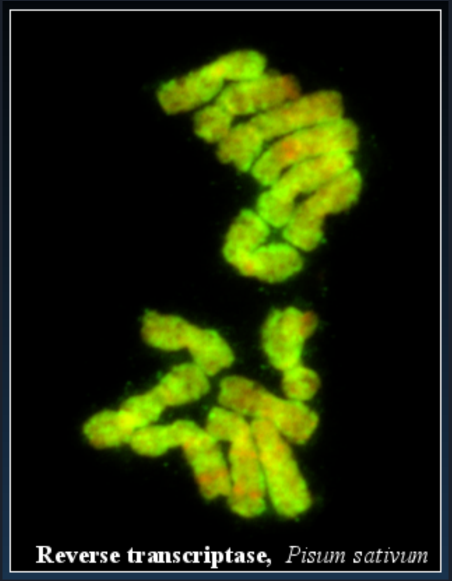




VicTR-A, *Vicia pannonica*

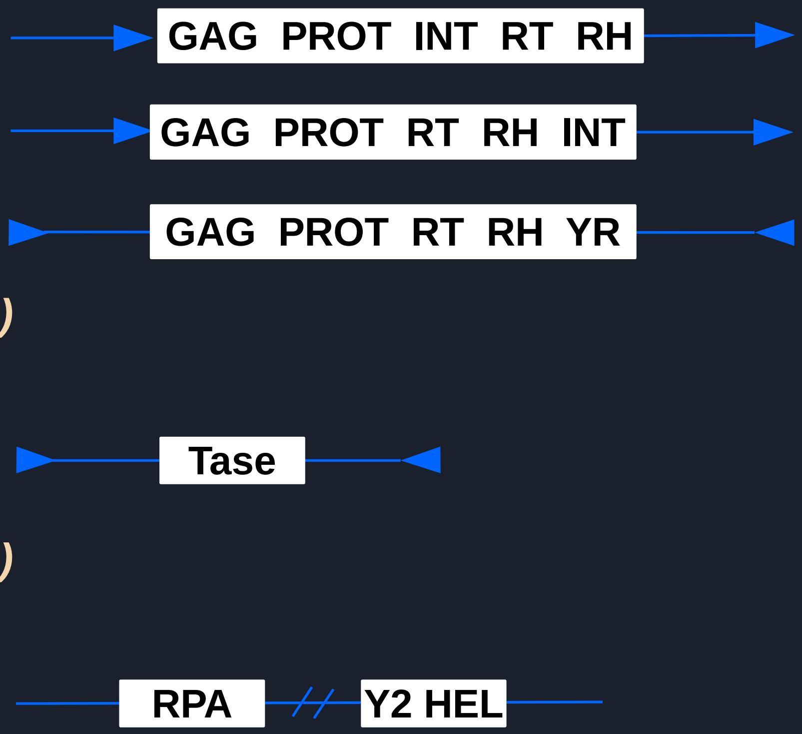
Mobile elements (transposons)



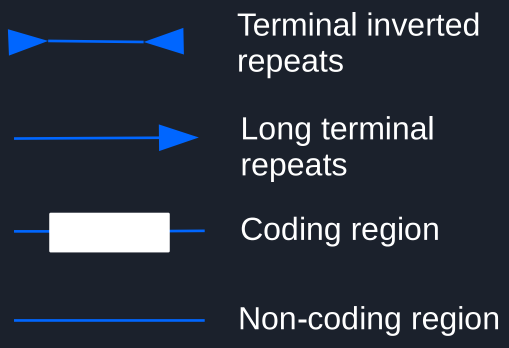


Reverse transcriptase, *Pisum sativum*

Mobile elements



Classification		Structure
Order	**Superfamily**	
Class I (retrotransposons)		
LTR	*Ty1/copia*	GAG PROT INT RT RH
	Ty3/gypsy	GAG PROT RT RH INT
DIRS	*DIRS*	GAG PROT RT RH YR
Class II (DNA transposons) – Subclass I		
TIR	*Tc1-Mariner*	Tase
Class II (DNA transposons) – Subclass II		
Helitron	*Helitron*	RPA // Y2 HEL



Terminal inverted repeats

Long terminal repeats

Coding region

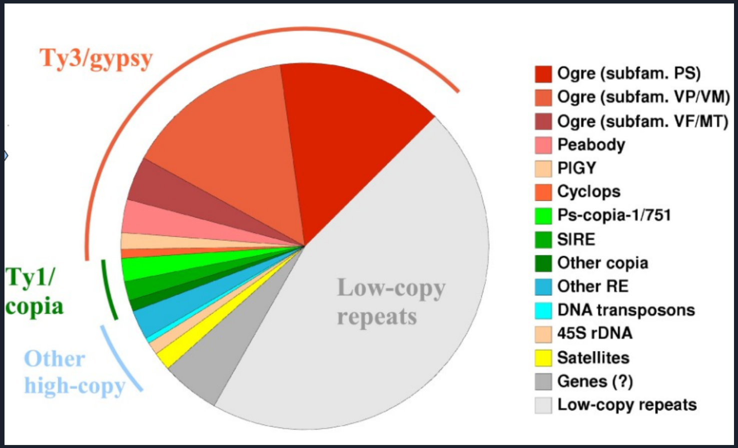
Non-coding region

GAG	Capside protein
PROT	Proteinase
RT	Reverse transcriptase
INT	Integrase
RH	RNase H
Tase	Transposase
YR	Tyrosine recombinase
Y2	YR with YY motif
RPA	Replication protein A
HEL	Helicase

NGS reads

REPEAT EXPLORER

Representation of repeats in the genome

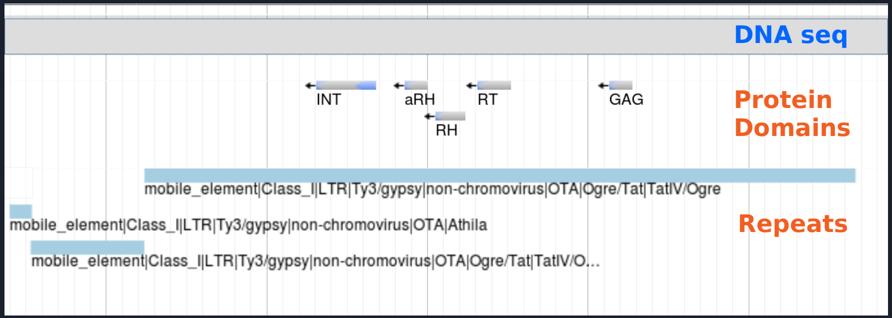


DNA sequence

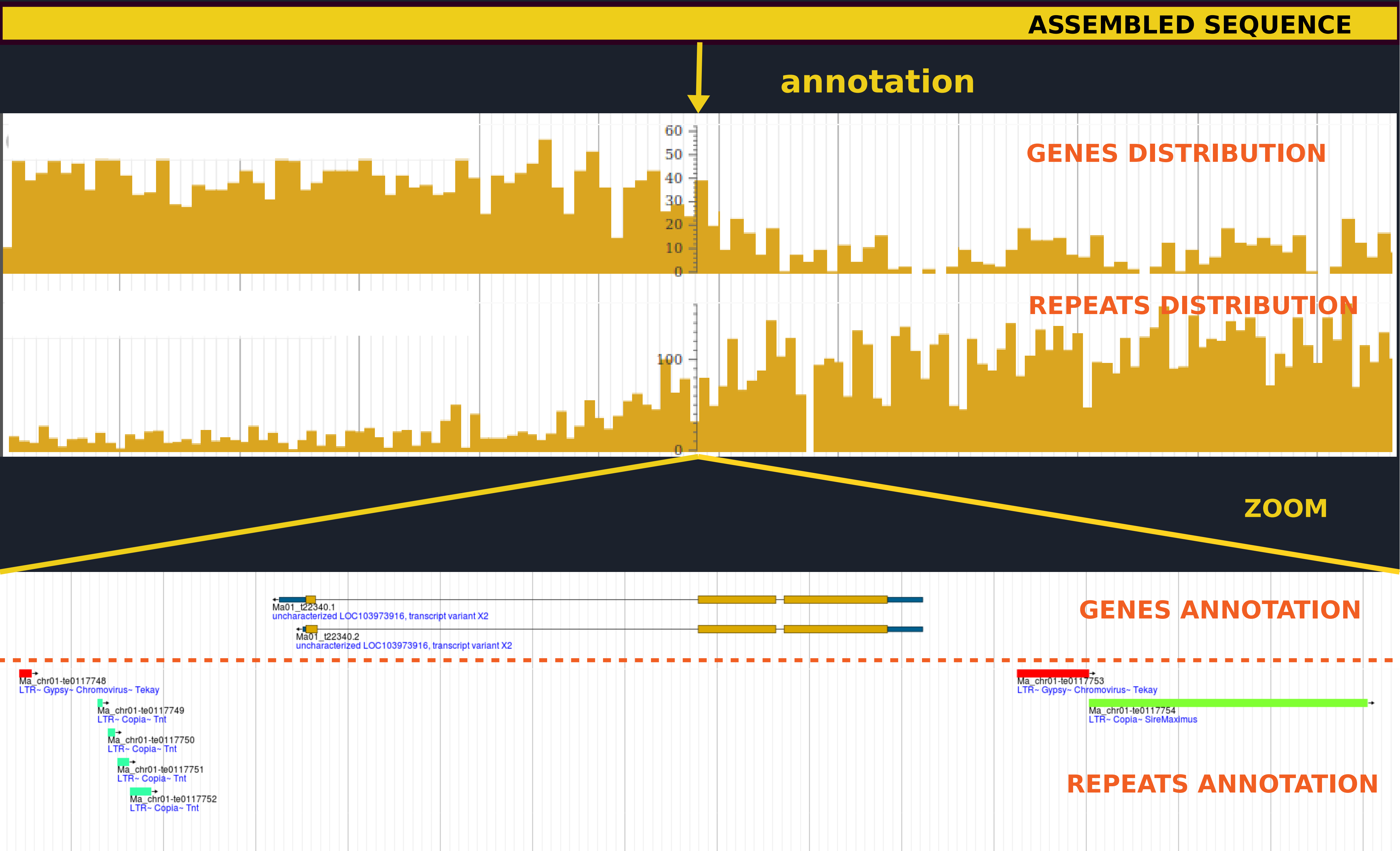
PROFREP

DANTE

annotation



Assembly Annotation



ASSEMBLED SEQUENCE

annotation

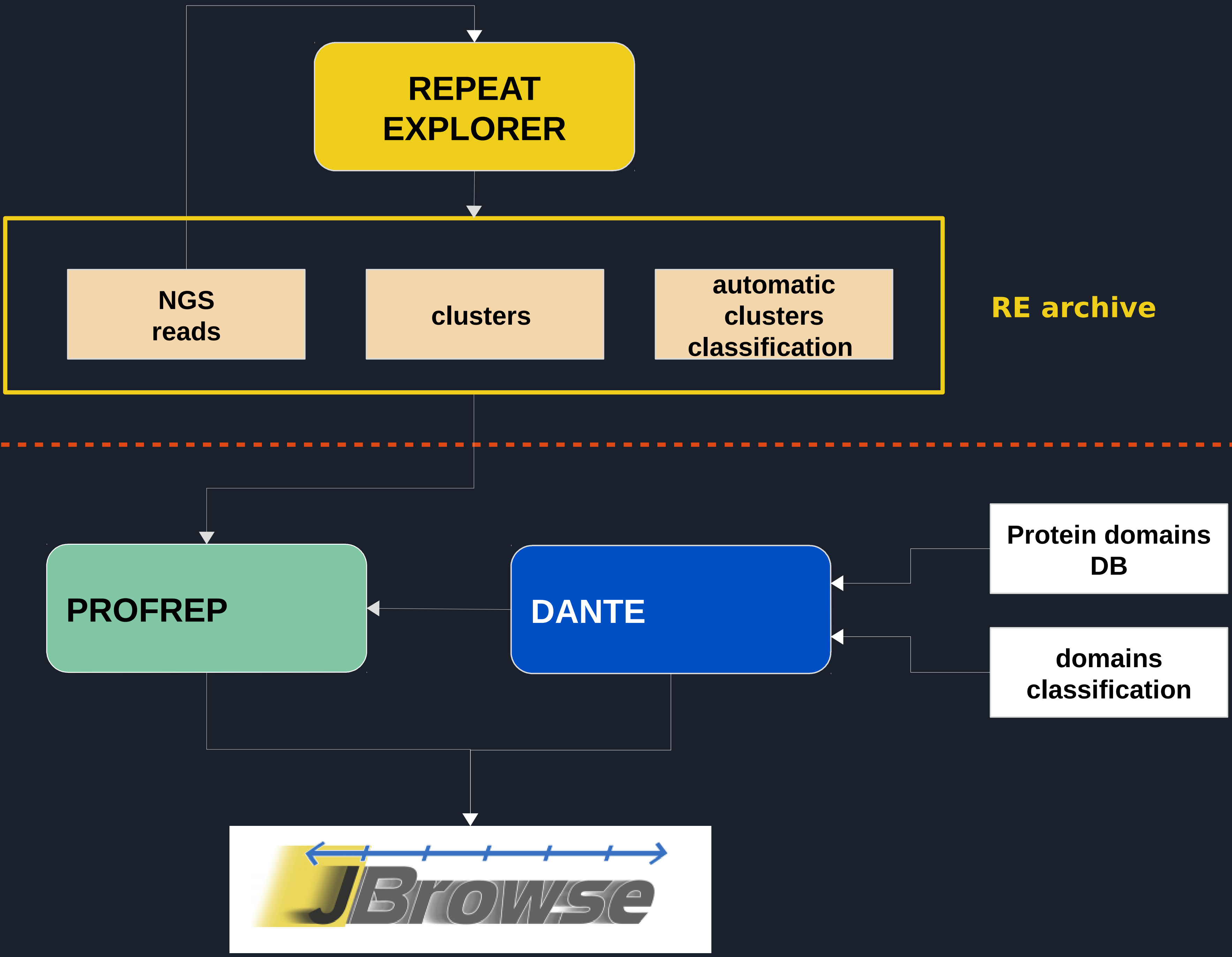
GENES DISTRIBUTION

REPEATS DISTRIBUTION

ZOOM

GENES ANNOTATION

REPEATS ANNOTATION



REPEAT EXPLORER

NGS reads

clusters

automatic clusters classification

RE archive

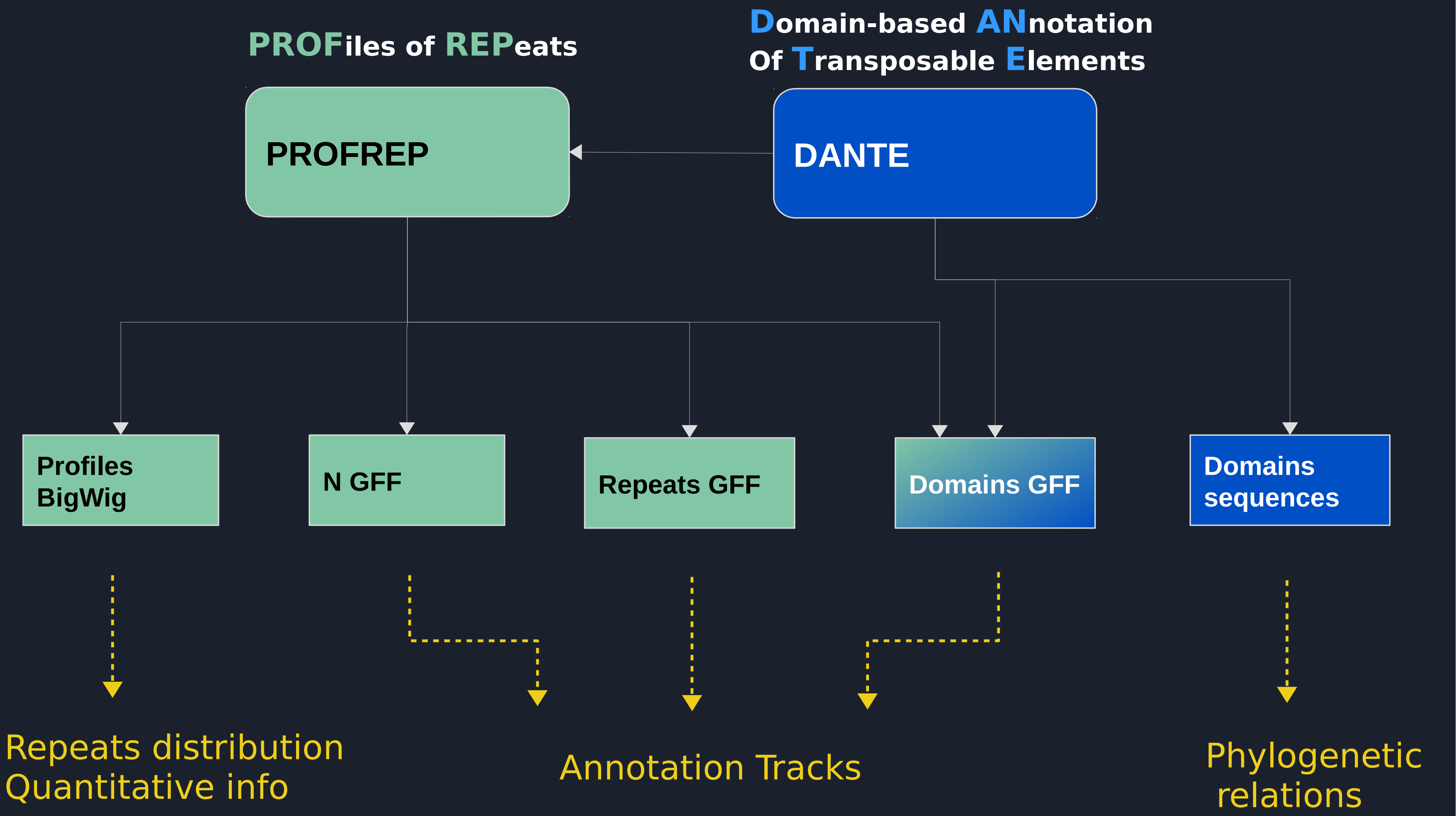
PROFREP

DANTE

Protein domains DB

domains classification

JBrowse



PROFiles of **REP**eats

Domain-based **AN**notation
Of **T**ransposable **E**lements

PROFREP

DANTE

Profiles BigWig

N GFF

Repeats GFF

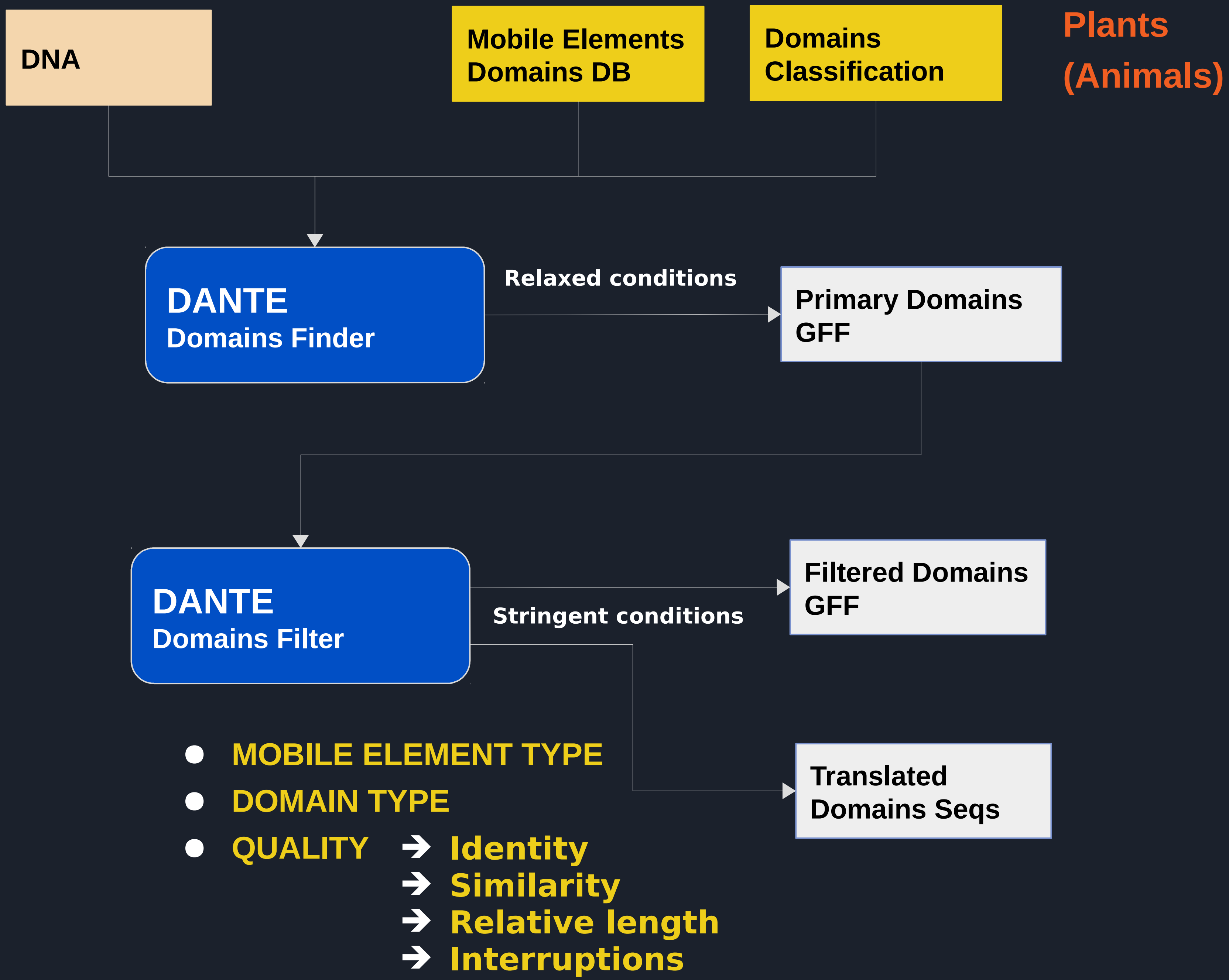
Domains GFF

Domains sequences

Repeats distribution
Quantitative info

Annotation Tracks

Phylogenetic
relations



DNA

Mobile Elements Domains DB

Domains Classification

Plants (Animals)

DANTE
Domains Finder

Relaxed conditions

Primary Domains GFF

DANTE
Domains Filter

Stringent conditions

Filtered Domains GFF

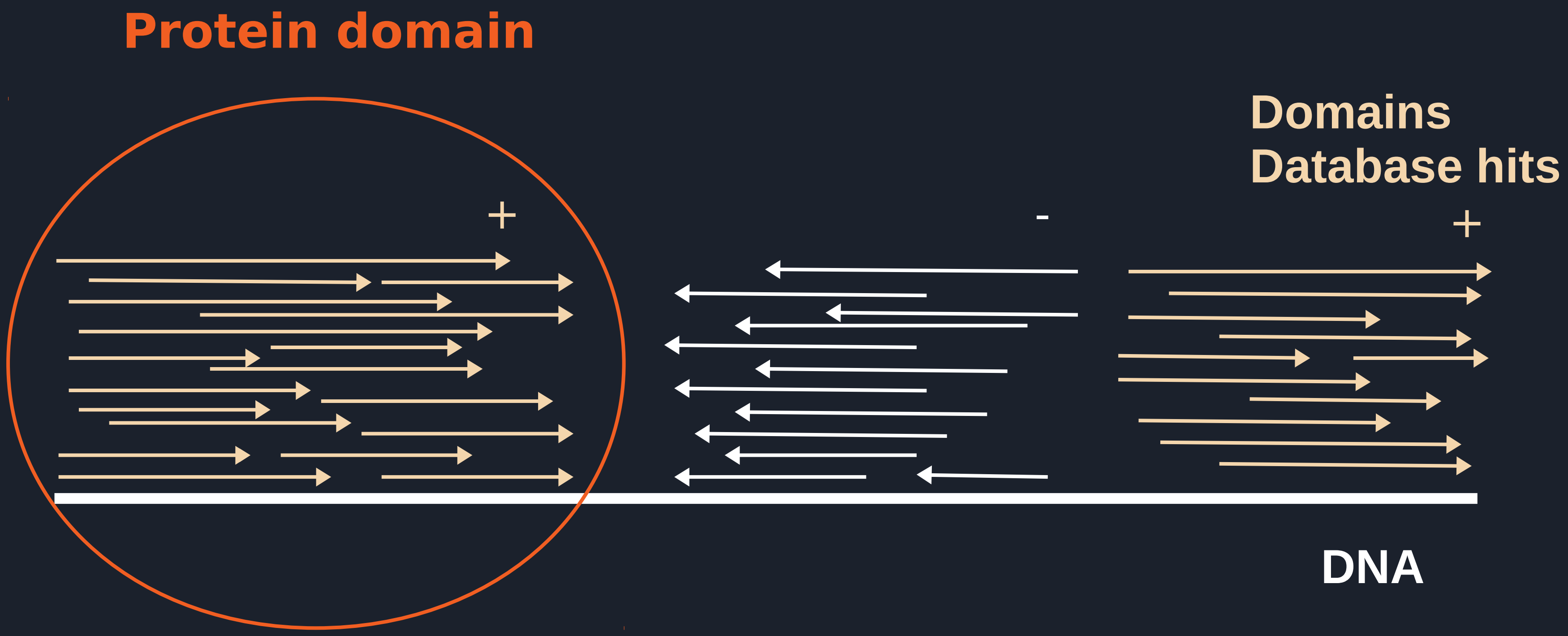
Translated Domains Seqs

- MOBILE ELEMENT TYPE
- DOMAIN TYPE
- QUALITY
  - → Identity
  - → Similarity
  - → Relative length
  - → Interruptions

Principle

-DANTE-

DNA x PROT similarity search



Protein domain

Domains
Database hits

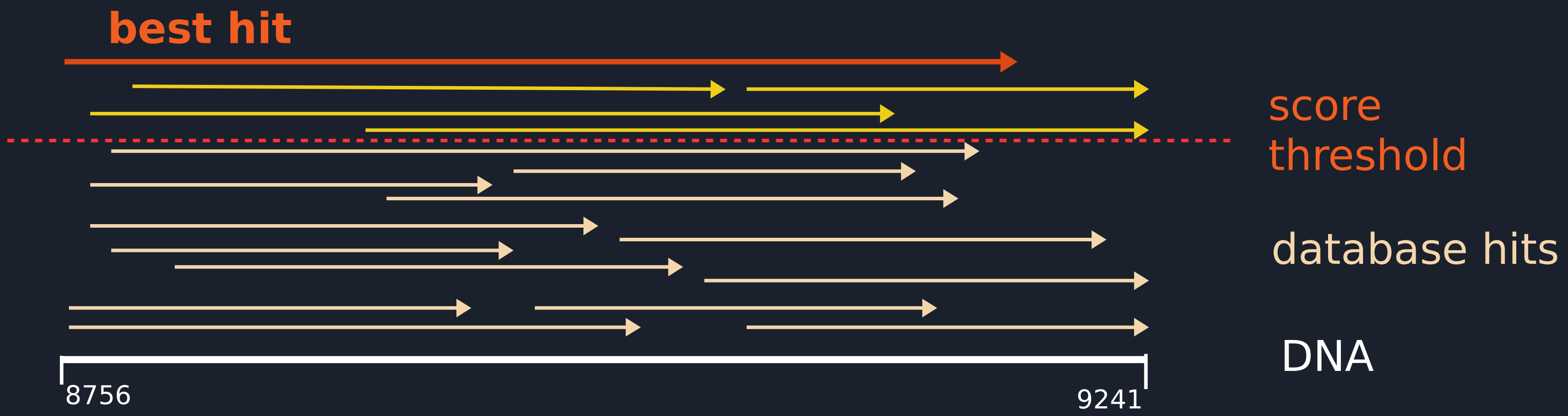
+

-

+

DNA

Principle -DANTE-



Hits classifications

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

Common classification level

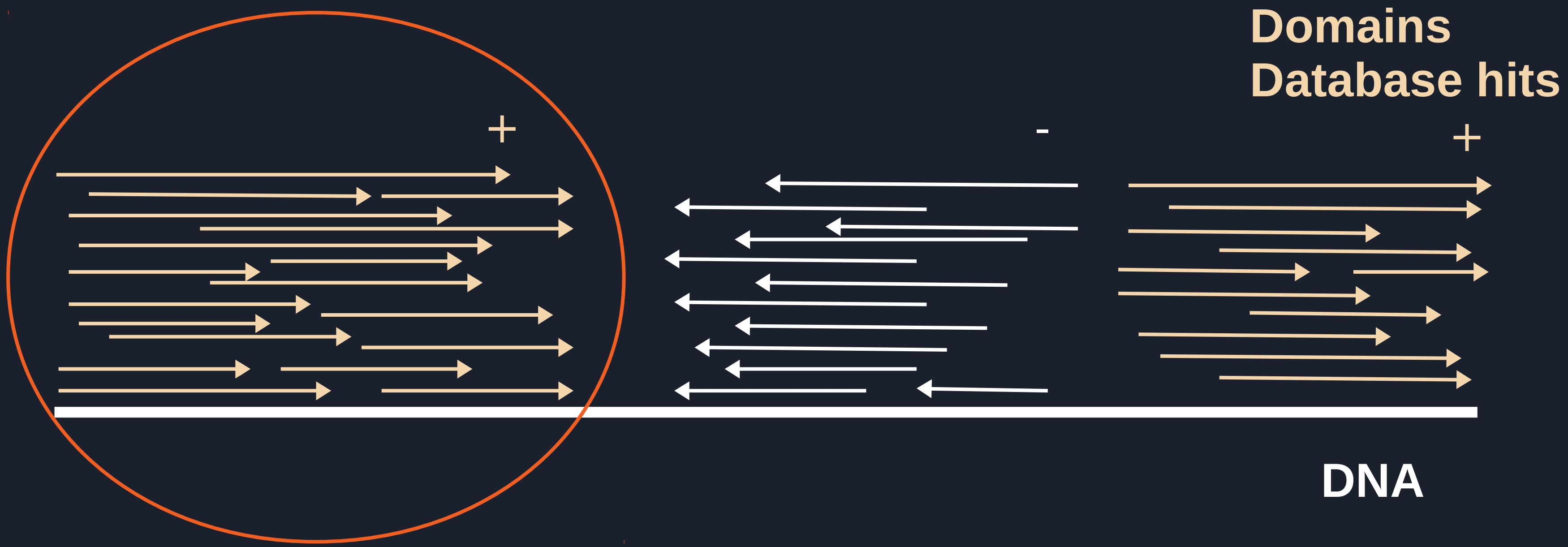
RT Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat

Principle

-DANTE-

DNA x PROT similarity search

Protein domain

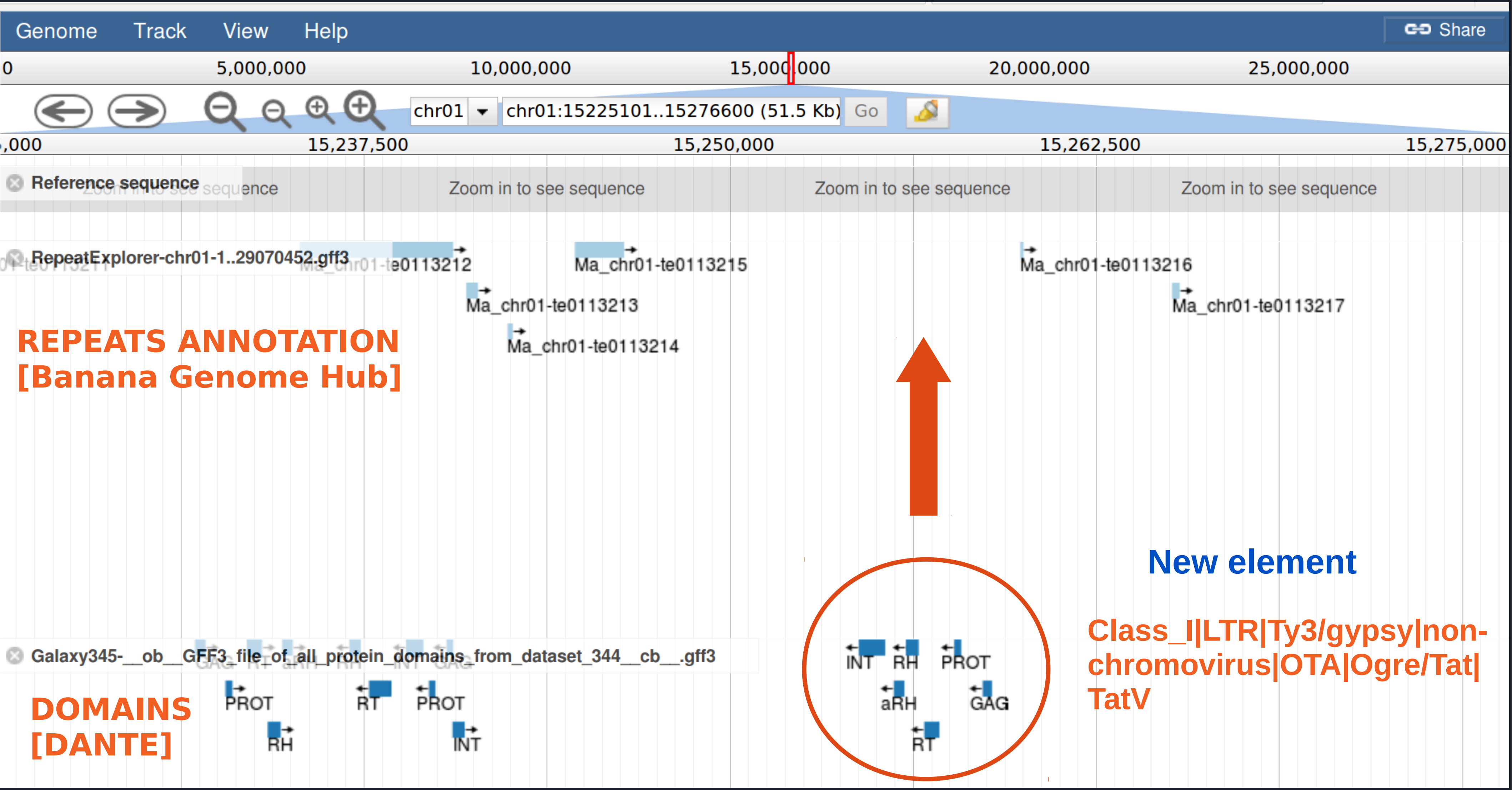


Reverse Transcriptase (RT)

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

Banana assembly (2016), chromosome 1

TatV element



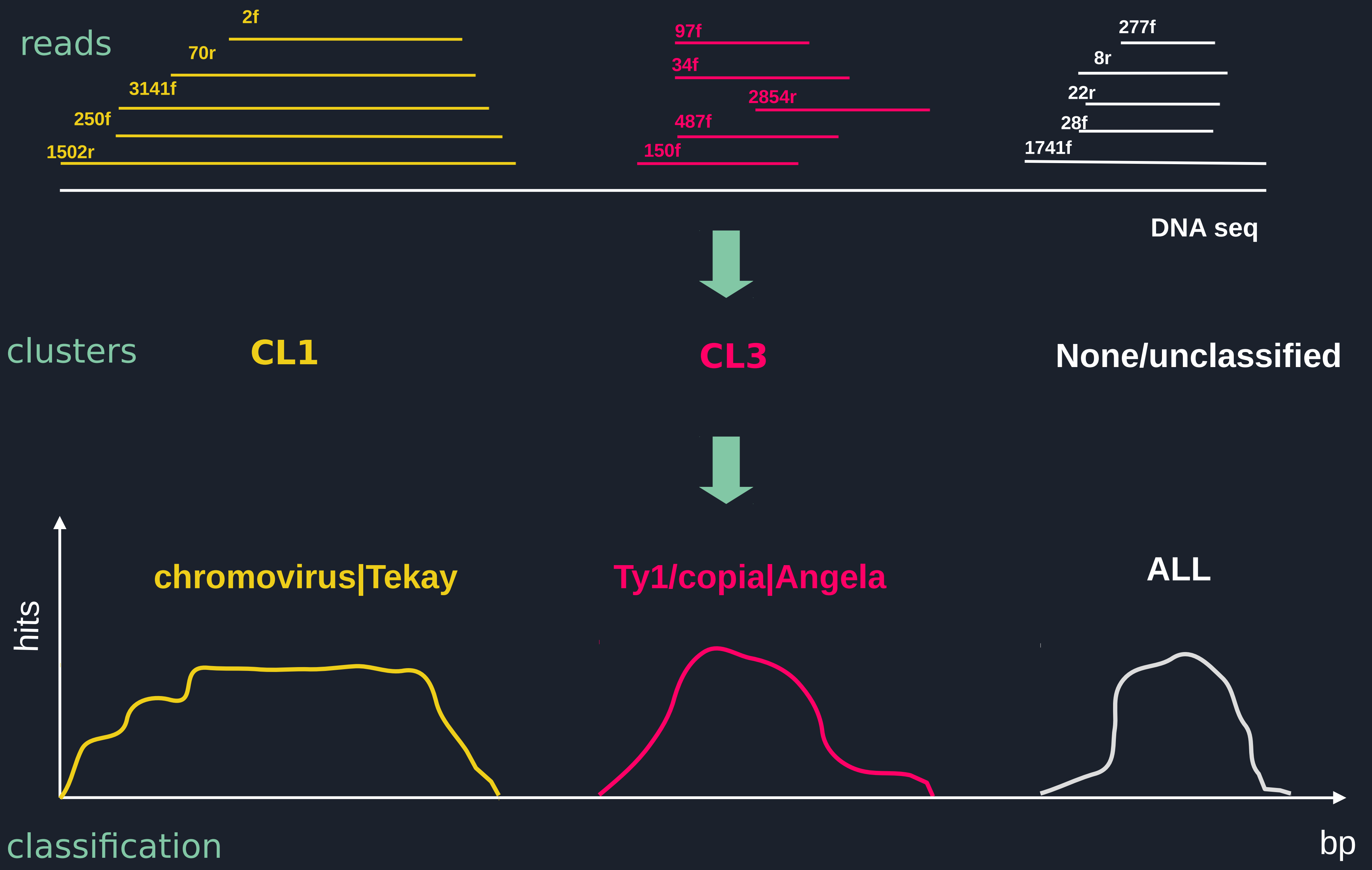
REPEATS ANNOTATION [Banana Genome Hub]

DOMAINS [DANTE]

New element

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

Principle -PROFREP-



clusters

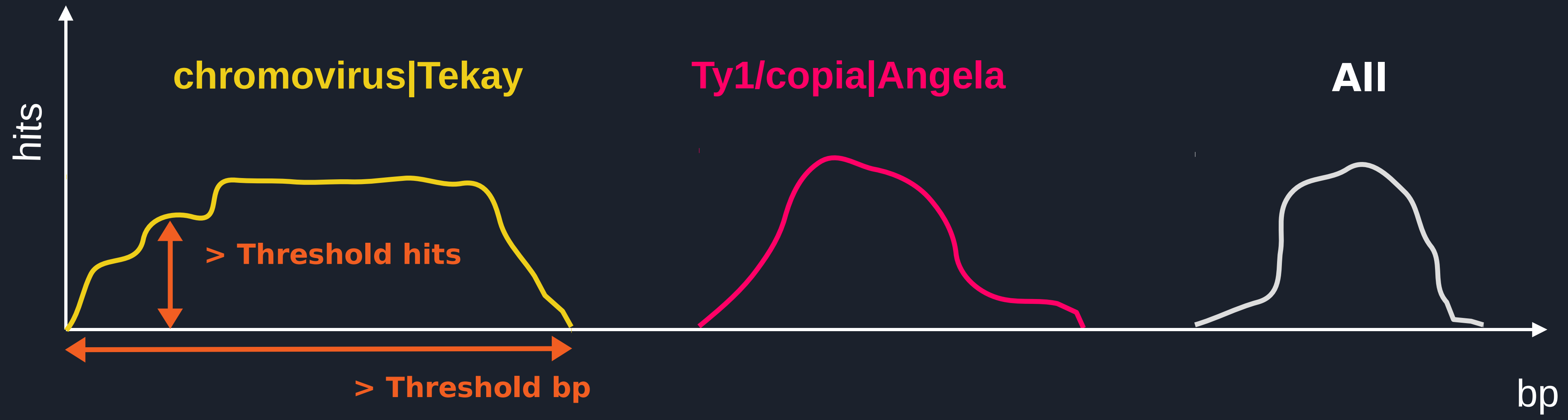
CL1 **CL3** **None/unclassified**

chromovirus|Tekay **Ty1/copia|Angela** **ALL**

classification

Principle -PROFREP-

1. PROFILES



2. GFF3

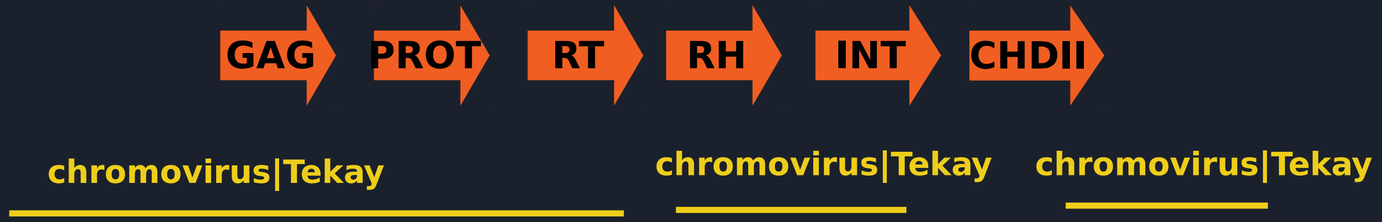
Annotation Refining -PROFREP-

1. Removing low confidence nested regions

PID diff > 5 %

Overhang < 10b

2. Interconnecting regions



Min 2 domains

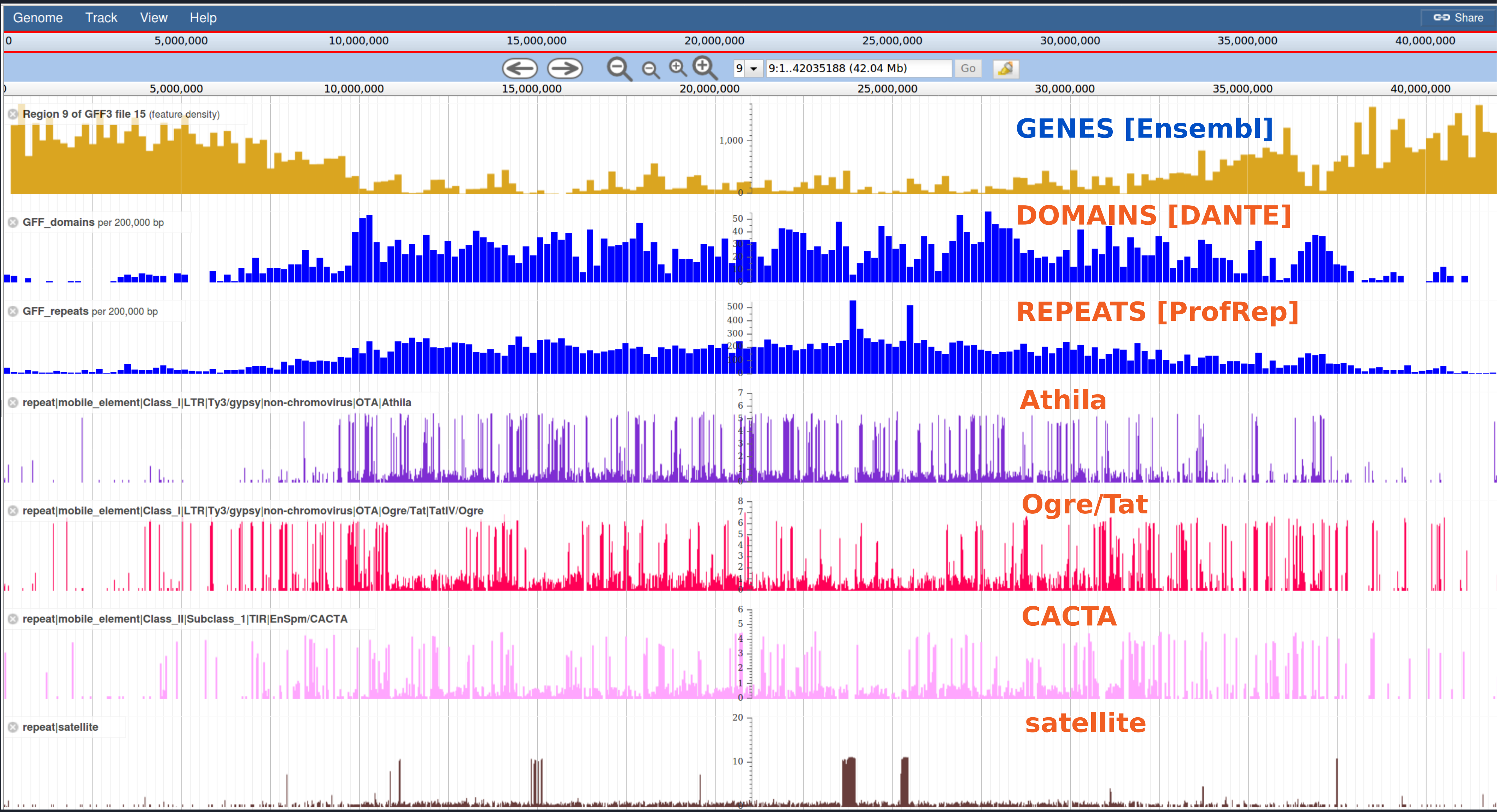
Gap < 250b

Same class

Last class level

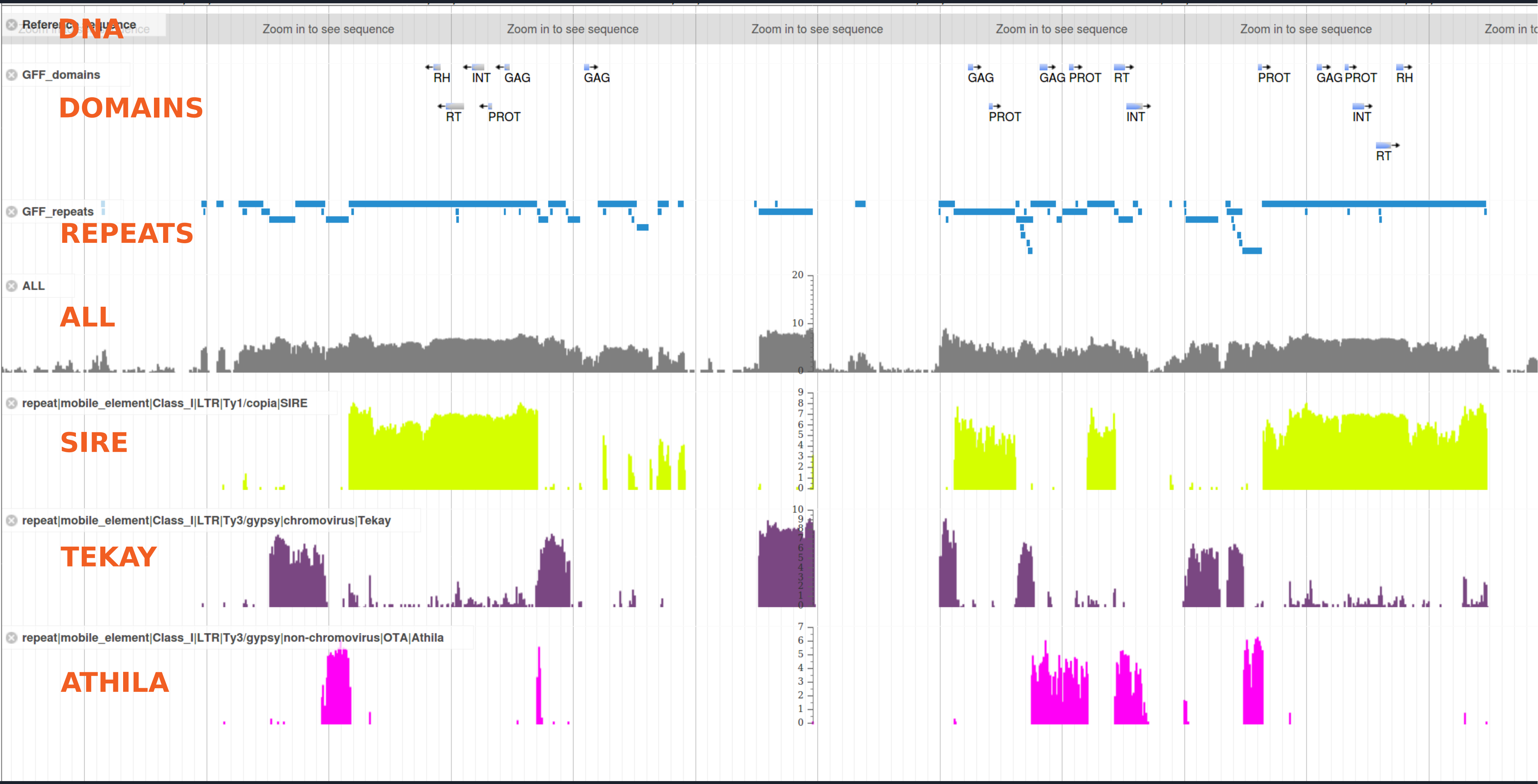
Cacao assembly (2014), chromosome 9

Distributions



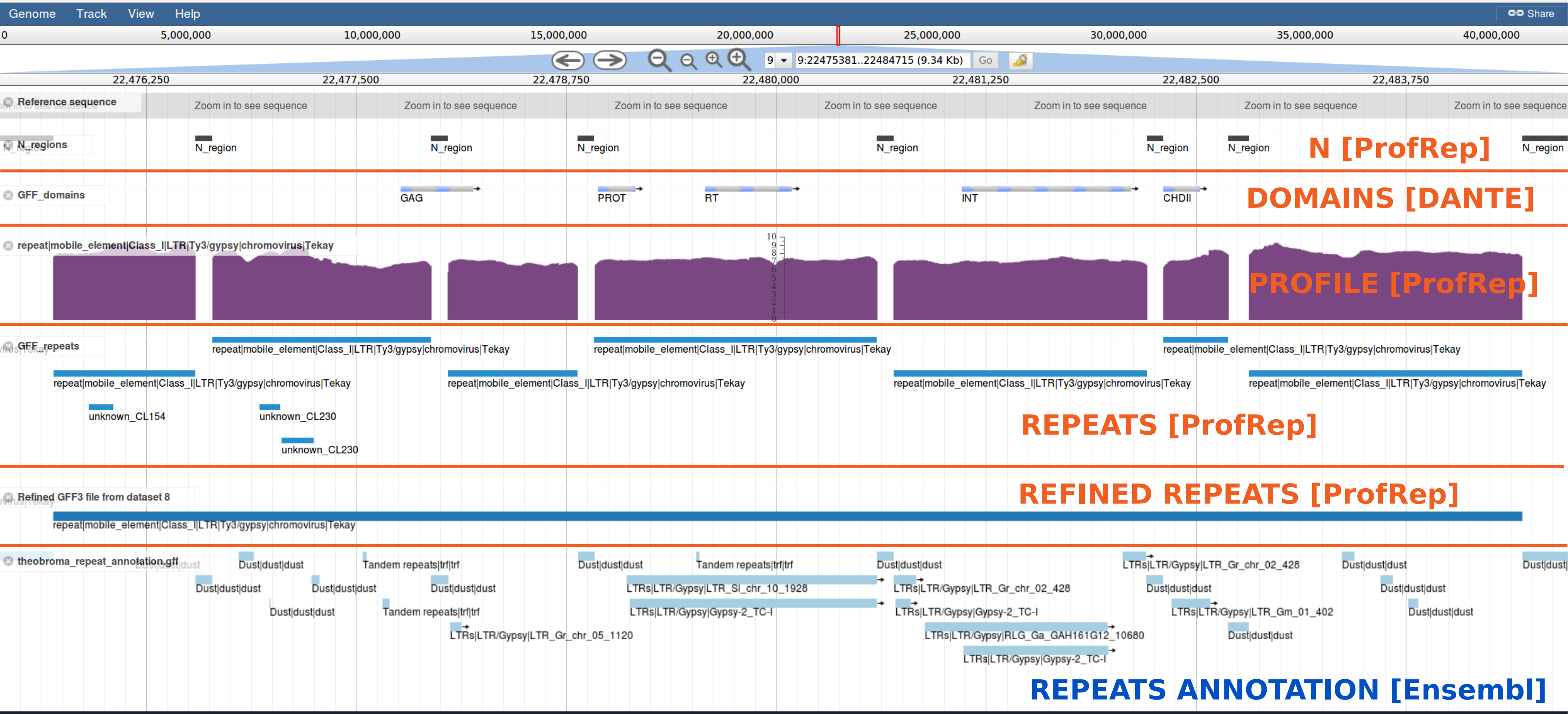
Cacao assembly (2014), chromosome 9

Annotations



Cacao assembly (2014), chromosome 9

Chromovirus|Tekay element



N [ProfRep]

DOMAINS [DANTE]

PROFILE [ProfRep]

REPEATS [ProfRep]

REFINED REPEATS [ProfRep]

REPEATS ANNOTATION [Ensembl]

THANK YOU FOR YOUR ATTENTION !

https://repeatexplorer-elixir.cerit-sc.cz/

BIOLOGY CENTRE ASCR

elixir

cerit scientific cloud

CESNET