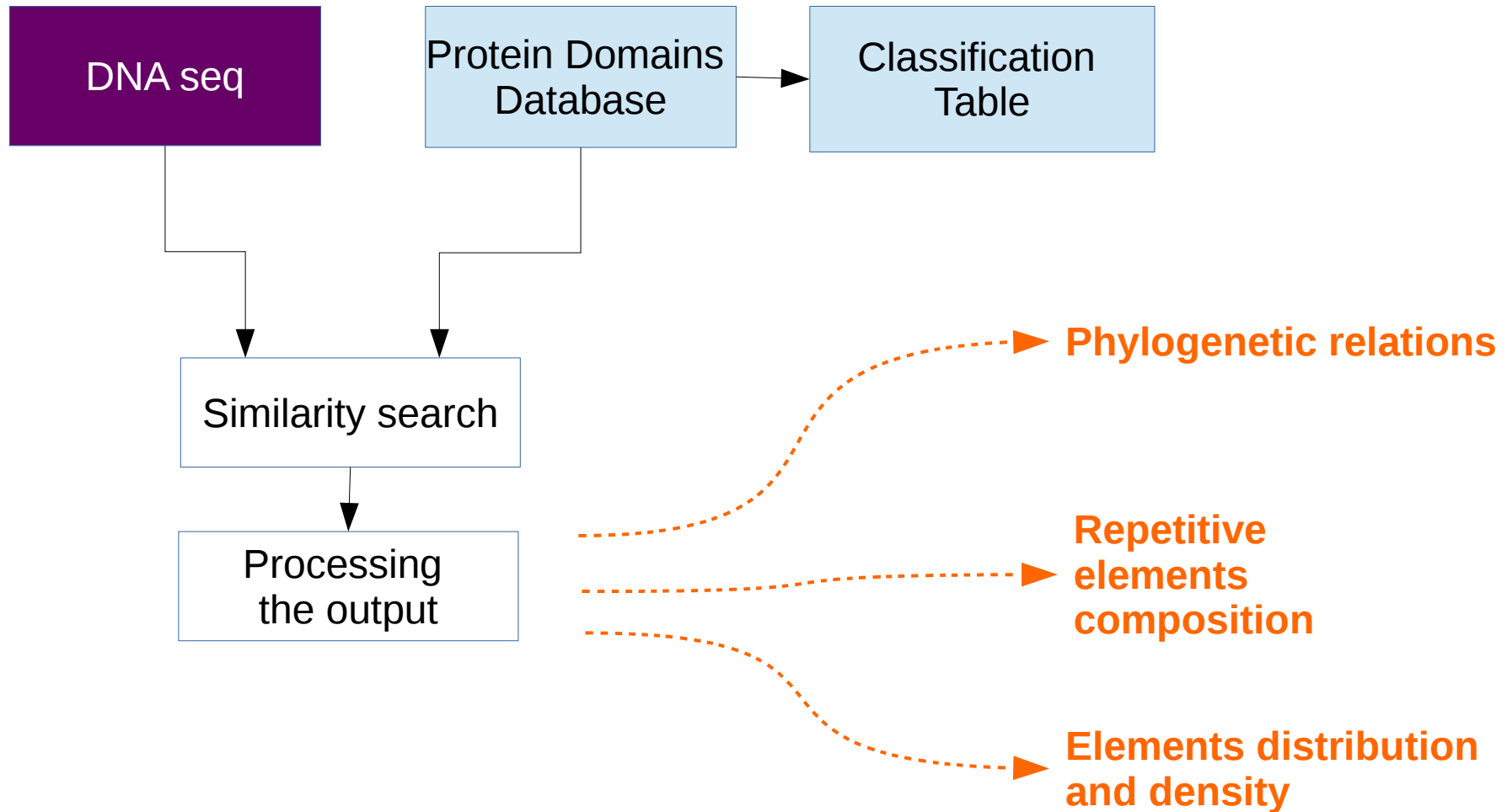


PROTEIN DOMAINS TOOLS



A word cloud of protein domain names. The words are arranged in a roughly triangular shape, pointing to the right. The colors are primarily blue, purple, and orange. The words include: HELIX, ENDOD, PROTI, INT, RT, CHDCR, GAG, aRH, NTPase, and YR. The words are of varying sizes and are oriented horizontally or vertically.

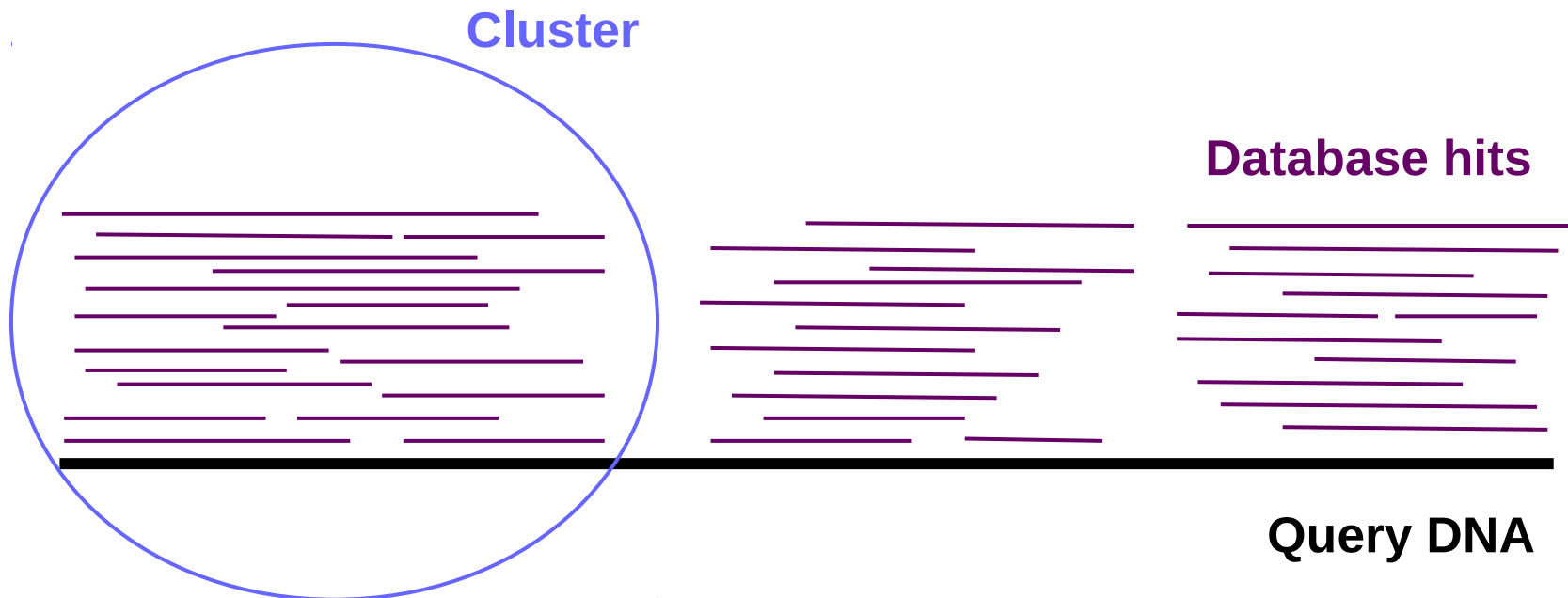
Purpose



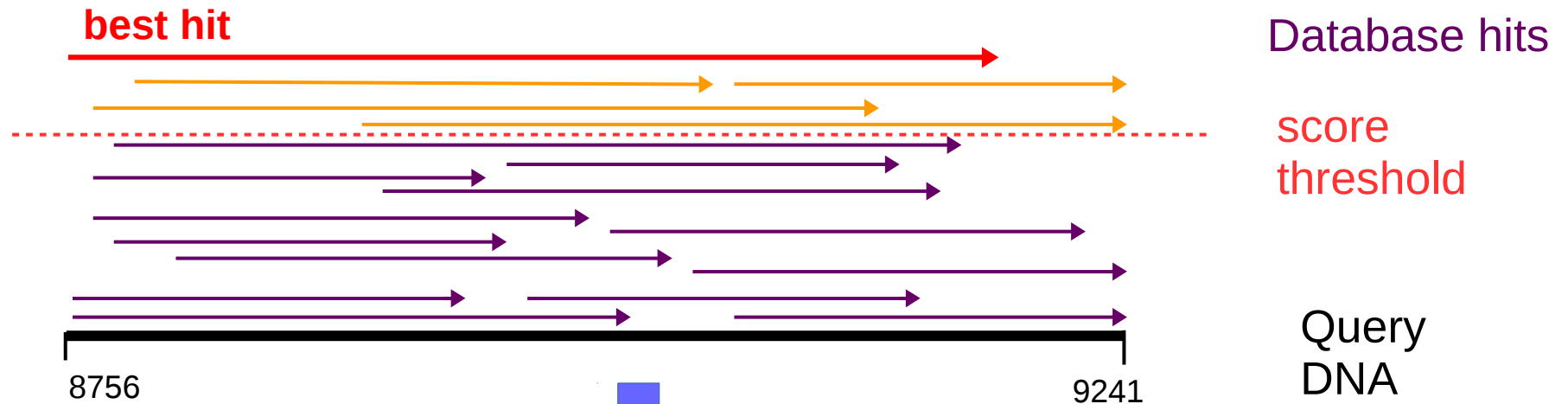
Principle – similarity search & building clusters

LAST

- Database indexed files
- Alignment – spaced adaptive seeds



Principle – cluster processing



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

GFF3 Standard Fields

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
scaffold146.1 size86774	profrep	protein_domain	976	1289	293	+	.	Name=RH;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	6810	7049	153	+	.	Name=PROT;Final_Classification=Cl
scaffold146.1 size86774	profrep	protein_domain	7656	8296	.	+	.	Name=RT/INT;Final_Classification=A
scaffold146.1 size86774	profrep	protein_domain	8756	9241	538	+	.	Name=RT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	9433	9781	343	+	.	Name=RH;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	10810	11667	747	+	.	Name=INT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	14592	14828	289	+	.	Name=PROT;Final_Classification=Cl
scaffold146.1 size86774	profrep	protein_domain	14856	15032	82	+	.	Name=PROT;Final_Classification=Cl
scaffold146.1 size86774	profrep	protein_domain	15420	15995	871	+	.	Name=RT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	16188	16634	628	+	.	Name=RH;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	24522	24659	149	+	.	Name=PROT;Final_Classification=Cl
scaffold146.1 size86774	profrep	protein_domain	24873	25481	913	+	.	Name=INT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	26313	27428	1060	+	.	Name=RT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	27723	28124	581	+	.	Name=RH;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	9780	9956	178	-	.	Name=INT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	10299	10658	303	-	.	Name=aRH;Final_Classification=Clas
scaffold146.1 size86774	profrep	protein_domain	10701	10817	136	-	.	Name=RH;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	16797	17666	1057	-	.	Name=INT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	18554	18811	306	-	.	Name=INT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	19158	19478	197	-	.	Name=INT;Final_Classification=Class
scaffold146.1 size86774	profrep	protein_domain	19976	20212	259	-	.	Name=PROT;Final_Classification=Cl
scaffold146.1 size86774	profrep	protein_domain	28912	29124	216	-	.	Name=PROT;Final_Classification=Cl

GFF3 Attributes

Cluster

- **Name**=RT
- **Final_Classification**=Class_|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat
- **Region_Hits_Classifications**=RT|Class_|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat|TatV[486bp],RT|Class_|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat|TatIV|Ogre[441bp]

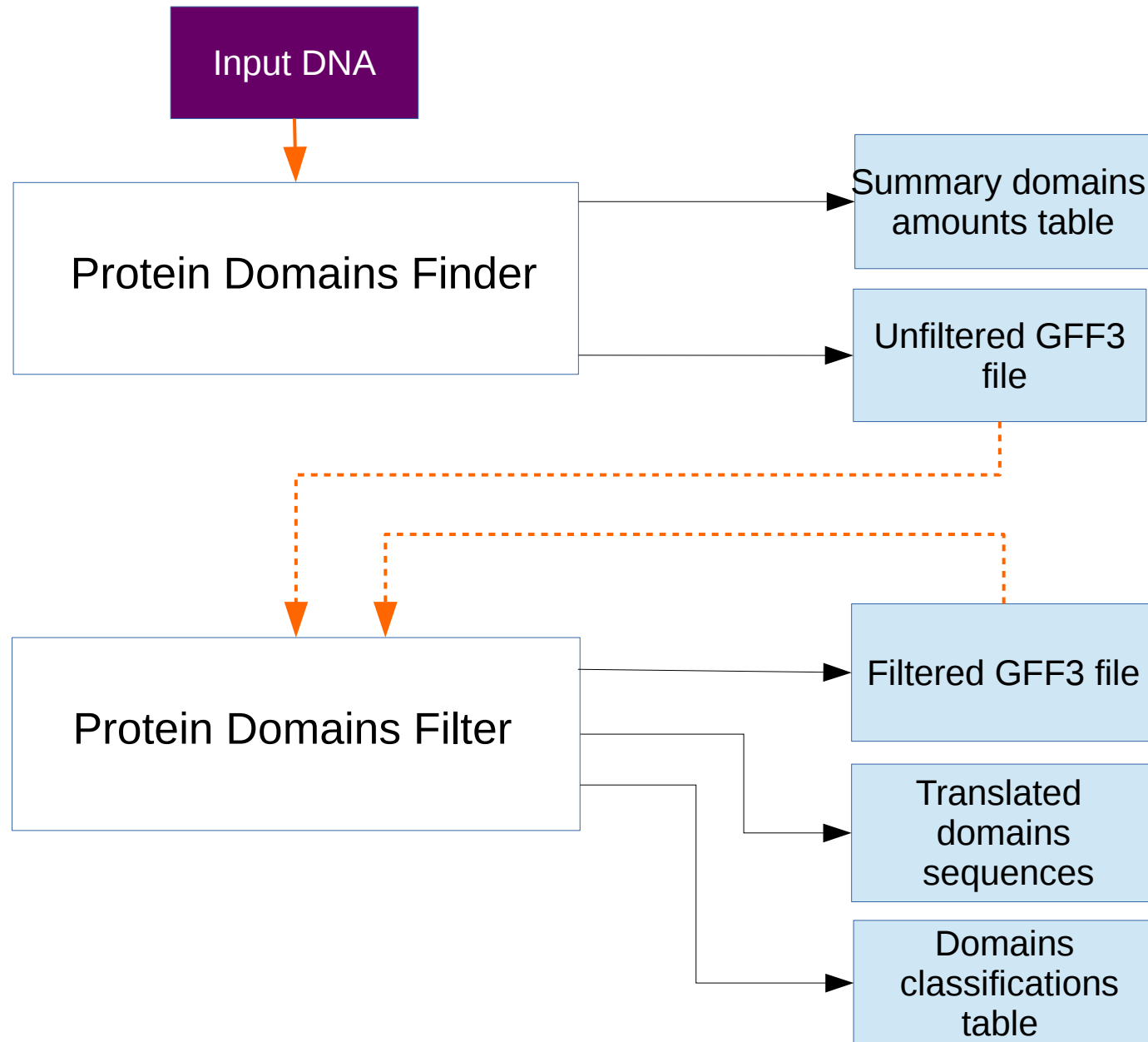
----- Ambiguous domain type

----- Conclusive domain type

- **Best_hit**=Ty3-RT__REXdb_ID8210|Class_|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat|TatV:8801-9241[90%]
- **DB_seq**=DFTDLNKACPKDSFPLPHIDRLVDSTAGNELLTFMDAFSGYNQIMMNPEDQEKTSFITDRGIYCYKVMPFGLKNAGATYQRLVNKMFHNHLGKTMEVYIDDMLVKSLKKEDHVKHLEECFDILNKYQMKLNPAKCTFGVPSGEFLGY
- **Query_seq**=DFKGVNKHQCQDPFPLPHIDRLVDAVAGSSLLSTMDAYSGYHQISLAREDQAKSSFLTEDGVFCYVVMPPFGLRNAGATYQRLVNKIFADLLGKEMEIIYVDDMIVKSLNDEDHIIYLSHCFEVCRTHRLKLNPAKCCFGVRSGKFLGY
- **Identity**=0.63
- **Similarity**=0.8
- **Relat_Length**=0.85
- **Relat_Interruptions**=0.0

Corresponds to the best alignment of the cluster

Tools & Outputs



Outputs Examples

• Translated domains sequences

```
>scaffold146.1|size86774:6810-7049 PR0T Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat|TatV
LVDSGASCNLSKRVKMGIPDEKLEFLDATLYAFDRRTIIPAGKIQLPVTLGEEERTR
SEMVEFIIIVMDLAYNAILG
>scaffold146.1|size86774:8801-9241 RT Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat
DFKGVNKHCCQDPFPLPHIDRLVDAVAGSSLLSTMDAYSQYHQISLAREDQAKSSFLTED
GVFCYVVMPPFGLRNAGATYQRLVNKIFADLLGKEMEYVDDMIVKSLNDEDHIIYLSHCF
EVCRTHLRLKLNPAKCCFGVRSKGFLGY
>scaffold146.1|size86774:10819-11667 INT Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Ogre/Tat|TatV
RDAMDCVRRQCSCQYFAPINRKPGAIEITLTELPCPFDRWGIDILGPFPQSVRQRRFCIVA
VEYHSKWIEAEAVASITSEAVKKFVMNNIIVRFGCPRLVSDNGPQFISDKFATFCEEYG
IQQRTSSVYHPQTNGQAEASNKIILHGLRRNLDLSLGGSWPDQLPHVLWAYRTTPKSSTGE
TPFSLVYGSEAVAPVESTIITPRIAAYMHTESANTEFRELDLDLLEERRNEVYGRVRKQQ
RALRKRYNQVRPRQFEKGDLILRSVESQGHKGKLDRAWEGPY
>scaffold146.1|size86774:14592-14828 PR0T Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Athila
MVDLGASINLMPYSIYSALQLGPLQGTAIIVIKLADRSNTHPEGVIEDVLVQVNNLVFPAD
FYVLKMGKAENNDCLLLG
>scaffold146.1|size86774:15420-15995 RT Class_I|LTR|Ty3/gypsy|non-chromovirus|OTA|Athila
IYAISDSDWSPVHVVPKKTGFTVERNKNLGVLPKRVTVNGWRVCIDYRKLNDATRKDHFP
LPFIDQMLERLAGKKFYCFLDGYSQYNQVAIAPEDQEKTFTCTYGTAFRKMPPFGLCNA
PATFQRCMLSIFSEFTGKFIEVFMDDFTVYGDSFEGALENLEKVLQRCVEKKLVLNSEKC
HFMVROGIVLGH
```

• Domains classifications table

1	2	3
DOMAINS_REPETITIVE_CLASSIFICATION	ORIGINAL_FILE_COUNTS	FILTERED_FILE_COUNTS
Ambiguous_domain	1	0
Class_I LTR Ty1/copia Bianca	6	4
Class_I LTR Ty3/gypsy non-chromovirus OTA Athila	8	5
Class_I LTR Ty3/gypsy non-chromovirus OTA Ogre/Tat	3	2
Class_I LTR Ty3/gypsy non-chromovirus OTA Ogre/Tat TatV	4	2

Filtering Options

- Type of domain
- Repetitive element classification
 - ♦ Custom string
- Quality of alignment
 - ♦ Identity
 - ♦ Similarity
 - ♦ Relative length
 - ♦ Interruptions

Galaxy implementation

Protein Domains Finder Find and classify all protein domains in your DNA sequence based on the database of repetitive elements

Options

protein domains (Galaxy Version 1.0.0)

Choose your input sequence

344: chr01_banana

Input DNA must be in proper fasta format, multi-fasta containing more sequences is allowed

Execute

Protein Domains Filter Filter Protein Domains Finder output according to a certain domain type and several alignment quality criteria

Options

(Galaxy Version 1.0.0)

Choose primary GFF3 file of all domains from Protein Domains Finder

345: GFF3 file of all protein domains from dataset 344

Minimum identity

0.35

Protein sequence identity threshold between input and mapped protein from db [0-1]

non-Viridiplantae: 0.3

Minimum similarity

0.45

Protein sequence similarity threshold between input and mapped protein from db [0-1]

non-Viridiplantae: 0.4

Minimum alignment length

0.8

Proportion of the hit length without gaps to the length of the database sequence [0-1]

Interruptions [frameshifts + stop codons]

3

tolerance threshold per every starting 100 amino acids of alignment sequence

Select protein domain type

All

GAG, INT, PROT, RH, RT, aRH,
CHDCR, CHDII, Tpase, YR,
HEL1, HEL2, ENDO

Filter a custom classification substring

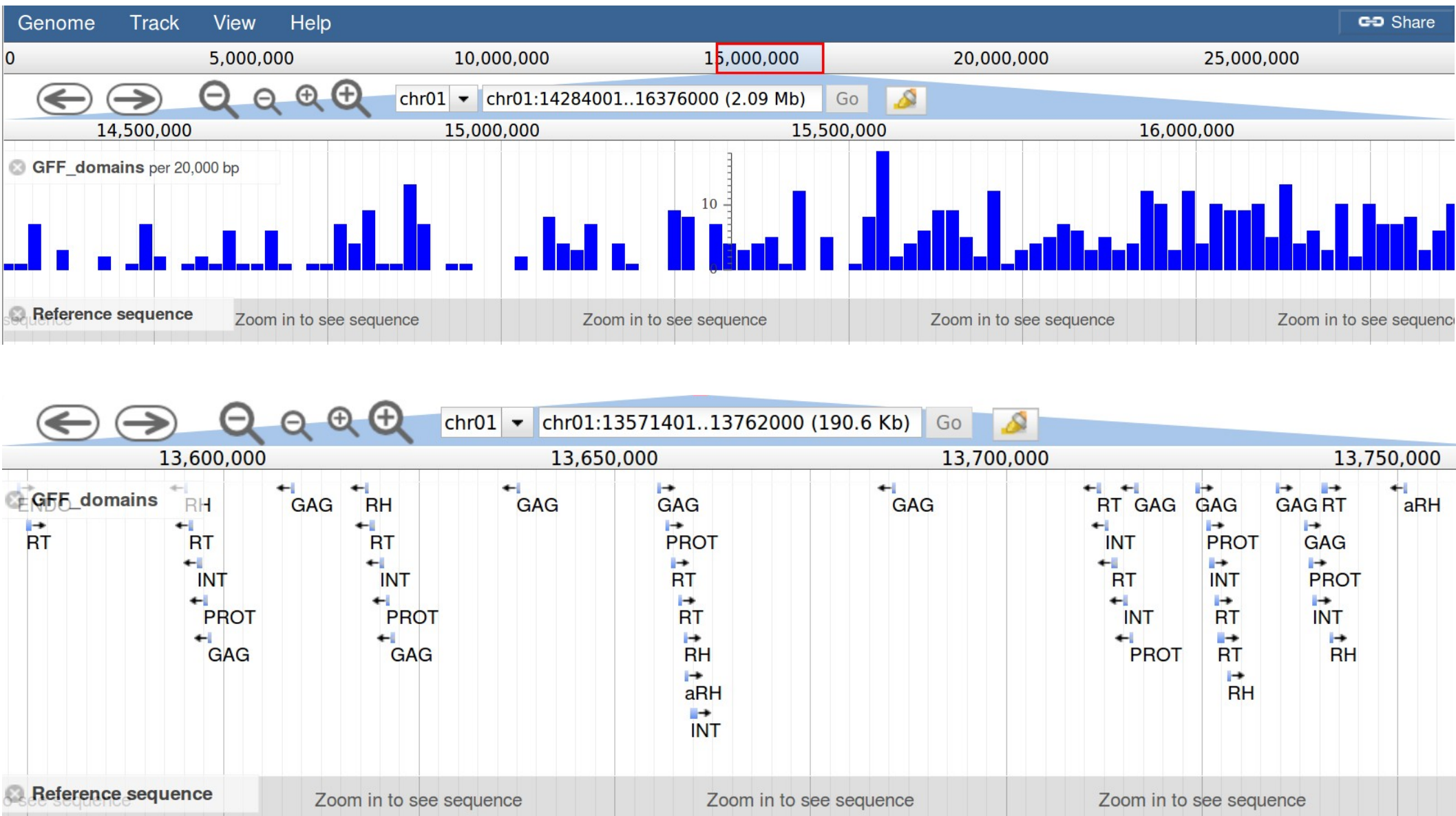
If empty, all classifications will be reported

You can type in an arbitrary string to filter a certain repetitive element type of any level. It must be a continuous substring in a proper format of Final_Classification attribute of GFF3 file. Classification levels are separated by | character

LTR, Ty1/copia|Bianca,
Ty3/gypsy|non-chromovirus,
Class_II

Execute

Results visualization



5,000,000
10,000,000
15,000,000
20,000,000
25,000,000

←
→
⊖
⊕

Go

125
13,637,450
13,637,475
13,637,500
13,637,525

GFF_domains

Reference sequence

protein_domain GAG
✕

Primary Data

Name	GAG
Type	protein_domain
Score	240
Position	chr01:13637485..13637748 (- strand)
Length	264 bp

Attributes

Best_hit	Ty1-GAG_REXdb_ID4602 Class_I LTR Ty1/copia SIRE:13637485-13637748[100%]
Db_seq	DRNEYNRICQCKSAKXIXRLLEITHKGTSRVKEFKFNLLIYSFKLFRMKQSKTIGDMYTRFMDVINSLKALGKCFNSLELI
Final_classification	Class_I LTR Ty1/copia SIRE
Identity	0.51
Query_seq	EKNEFNRVSI CETPYN I*HILEITHKGTSRVKEFKFNLLIYSFKLFRMKQSKTIGDMYTRFMDVINSLKALGKCFNSLELI
Region_hits_classifications	GAG Class_I LTR Ty1/copia SIRE
Relat_interruptions	1.0
Relat_length	0.946
Seq_id	chr01
Similarity	0.75
Source	profrep

Region sequence

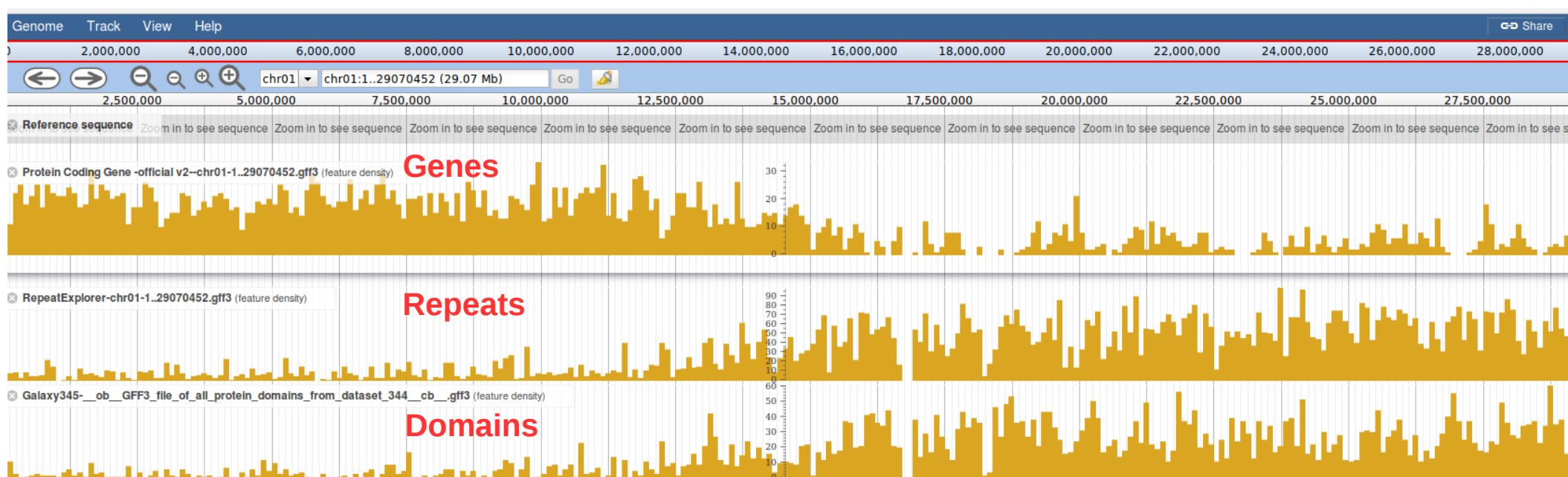
FASTA

```

>chr01 chr01:13637485..13637748 (- strand)
class=protein_domain length=264
GAAAAAATGAGTTTAATCGTGTTCATATGTGAAACTCCATATAATATTTAACACATACT
    
```

Results visualization

JBrowse comparison of different tracks



Musa acuminata assembly(2016) - chr01

Results visualization

Comparison of repeats and domains annotation

