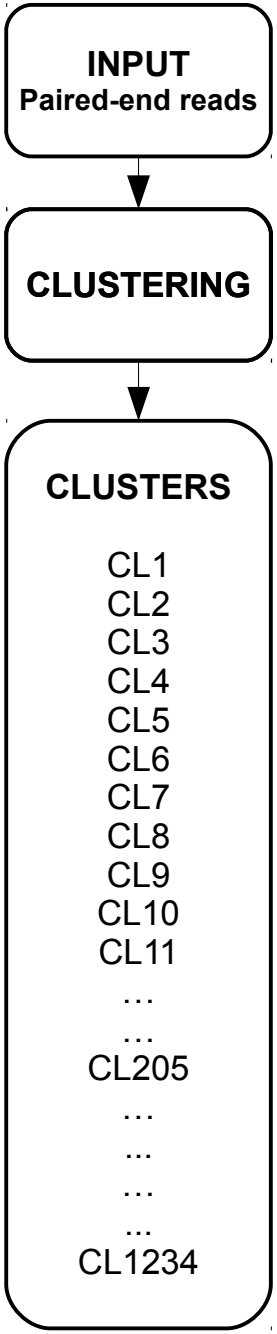
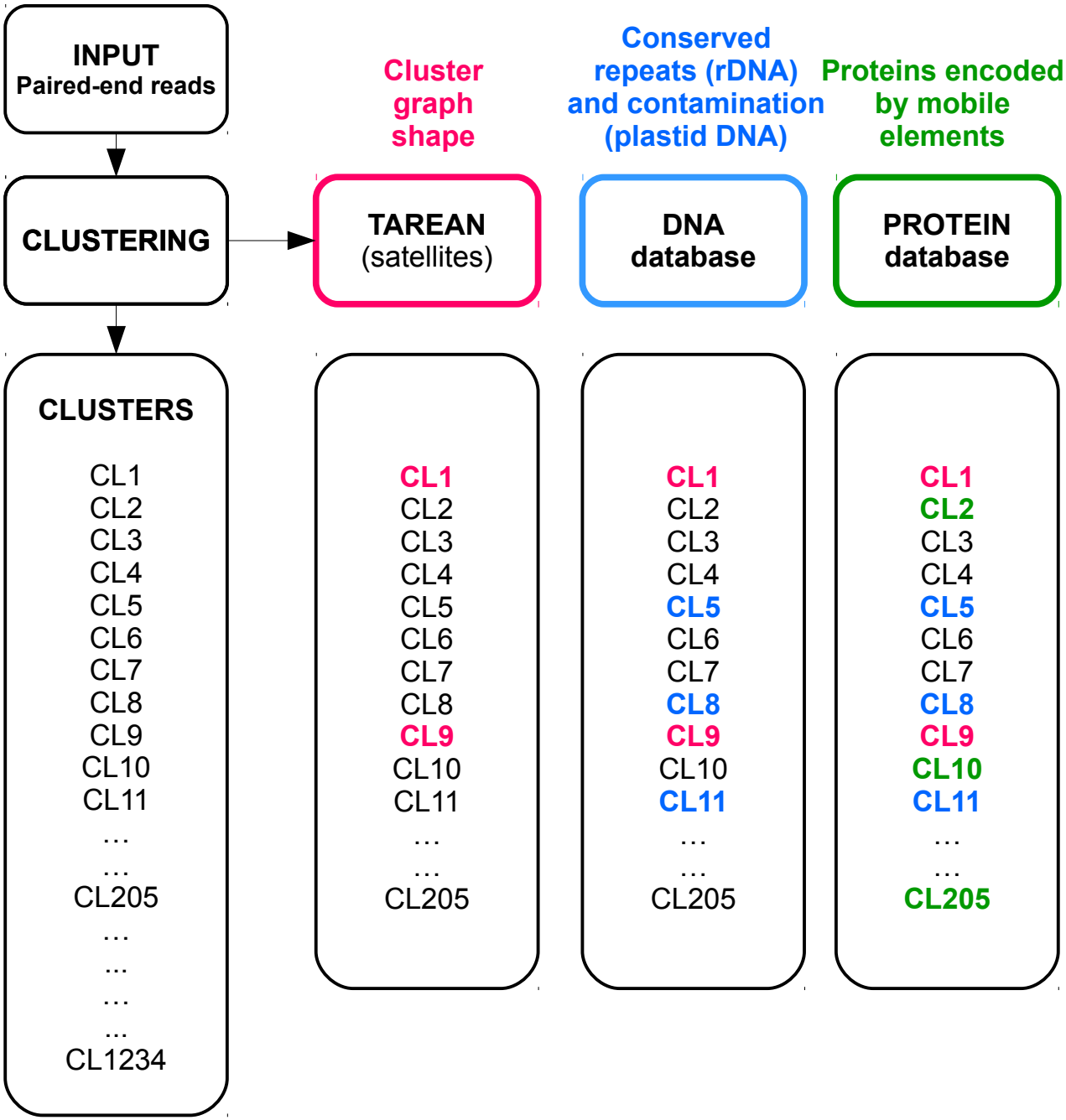
Using *RepeatExplorer* output for repeat annotation and quantification

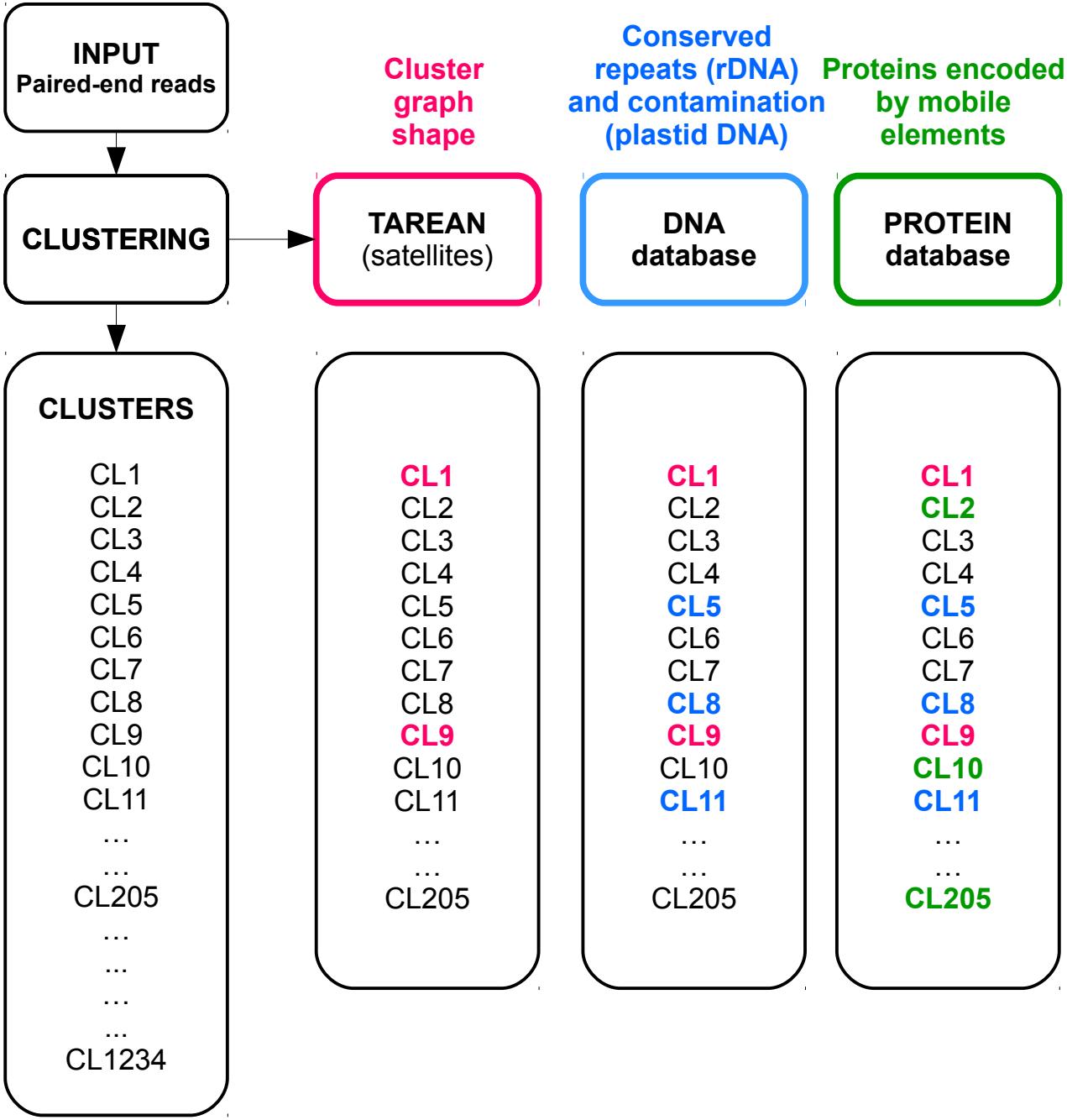
RepeatExplorer ver. 2



RepeatExplorer ver. 2



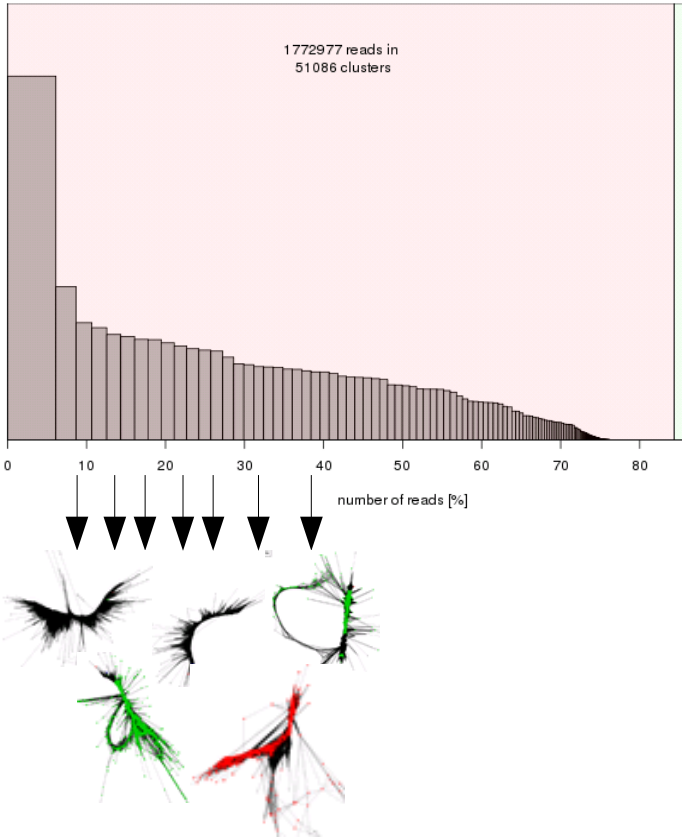
RepeatExplorer ver. 2



Many repeats are split to multiple clusters

(feature of the clustering algorithm; depends on many factors)

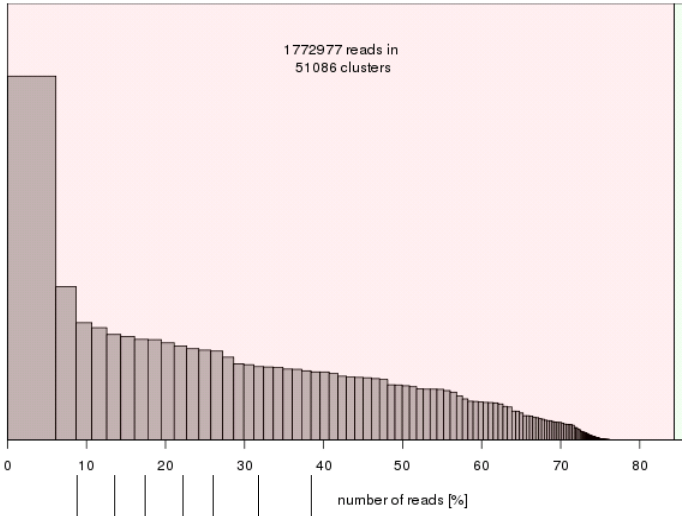
Cluster fragmentation

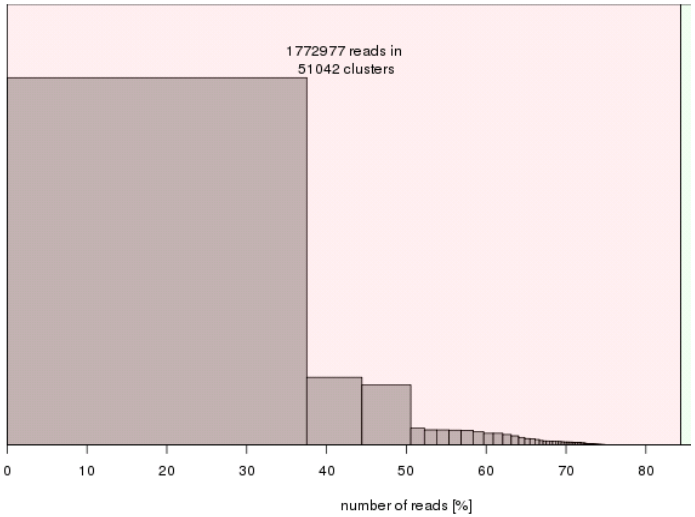


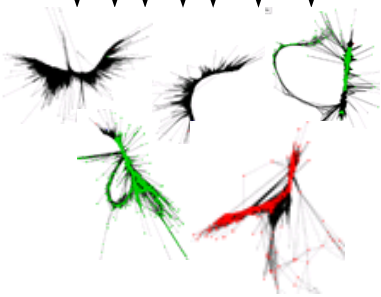
LTR-retrotransposon Ogre in *Vicia pannonica* (~ 40% of the genome)

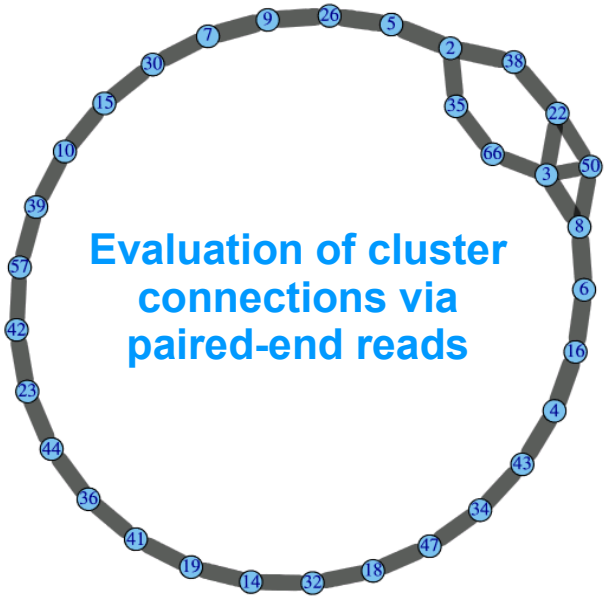
Cluster fragmentation

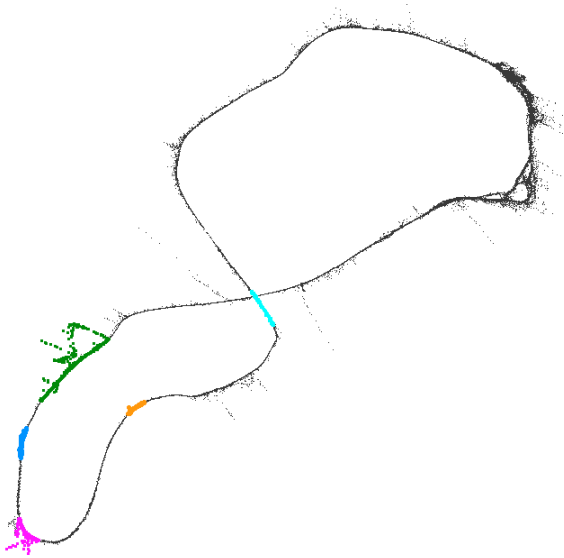
Repeats can be split to multiple clusters !







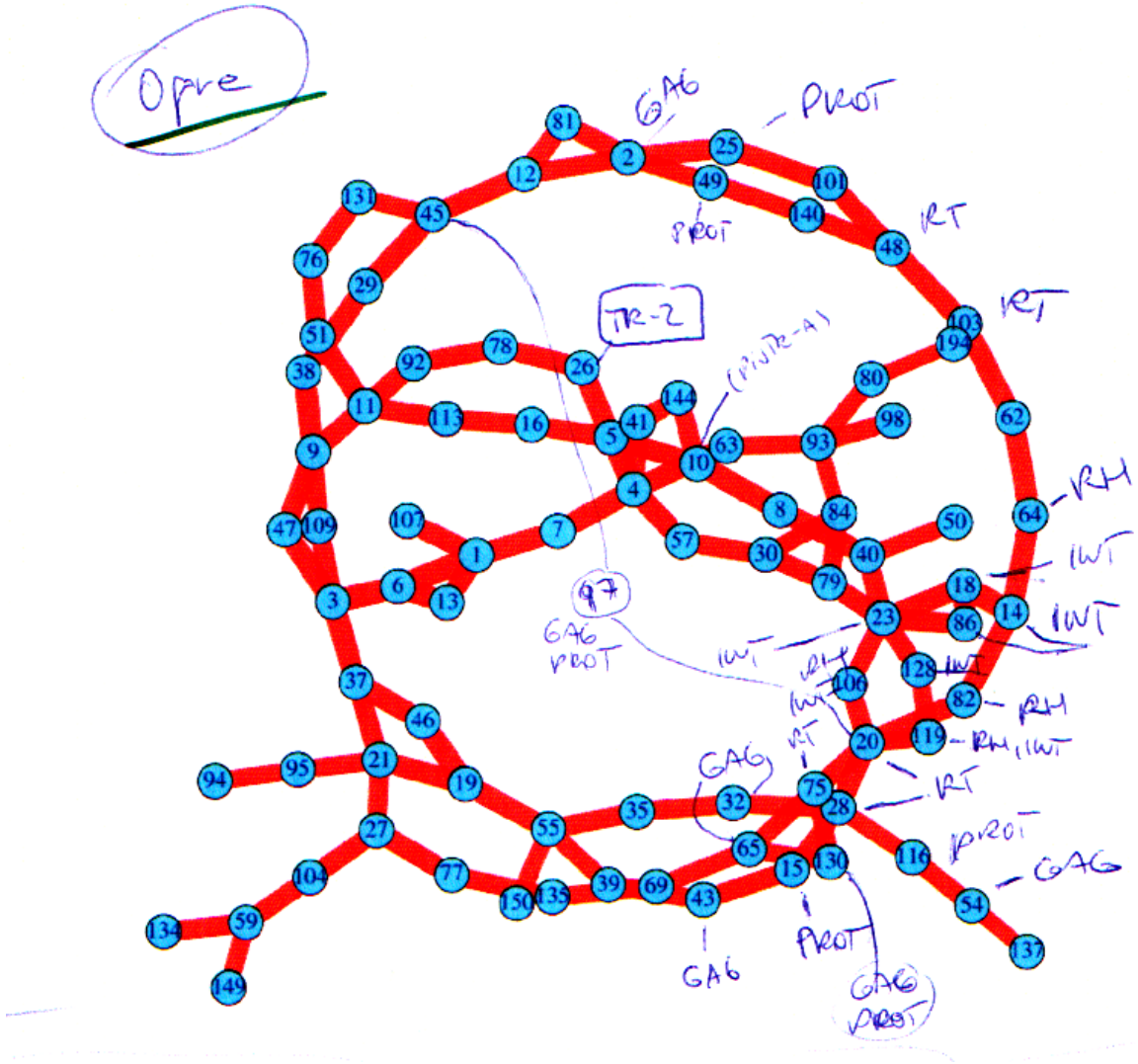




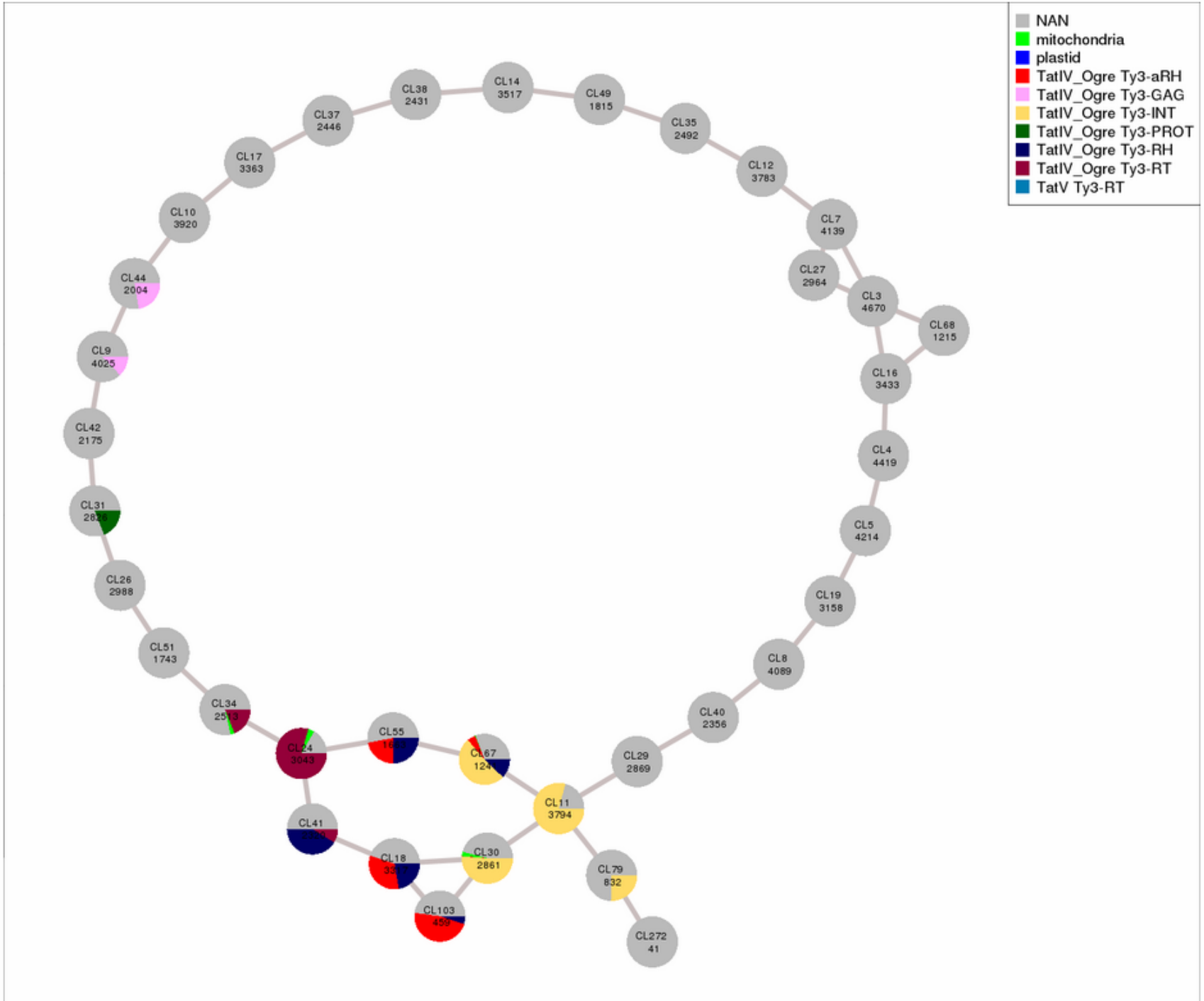
LTR-retrotransposon Ogre in *Vicia pannonica* (~ 40% of the genome)

Superclusters in the past...

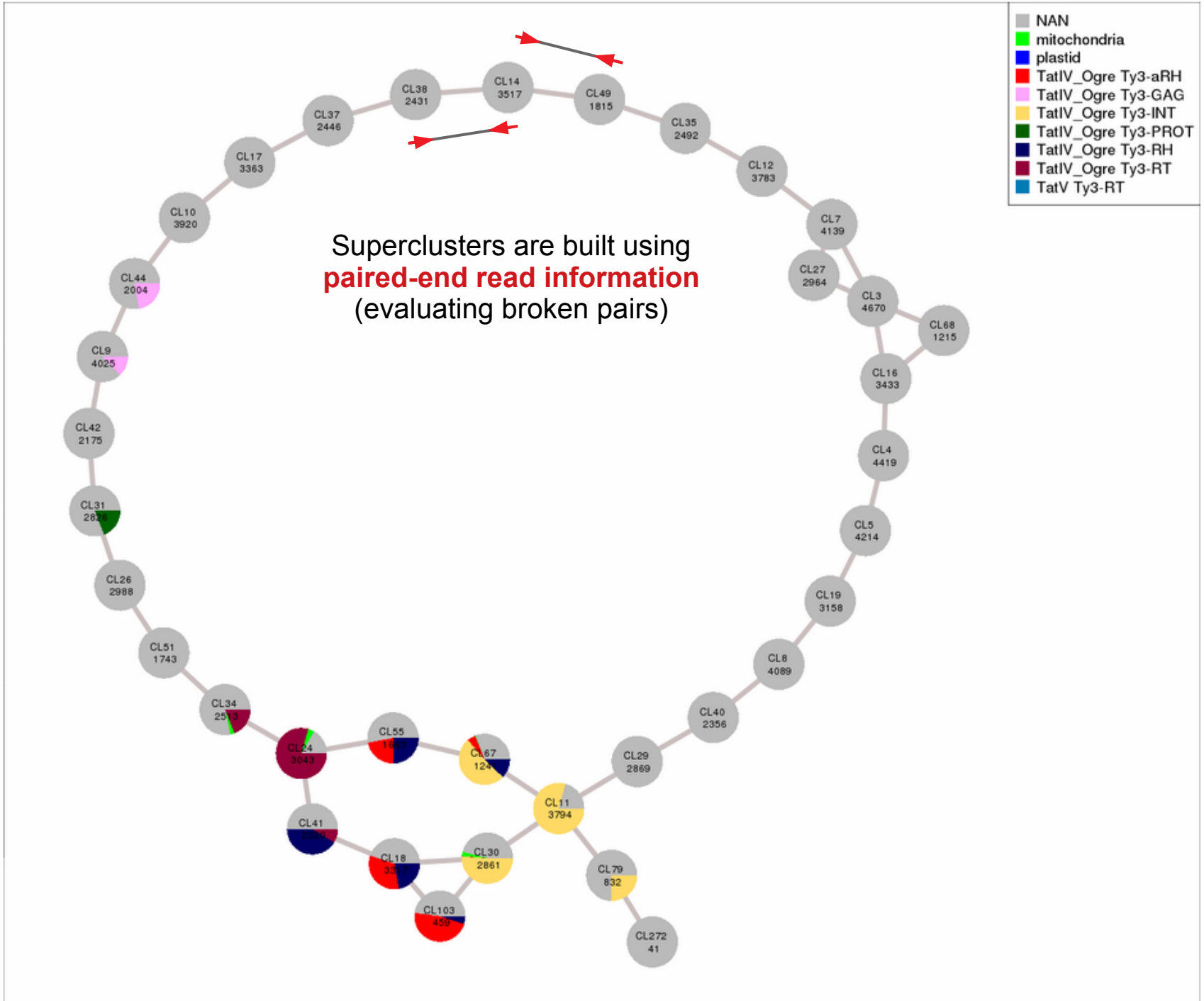
PFL 0.1



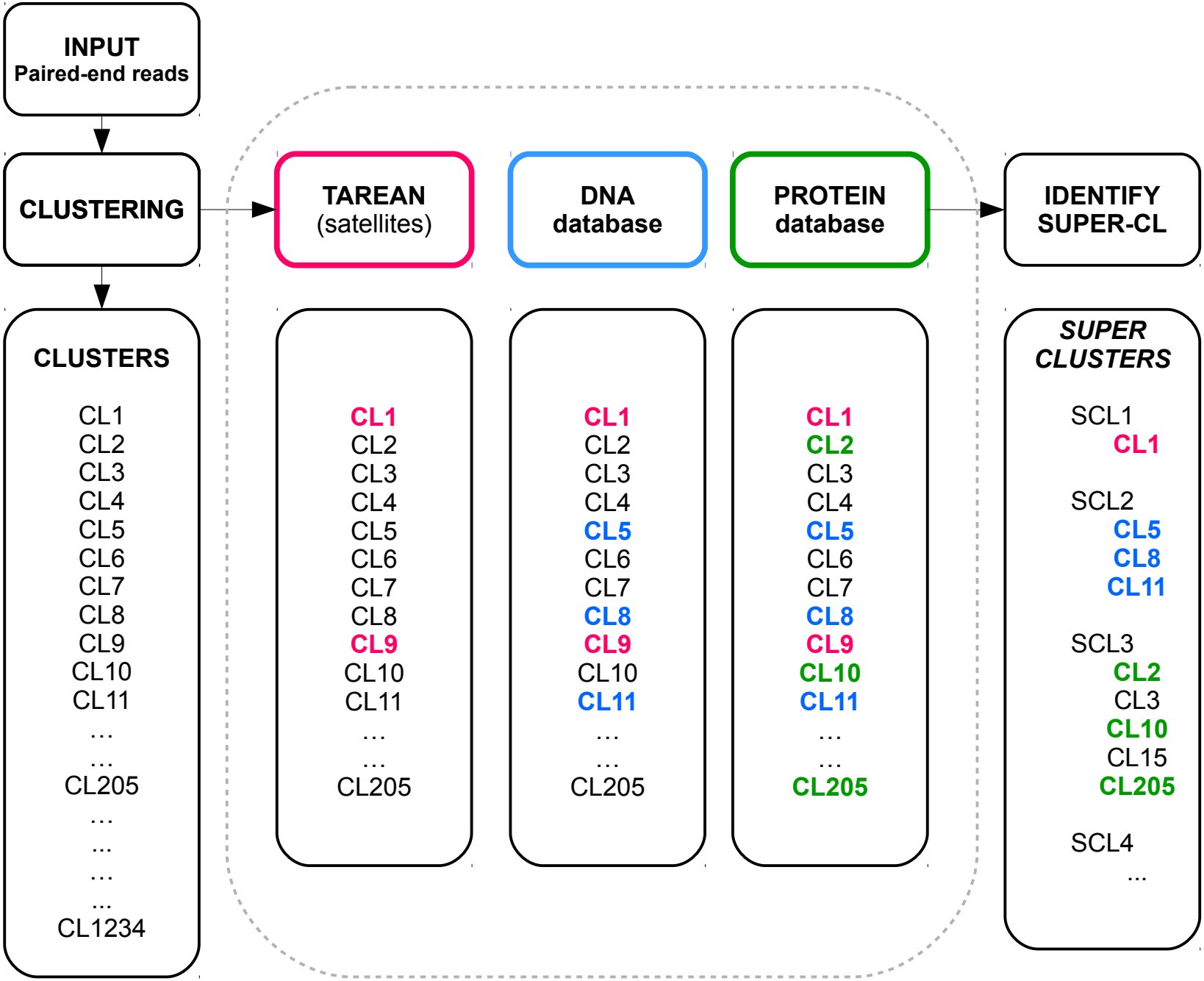
...and now in RepeatExplorer 2



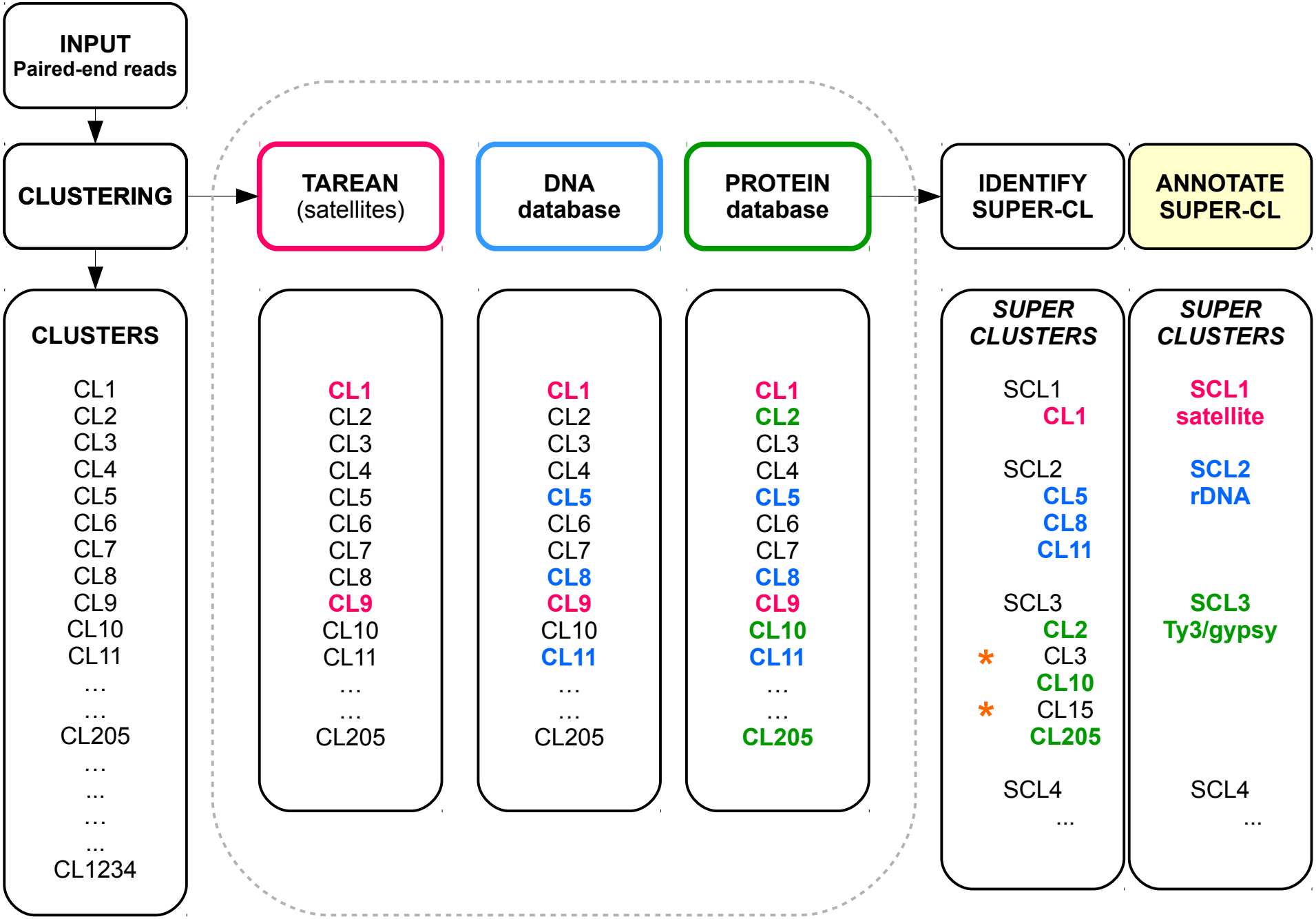
...and now in RepeatExplorer 2



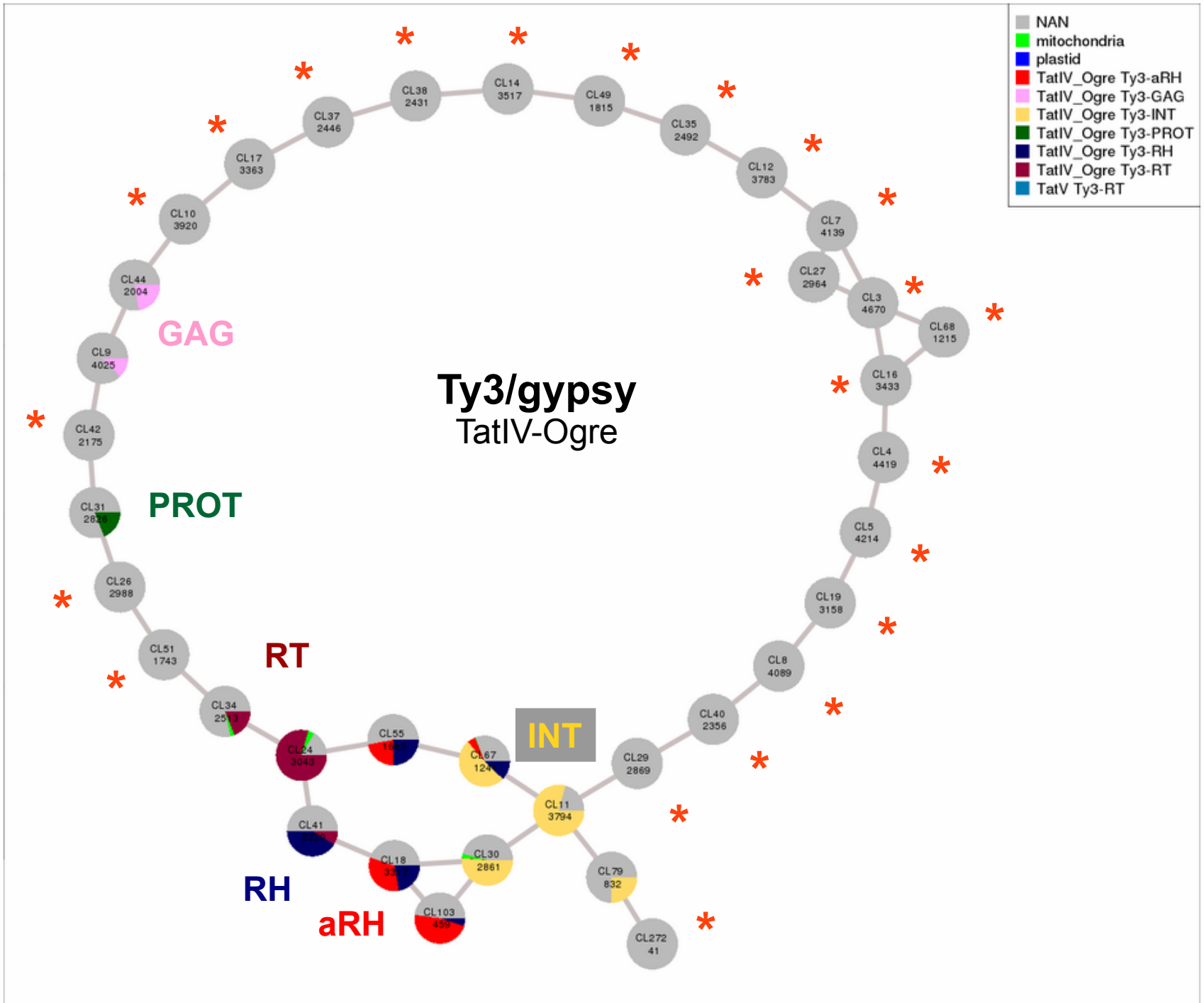
RepeatExplorer ver. 2



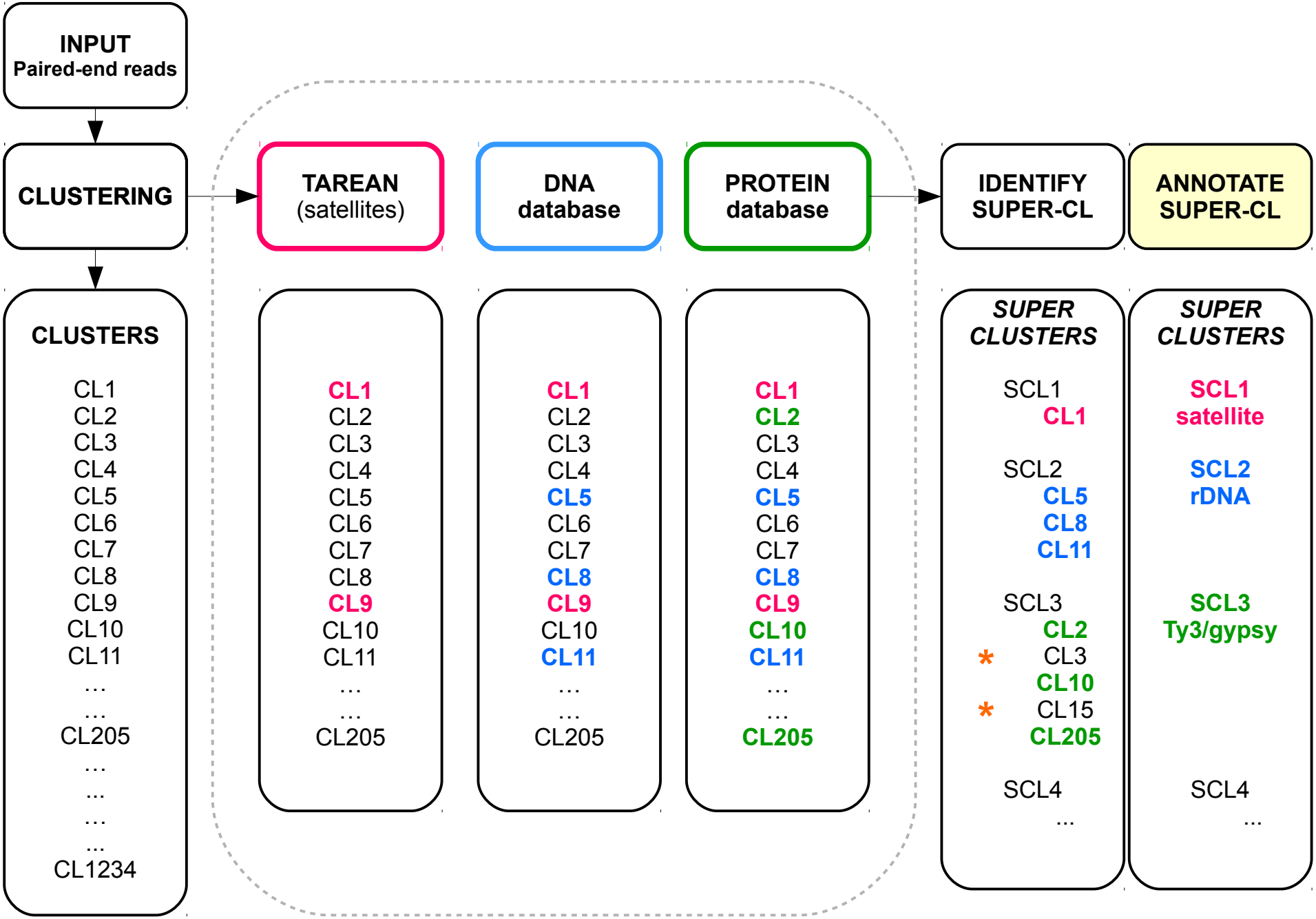
RepeatExplorer ver. 2



Superclusters provide more complete annotation



RepeatExplorer ver. 2



Repeat quantification

Cluster sizes (numbers of reads) are proportional to genomic abundance of corresponding repeats

	7192836 (= 100%)	64.1
CL	**reads**	**genome %**
1	304159	4.229
2	234749	3.264
3	216307	3.007
4	202822	2.820
5	149693	2.081
6	145911	2.029
7	143766	1.999
8	142608	1.983
9	141836	1.972
10	123886	1.722
11	79345	1.103
12	72781	1.012
13	67096	0.933
14	65455	0.910
15	62334	0.867
16	53845	0.749
17	49341	0.686
18	45062	0.626
19	44762	0.622
20	43332	0.602
21	42344	0.589
22	40125	0.558
23	39923	0.555
24	36353	0.505
25	35977	0.500
26	35674	0.496
27	34829	0.484
28	34534	0.480
29	34302	0.477
30	33114	0.460
31	32930	0.458

IDENTIFY SUPER-CL → ***SUPER CLUSTERS***

SCL1
CL1

SCL2
CL5
CL8
CL11

SCL3
CL2
CL3
CL10
CL15
CL205

SCL4
...

ANNOTATE SUPER-CL → ***SUPER CLUSTERS***

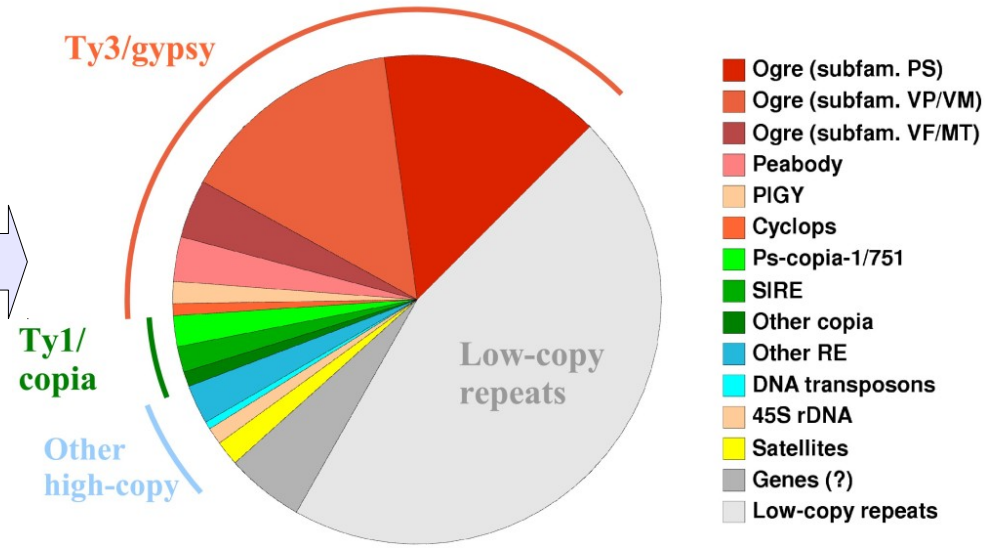
SCL1
satellite

SCL2
rDNA

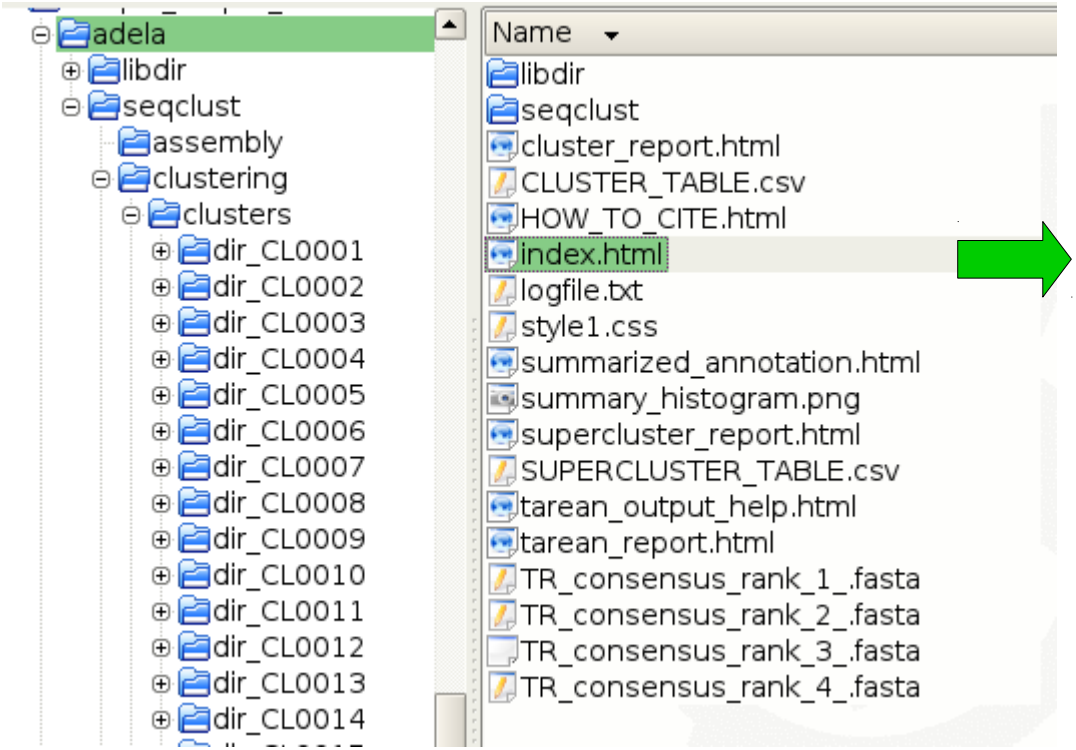
SCL3
Ty3/gypsy

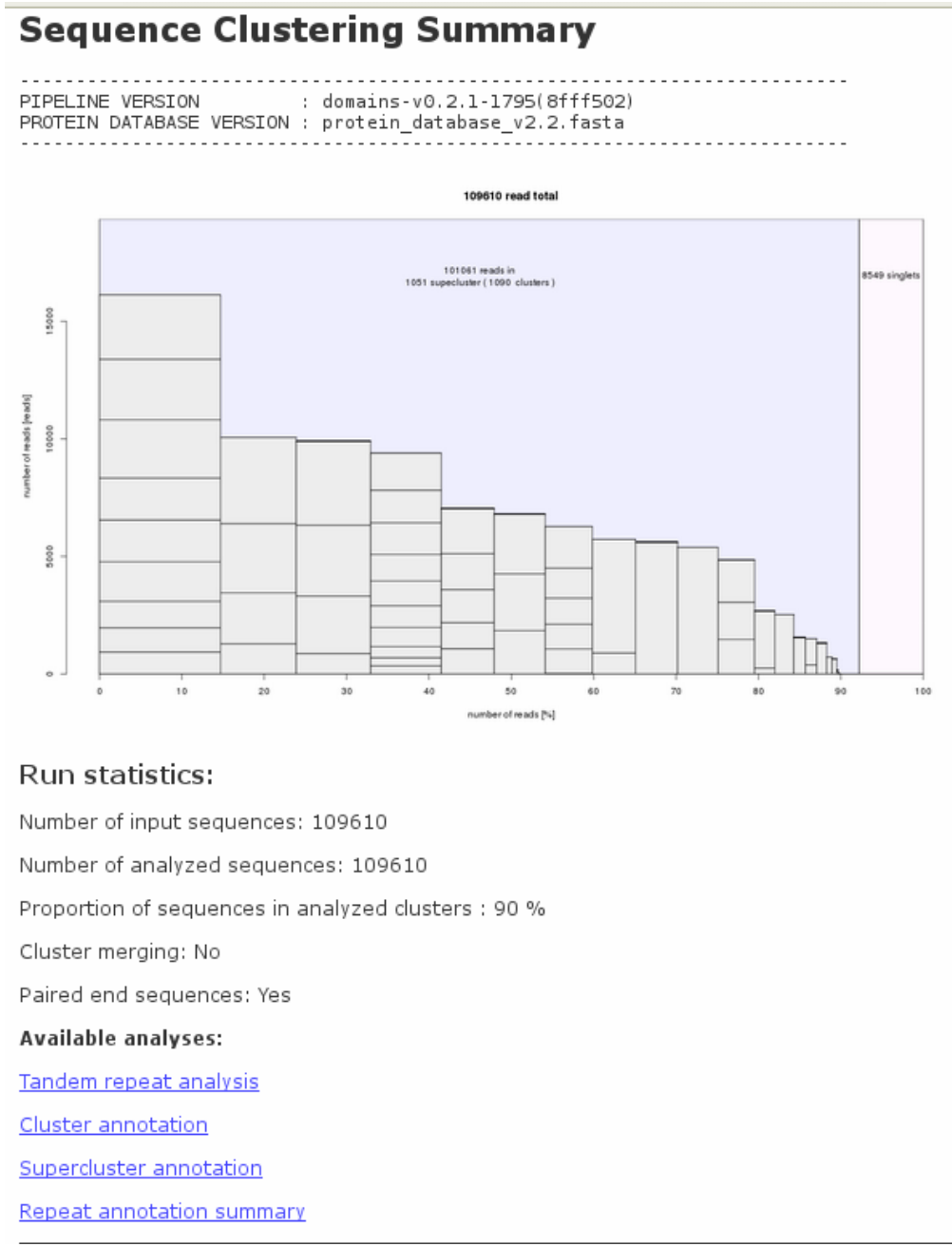
SCL4
...

Proportions of various repeat types in a genome



RepeatExplorer output files



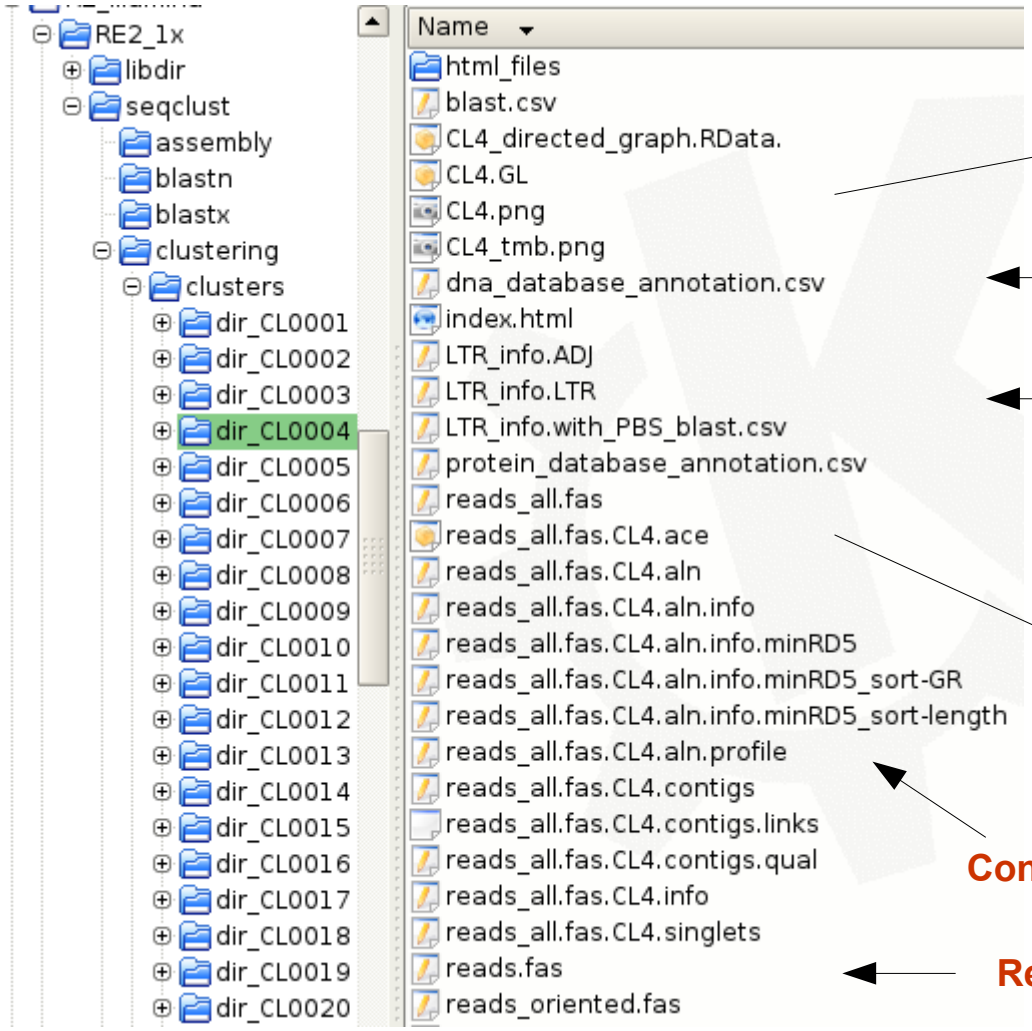


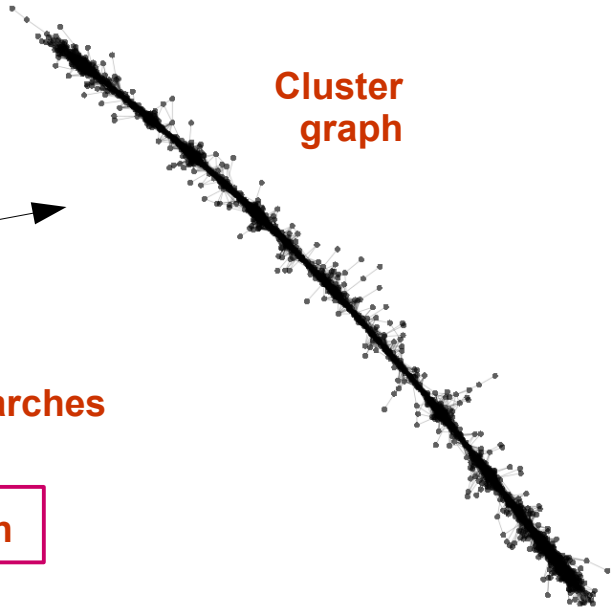
Always download and use archive with complete output of the analysis

- do not work with HTML only
- do not trust automatic annotation

RepeatExplorer output files

Cluster directories





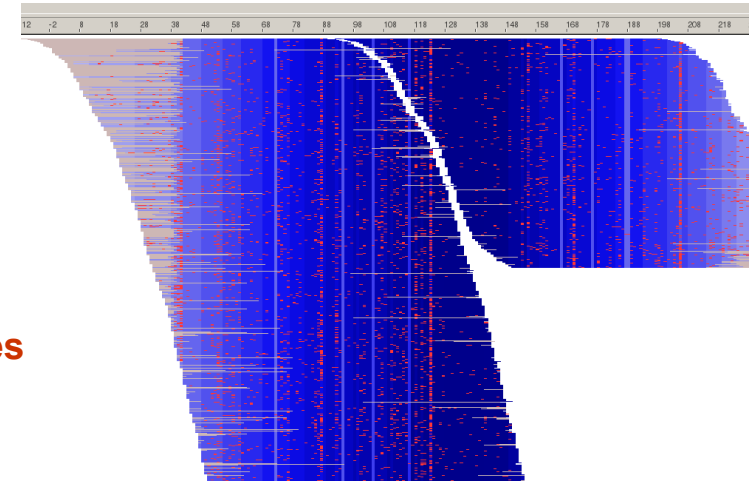
Cluster graph

Similarity searches

LTR-detection

Contigs

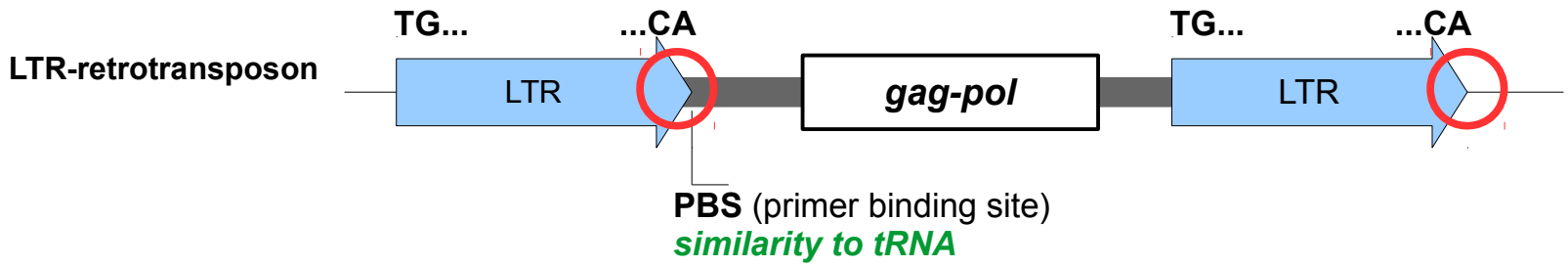
Read sequences



assembly

Insertion sites of mobile elements

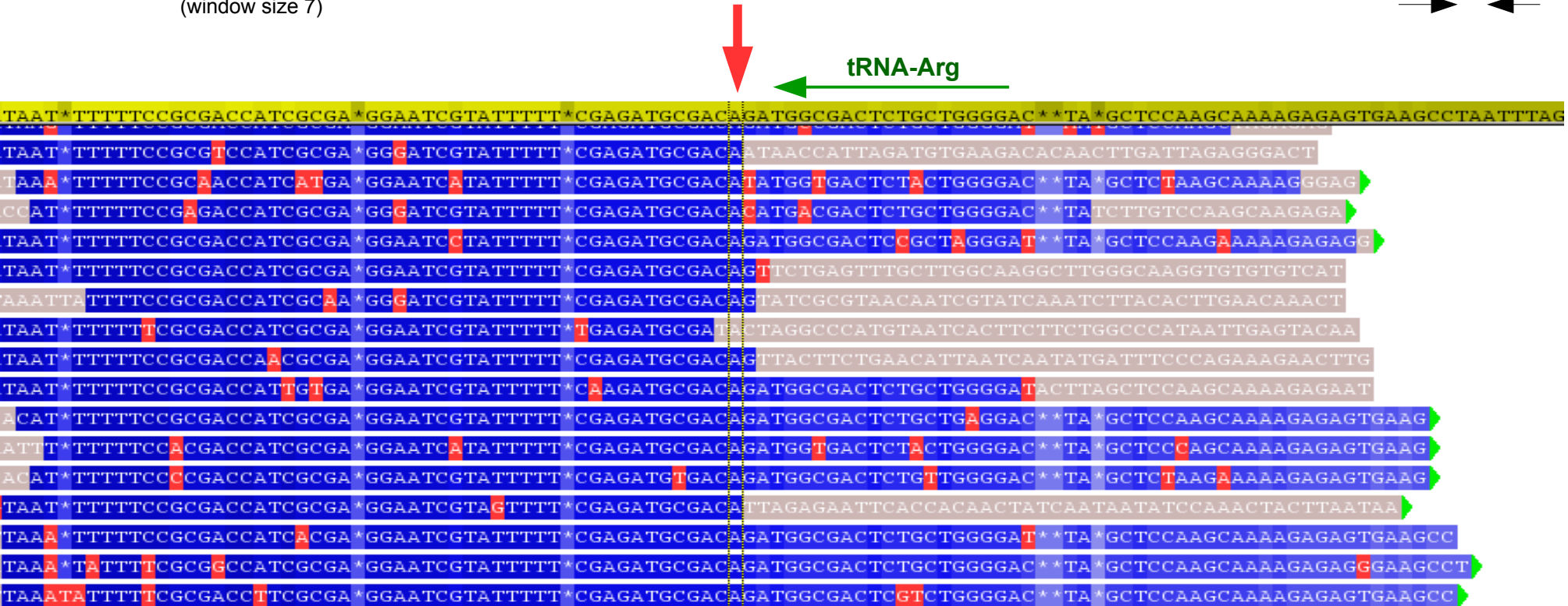
A tool for detection of LTR / PBS sites



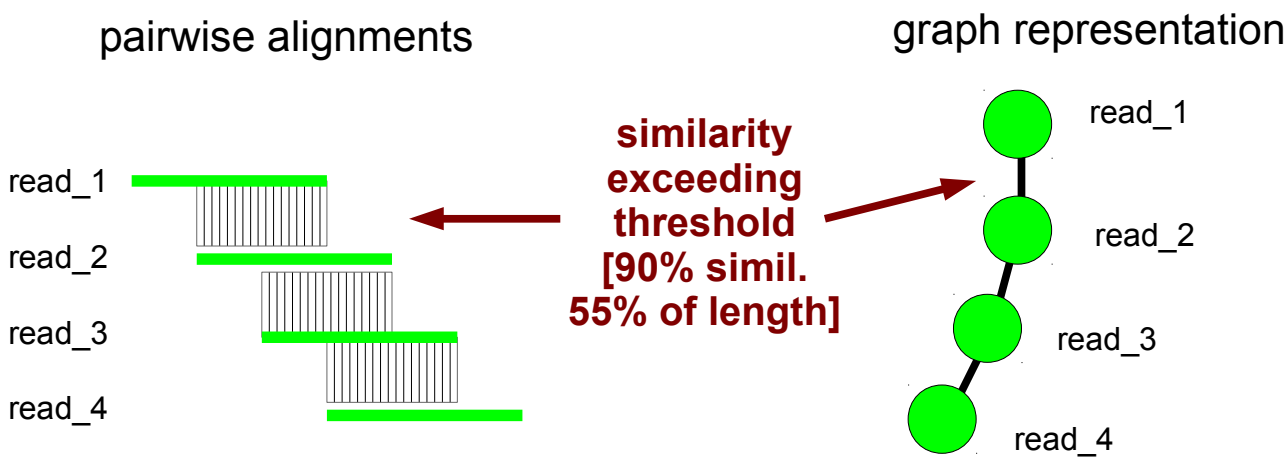
Program output:

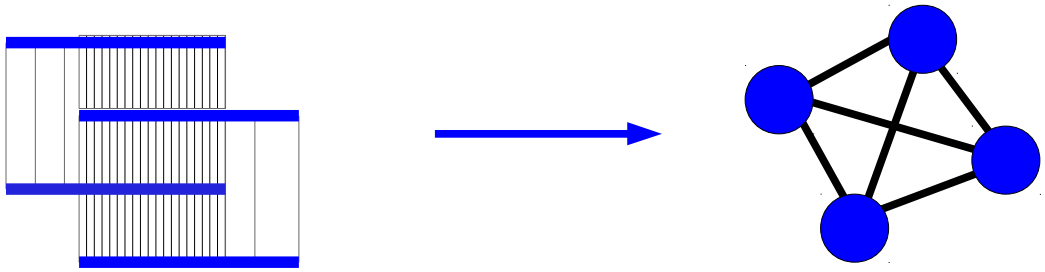
CL	contig	pos.	site	site_depth	out_ma	maske	maske	region_in	region_out	blast to tRNA	%	length			site from	site to	tRNA from	tRNA to	E-val
19	400	364	TGCGA**CA**	106.6	30.4	0.0362	0.2975	GCGAGGAA	GATGGCGA	**At-chr2.tRNA28-Arg**	100	18	0	0	3	20	23	6	7E-007

(window size 7)

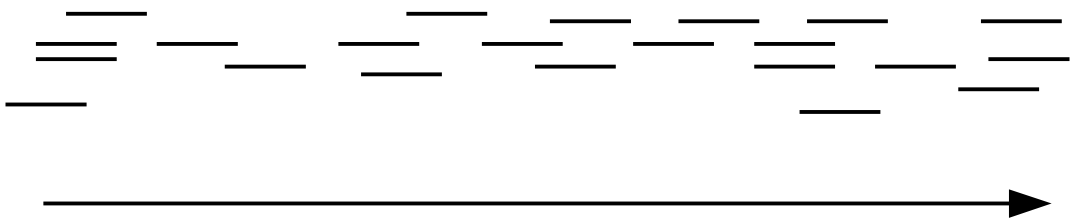


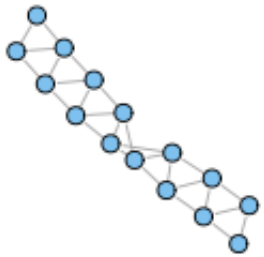
Graph shapes





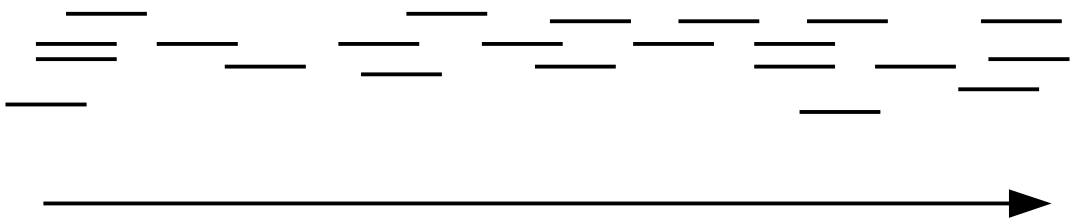
Linear graphs

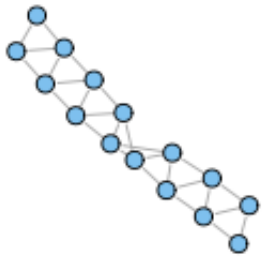




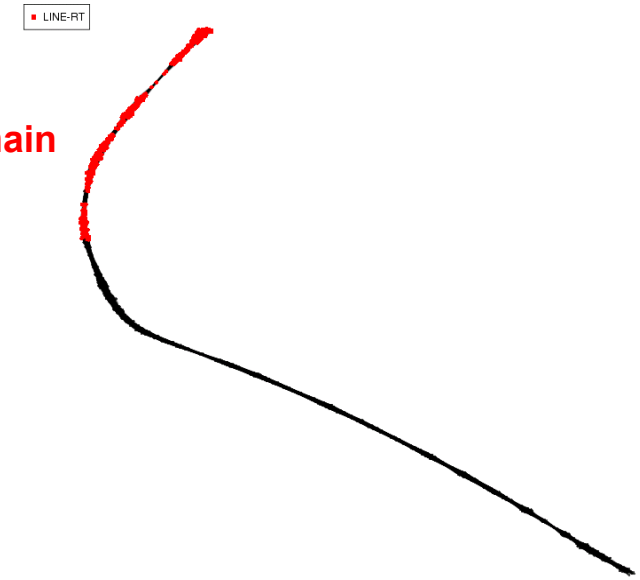
DNA transposons, LINEs, ...

Linear graphs

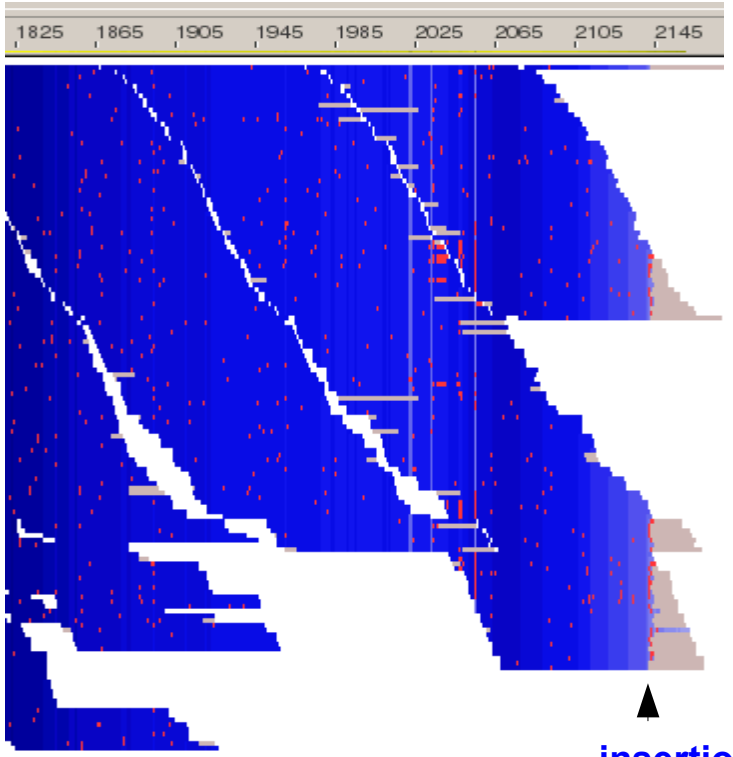




DNA transposons, LINEs, ...



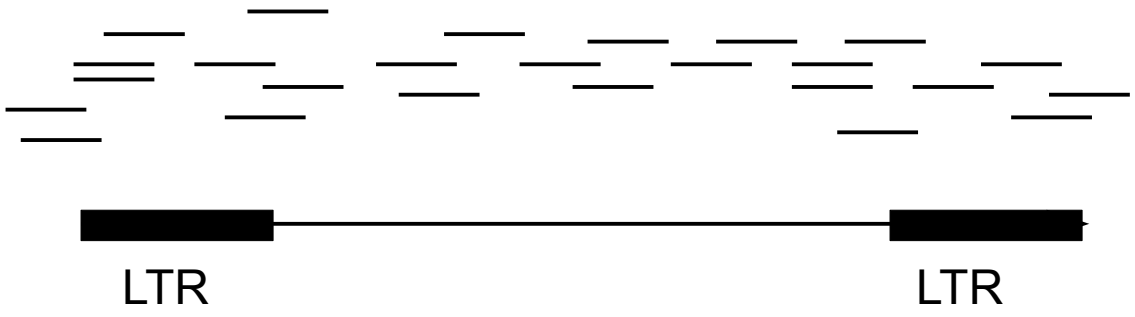
RT domain (LINE)

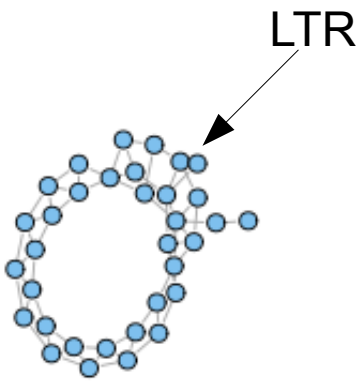


assembled contig

insertion sites

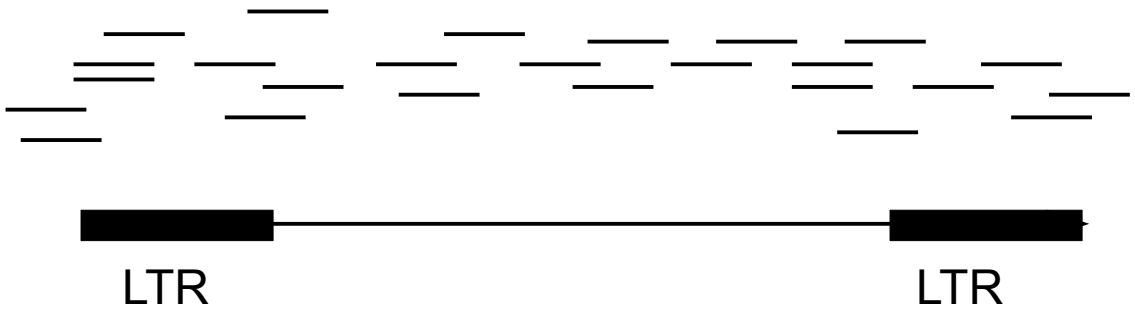
Linear / large circular graphs

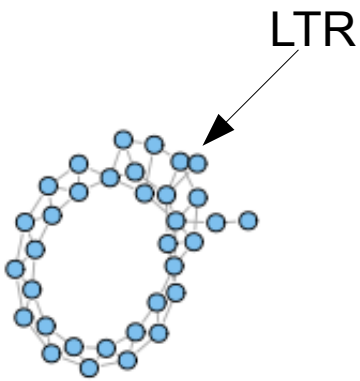




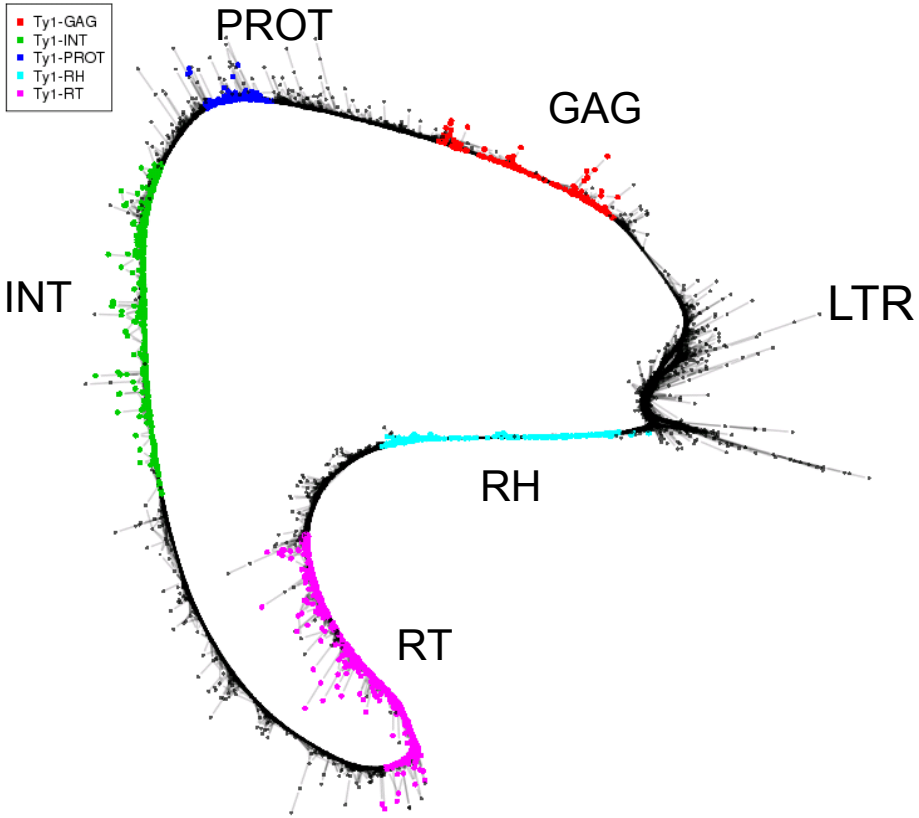
LTR-retrotransposons

Linear / large circular graphs

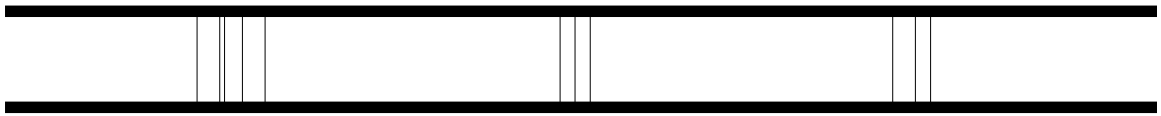


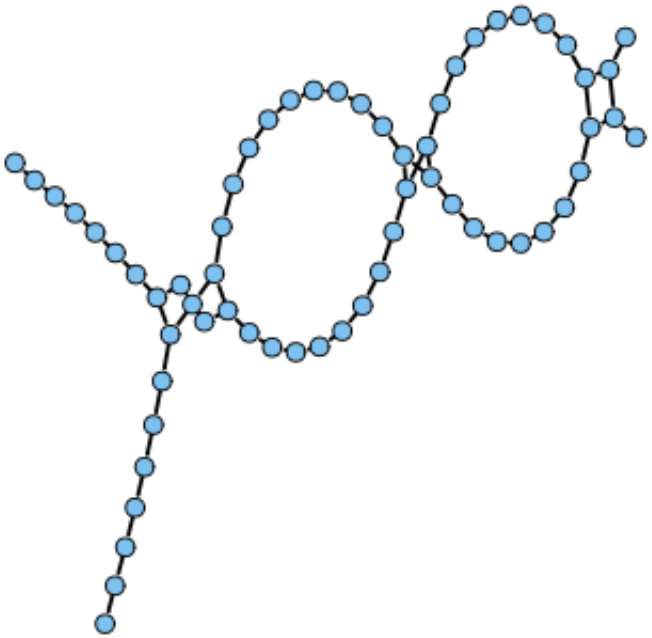


LTR-retrotransposons

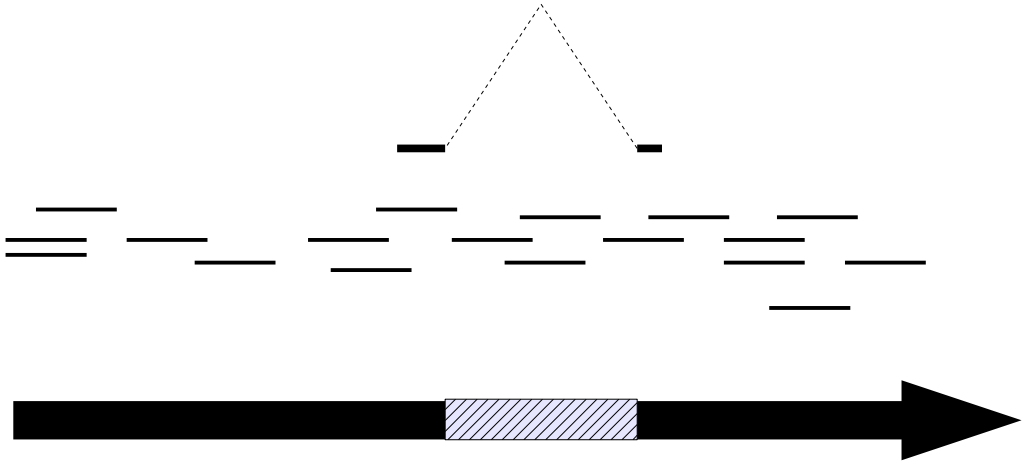


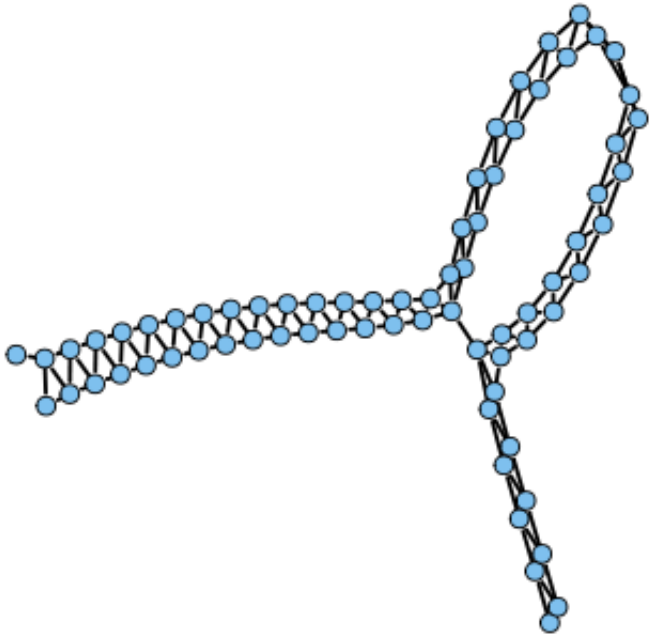
It's getting complicated...





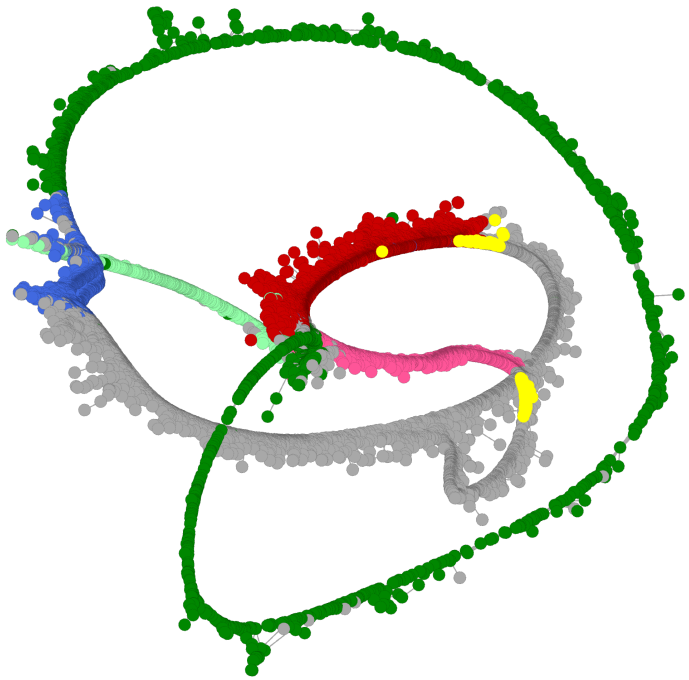
Repeat variants with partial similarity
(depends on similarity threshold)

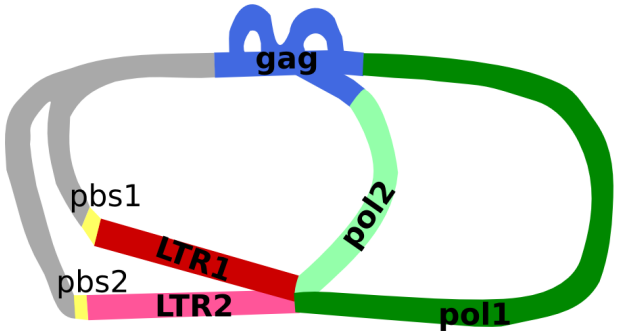




Variants differing in indel

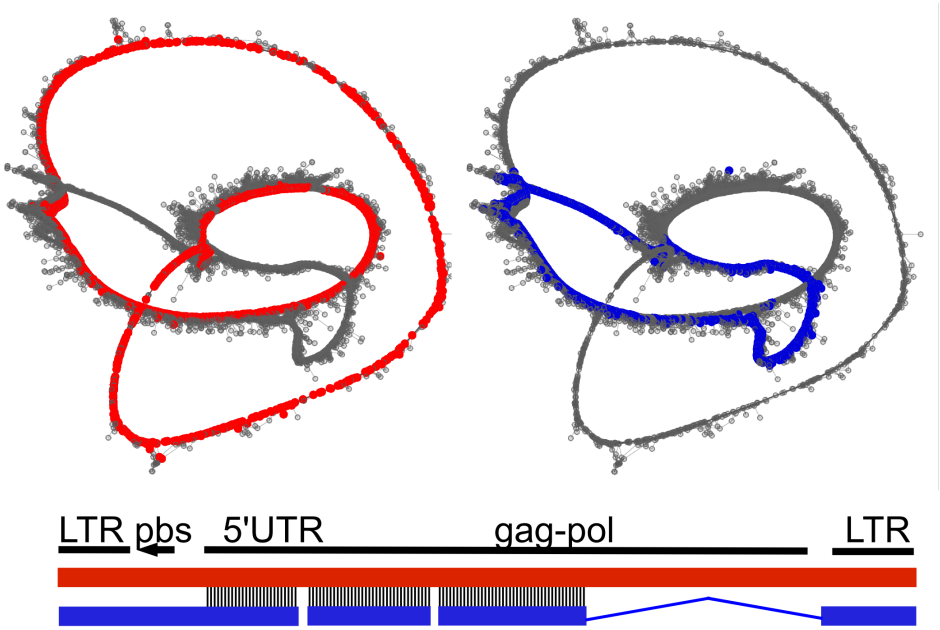
It's getting complicated...





...but it has some meaning

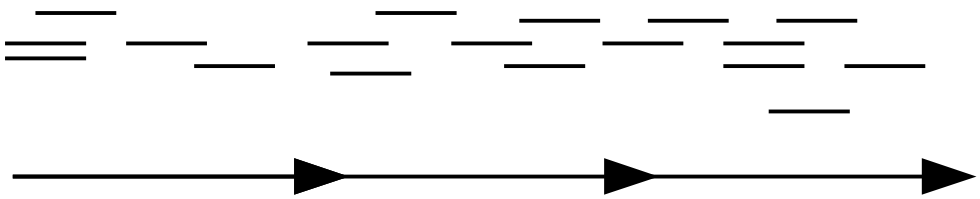
(in this case, presence of element variants differing in LTR sequence and in deletion within gag-pol region)

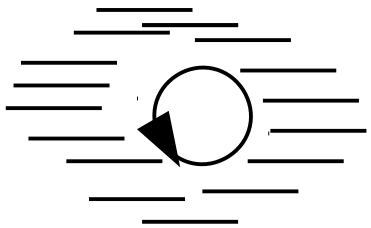


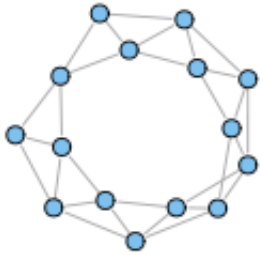
Circular graphs

TANDEM REPEATS

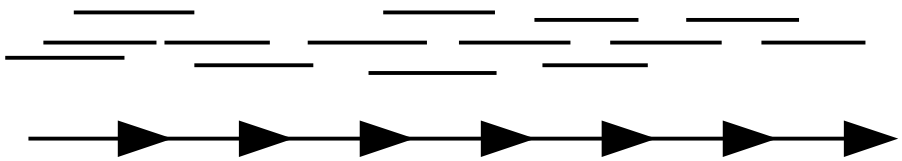
Read length << monomer

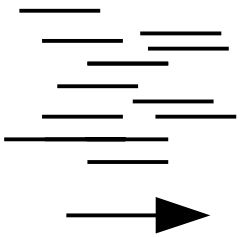


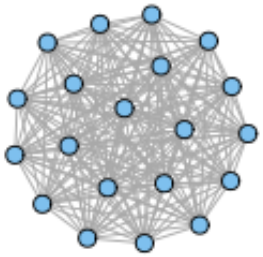




Read length >= monomer

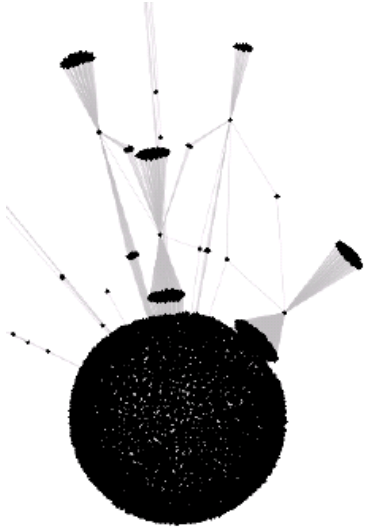


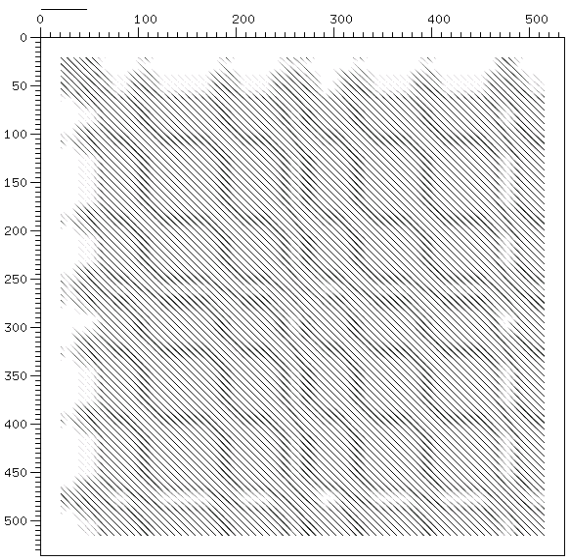




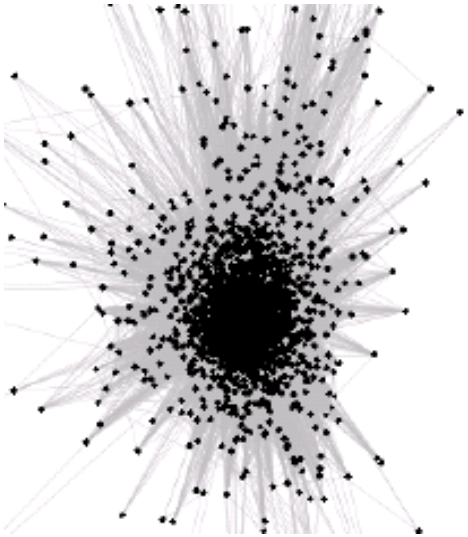
Circular graphs

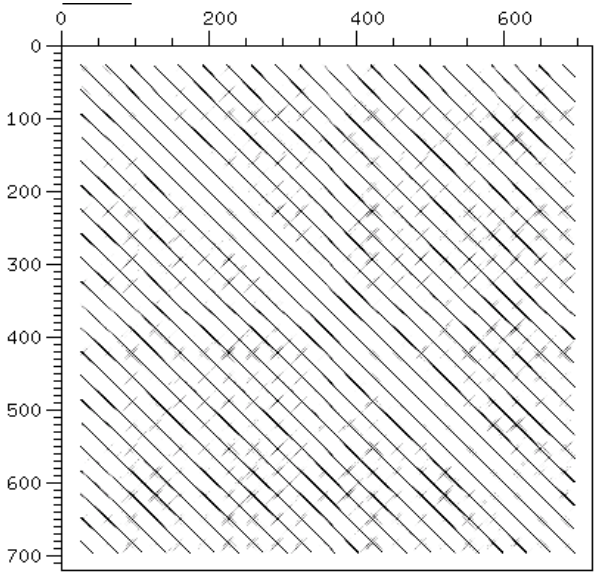
microsatellite (6bp)
(GAACCT)n



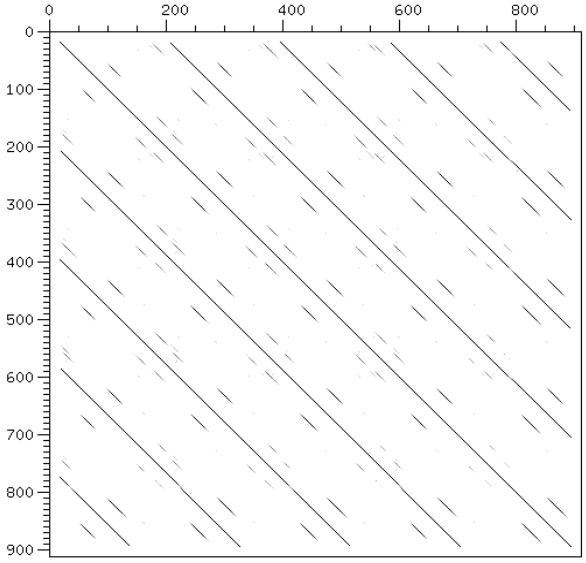


satellite 35 bp



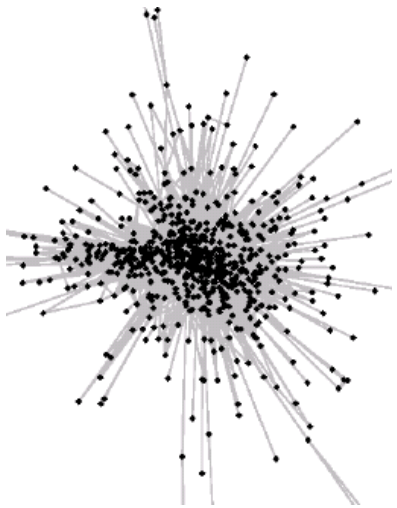


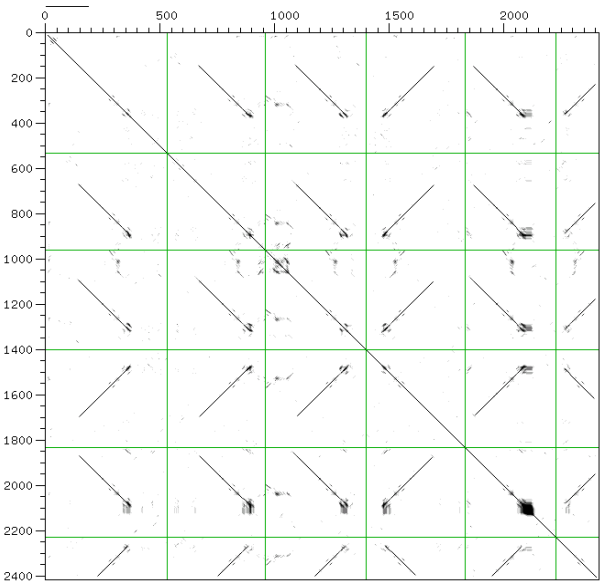
satellite 190 bp



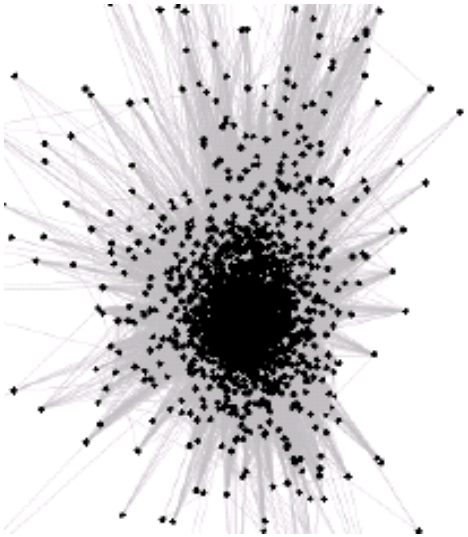
Circular graphs

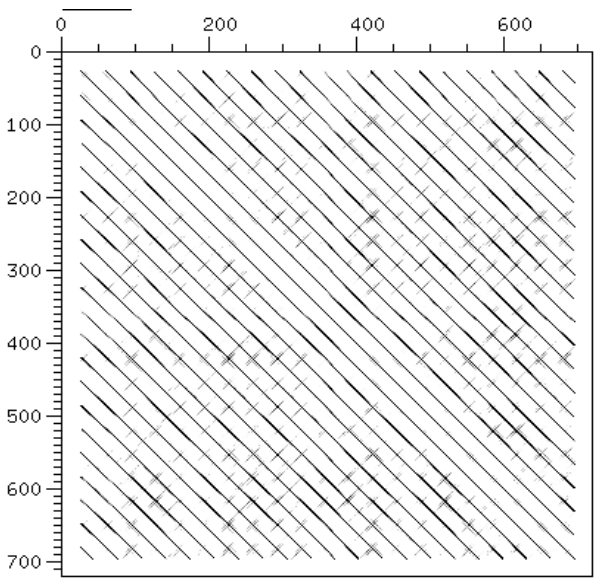
SINE



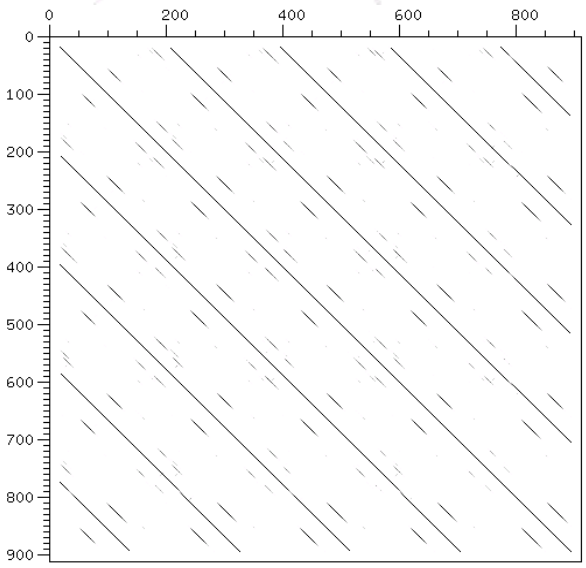


satellite 35 bp



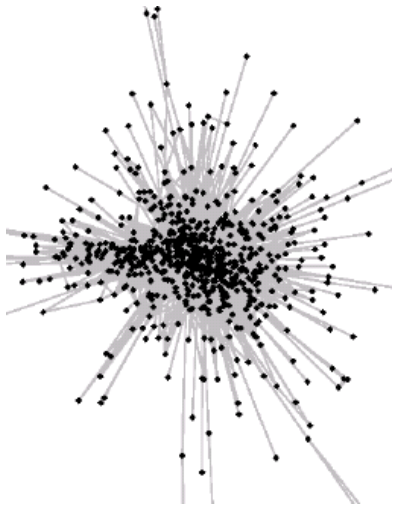


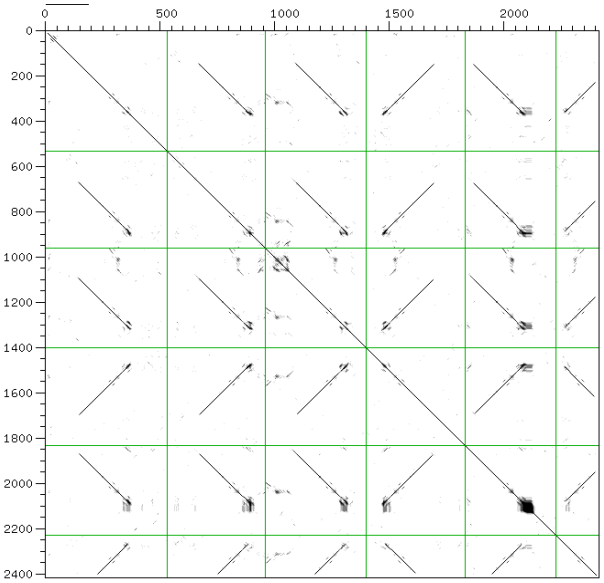
satellite 190 bp



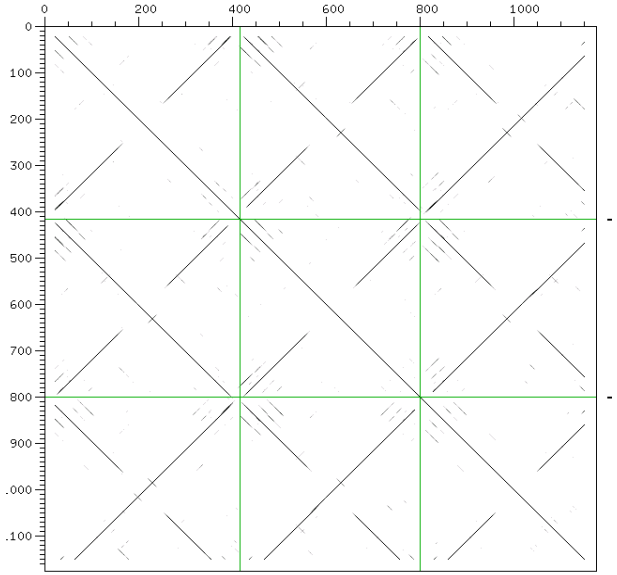
Circular graphs

SINE

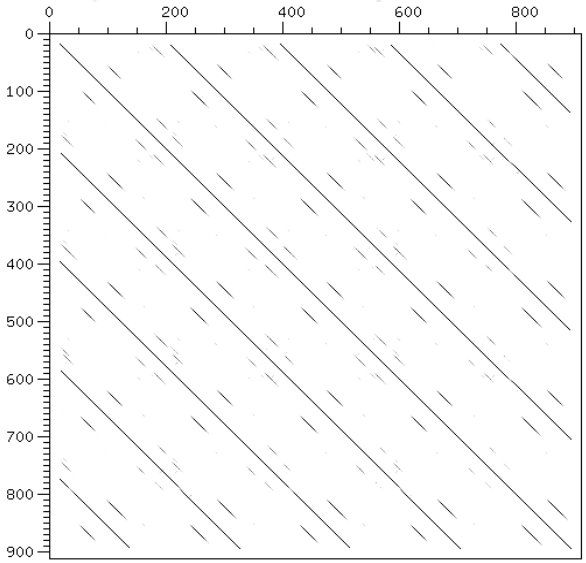




MITE
(foldback)

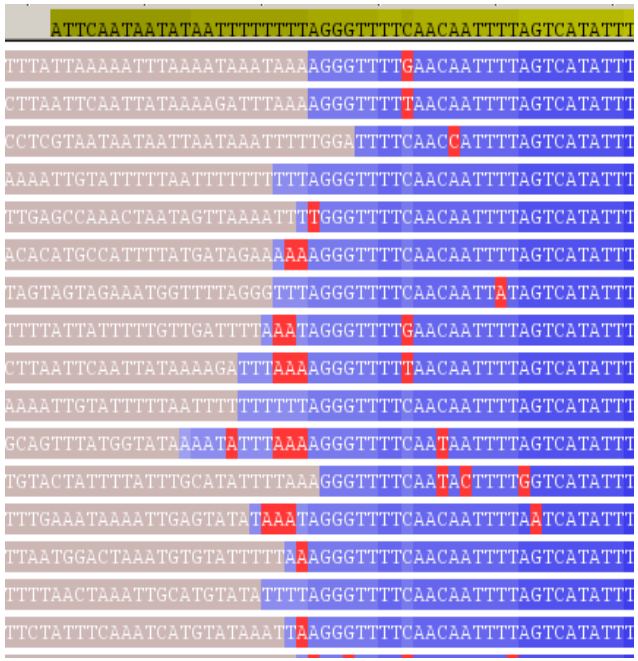


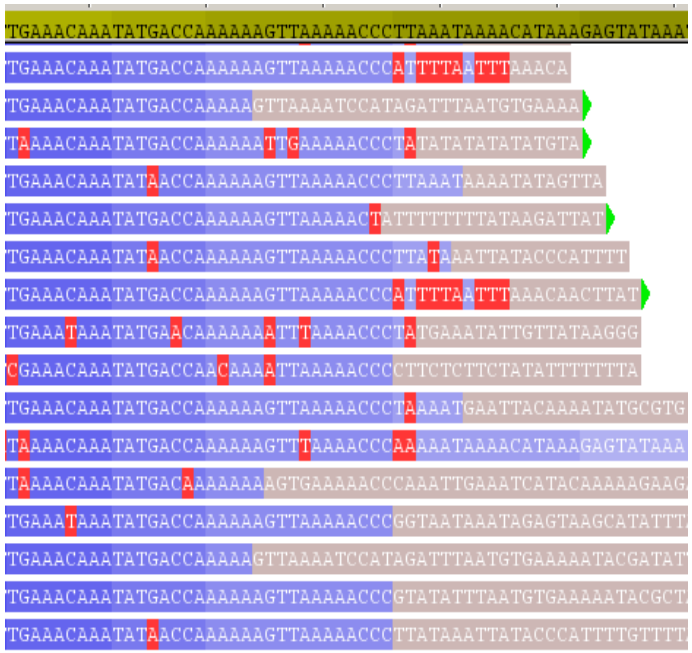
satellite 190 bp

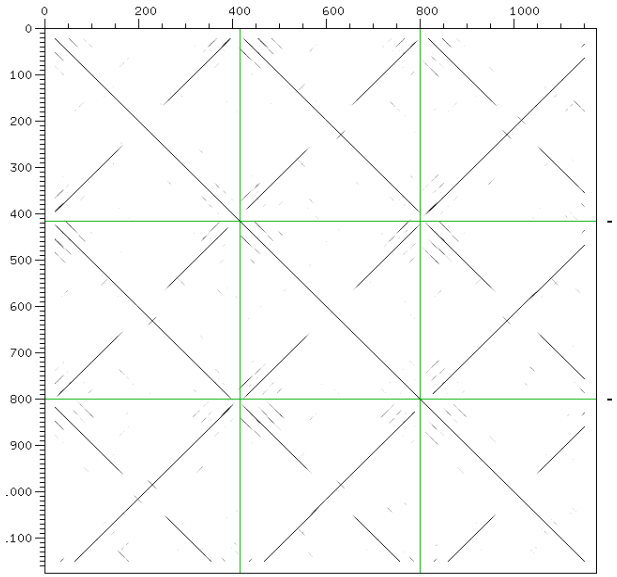


Insertion sites of mobile elements

MITE
(foldback)

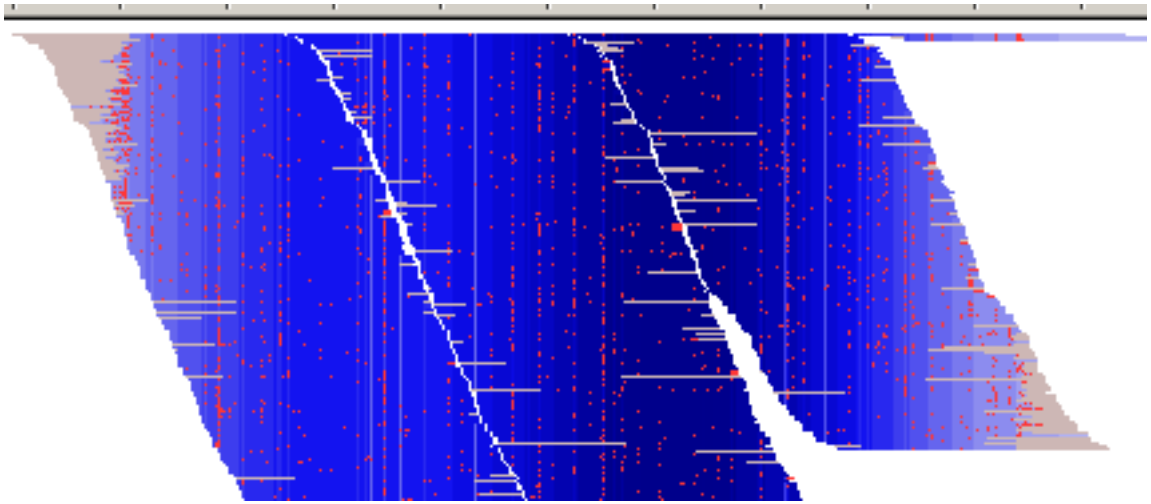






TIR

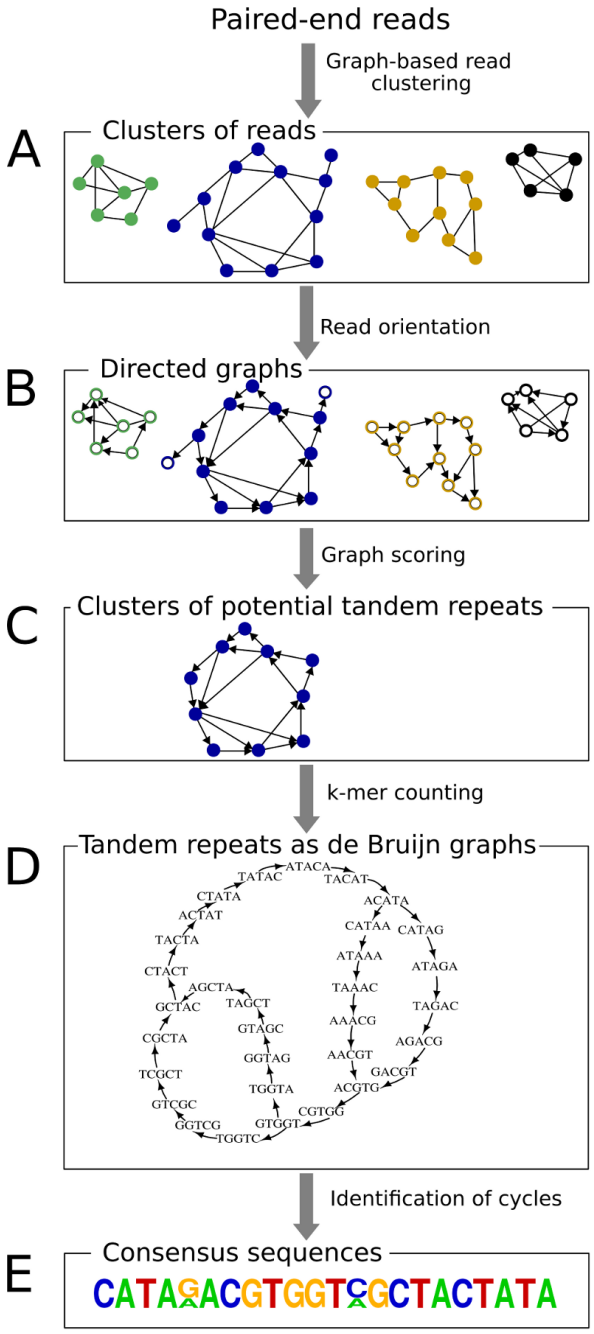
TIR



contig

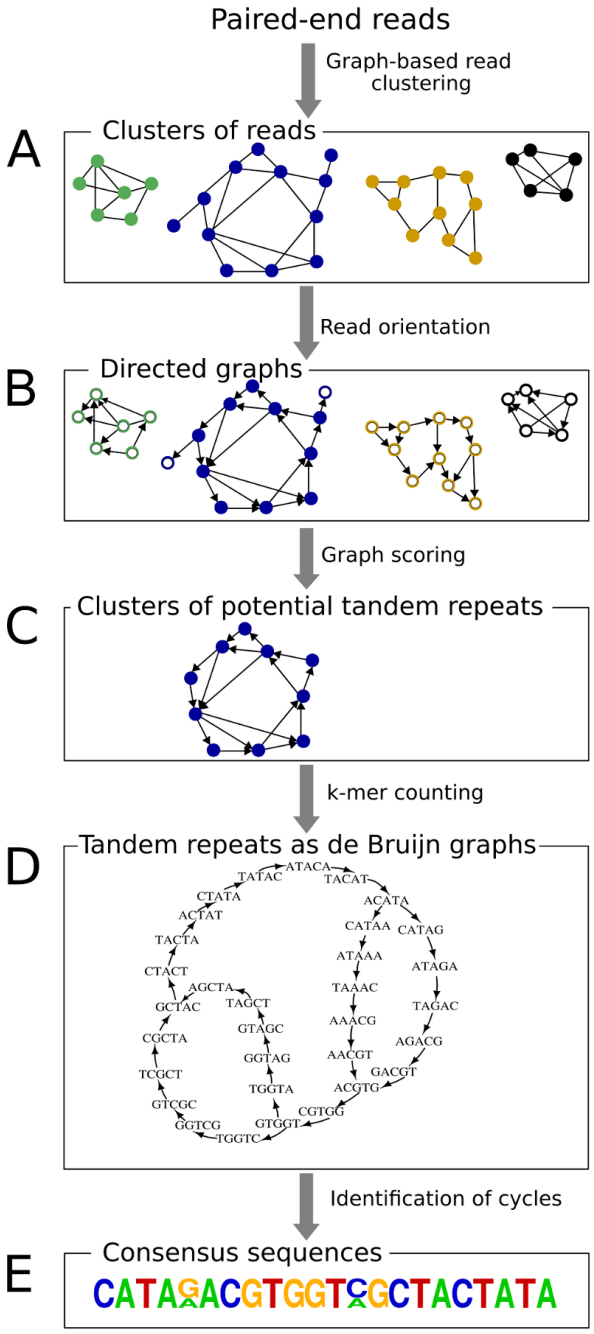
TAREAN

- Detects clusters with circular graphs automatically
- Calculates consensus sequences (alignment-free)
- Uses various parameters to distinguish tandem repeats from mobile elements
- *It is recommended to run TAREAN with cluster merging option as complementary analysis to RepeatExplorer*



TAREAN

- Detects clusters with circular graphs automatically
- Calculates consensus sequences (alignment-free)
- Uses various parameters to distinguish tandem repeats from mobile elements
- *It is recommended to run TAREAN with cluster merging option as complementary analysis to RepeatExplorer*

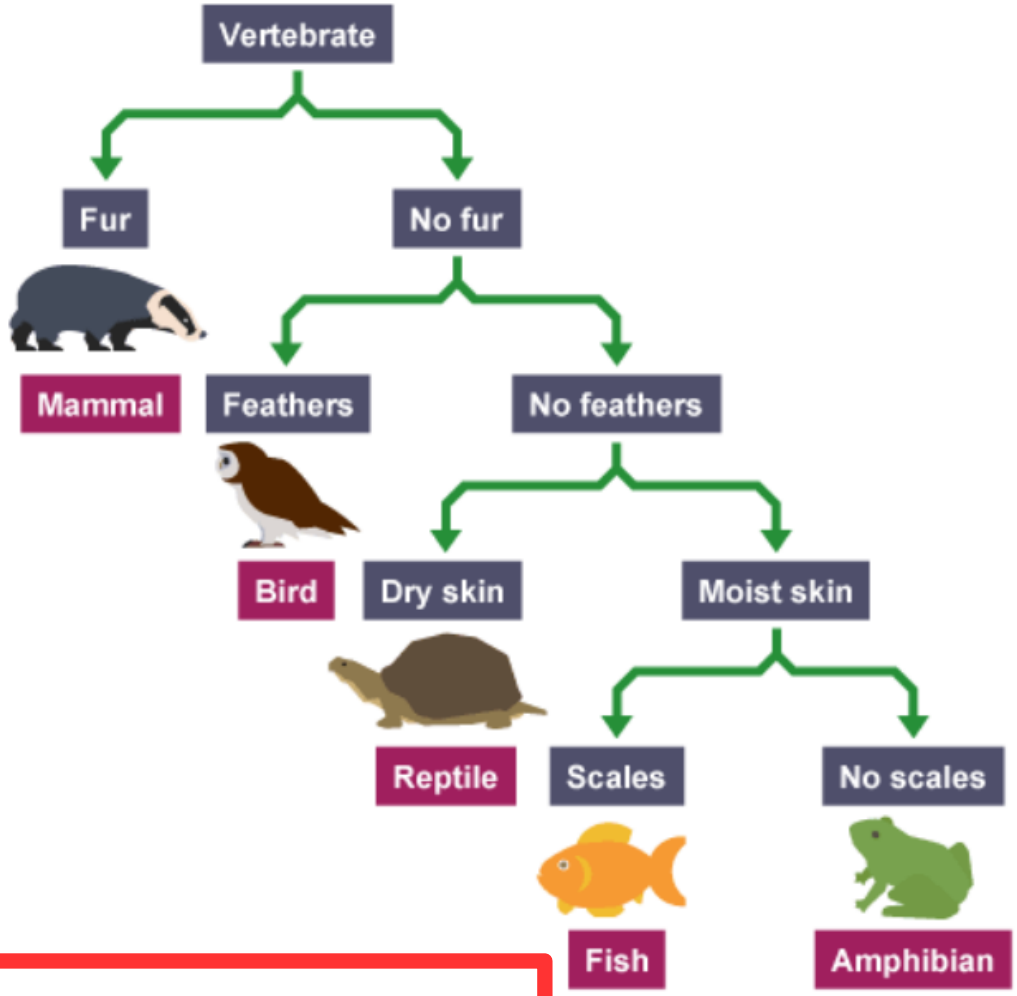


Repeat annotation and quantification

- always combine results of all tools
- check / correct automatic repeat classification
- consider experimental setup

Taxonomy of repetitive DNA

```
-----------------------------------------
Unclassified_repeat
 |--rDNA
 |    |--45S_rDNA
 |    |    |--18S_rDNA
 |    |    |--25S_rDNA
 |    |    '--5.8S_rDNA
 |    '--5S_rDNA
 |--satellite
 '--mobile_element
      |--Class_I
      |    |--SINE
      |    |--LTR
      |    |    |--Ty1_copia
      |    |    |    |--Ale
      |    |    |    |--Alesia
      |    |    |    |--Angela
      |    |    |    |--Bianca
      |    |    |    |--Bryco
      |    |    |    |--Gymco-I
      |    |    |    |--Gymco-II
      |    |    |    |--Ikeros
      |    |    |    |--Ivana
      |    |    |    |--Osser
      |    |    |    |--SIRE
      |    |    |    |--TAR
      |    |    |    |--Tork
      |    |    |    '--Ty1-outgroup
      |    |    '--Ty3_gypsy
      |    |         |--non-chromovirus
      |    |         |    |--nonchromo-outgroup
      |    |         |    |--Phygy
      |    |         |    |--Selgy
      |    |         |    '--OTA
      |    |         |        |--Athila
      |    |         |        '--Ogre_Tat
      |    |         |            |--TatI
      |    |         |            |--TatII
      |    |         |            |--TatIII
      |    |         |            |--TatIV_Ogre
      |    |         |            '--TatV
      |    |         '--chromovirus
      |    |              |--Chlamyvir
      |    |              |--Tcn1
      |    |              |--CRM
      |    |              |--Galadriel
      |    |              |--Tekay
      |    |              |--Reina
      |    |              |--chromo-outgroup
      |    |              '--chromo-unclass
      |    |--pararetrovirus
      |    |--DIRS
      |    |--Penelope
      |    '--LINE
      '--Class_II
           |--Subclass_1
           |    '--TIR
           |         |--MITE
           |         |--EnSpm_CACTA
           |         |--hAT
           |         |--Kolobok
           |         |--Merlin
```



Keep improving your knowledge of "repeat taxonomy"