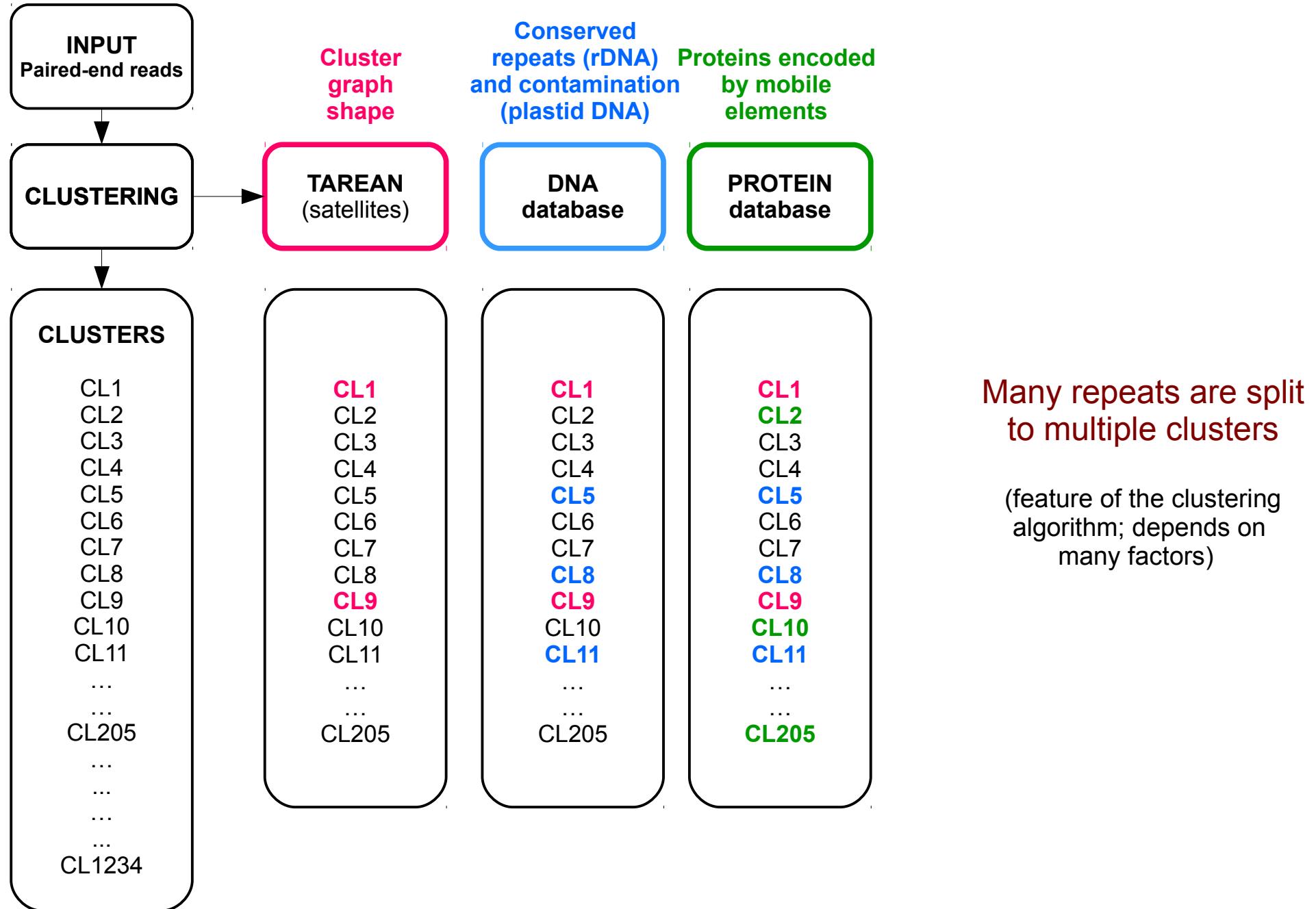


Using *RepeatExplorer* output for repeat  
annotation and quantification

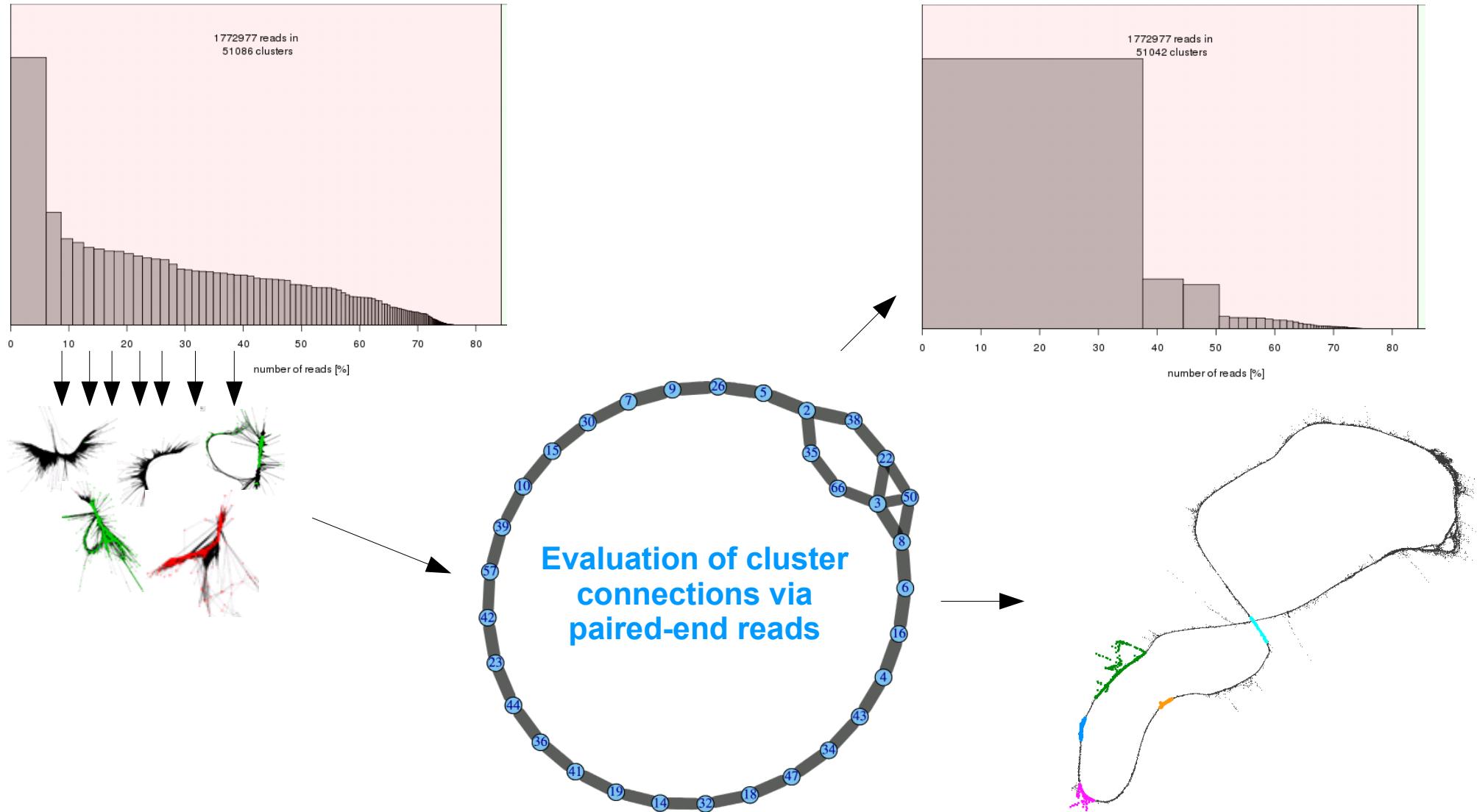
# RepeatExplorer ver. 2



# RepeatExplorer ver. 2

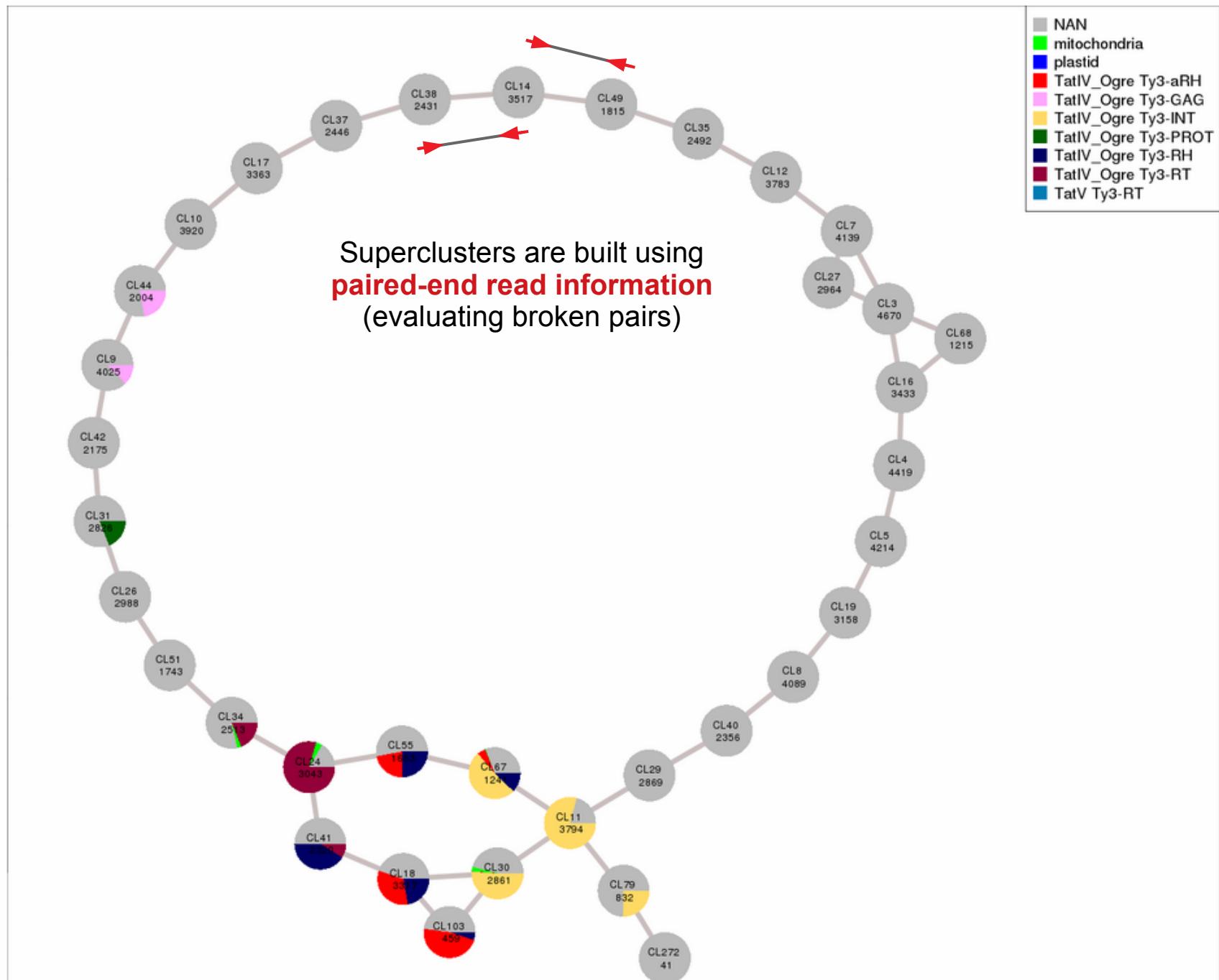


# Cluster fragmentation

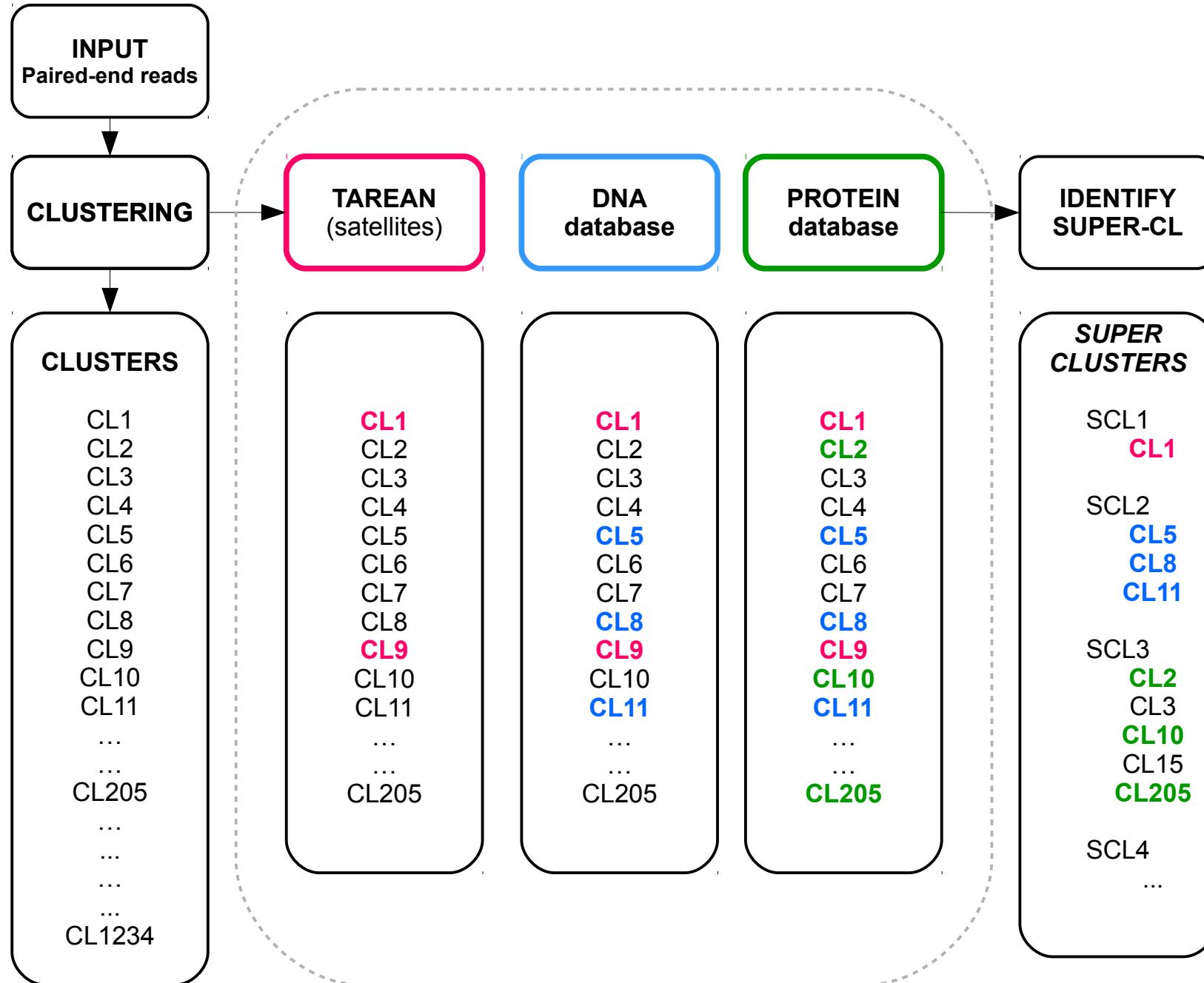


LTR-retrotransposon Ogre in *Vicia pannonica* (~ 40% of the genome)

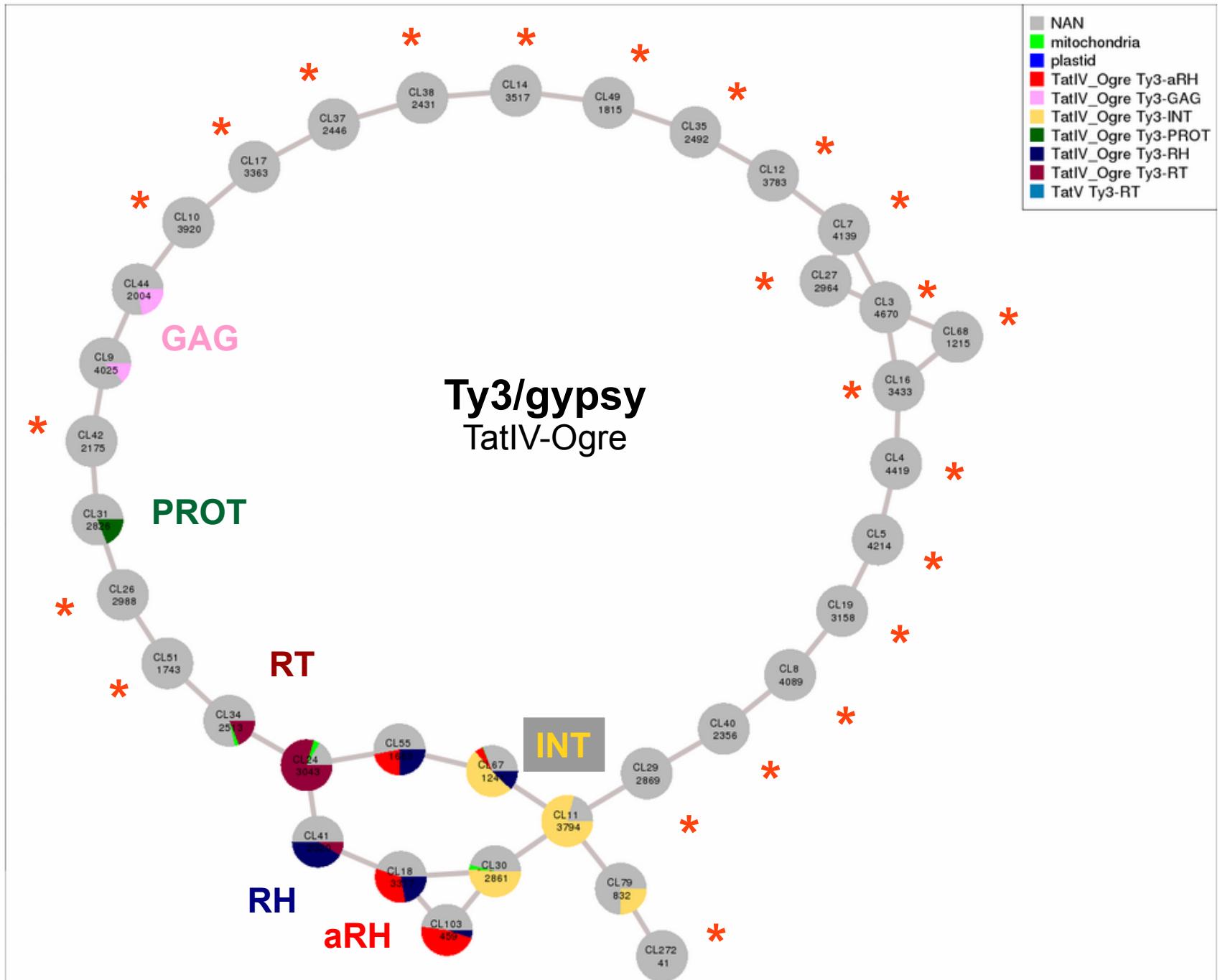
# RepeatExplorer 2 – automatic detection of superclusters



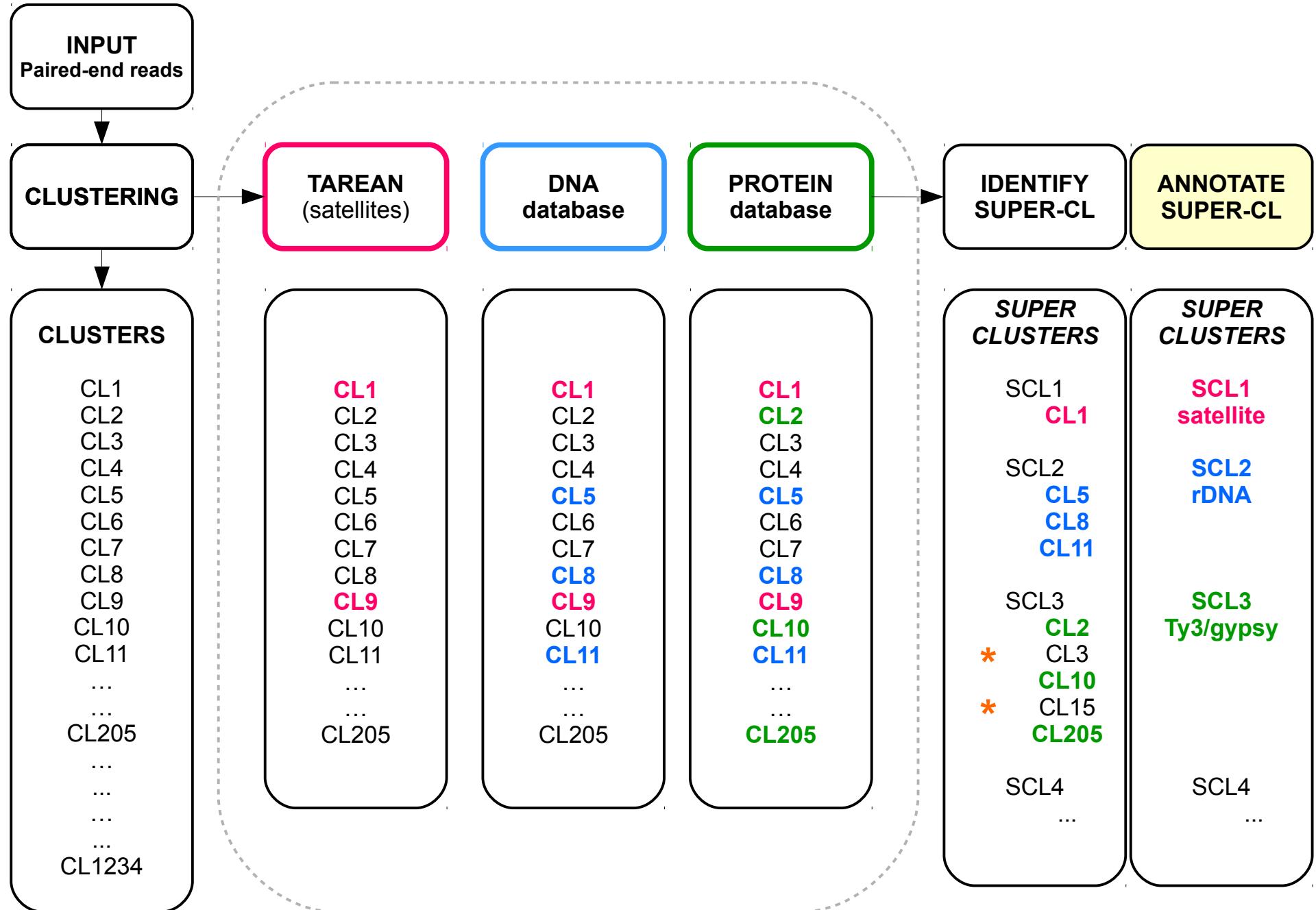
# RepeatExplorer 2 – automatic detection of superclusters



# Superclusters provide more complete annotation



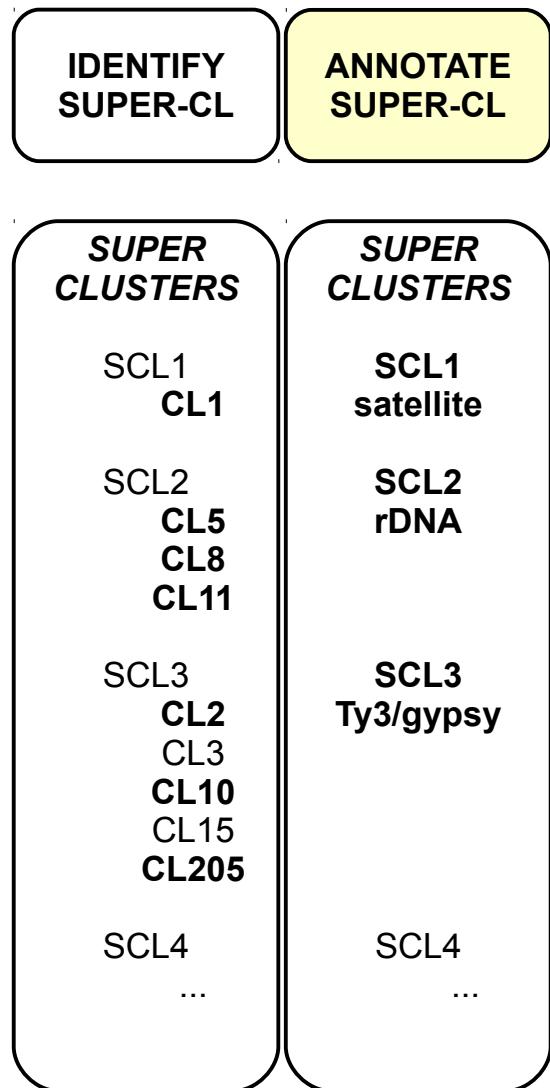
# Superclusters provide more complete annotation



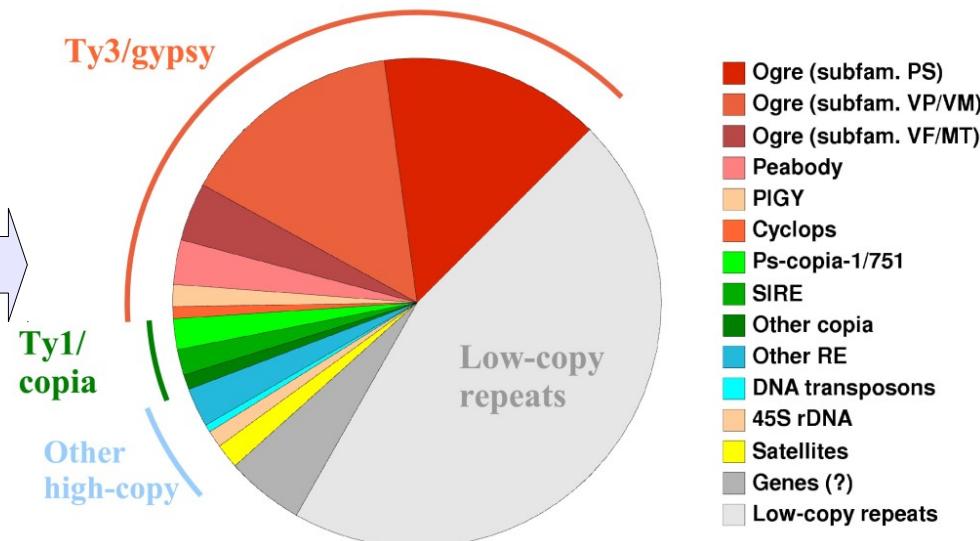
# Repeat quantification

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

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

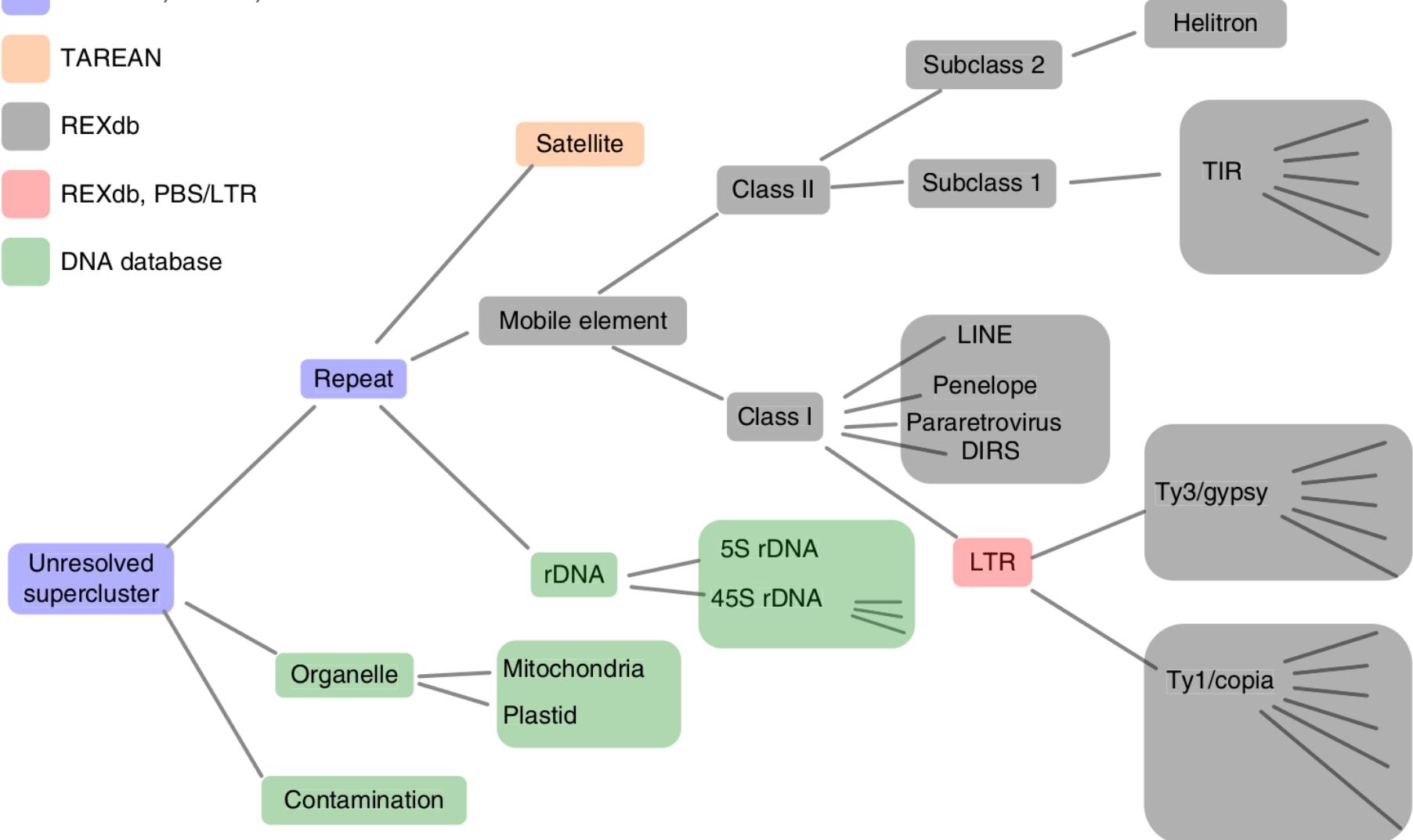


Proportions of various repeat types in a genome



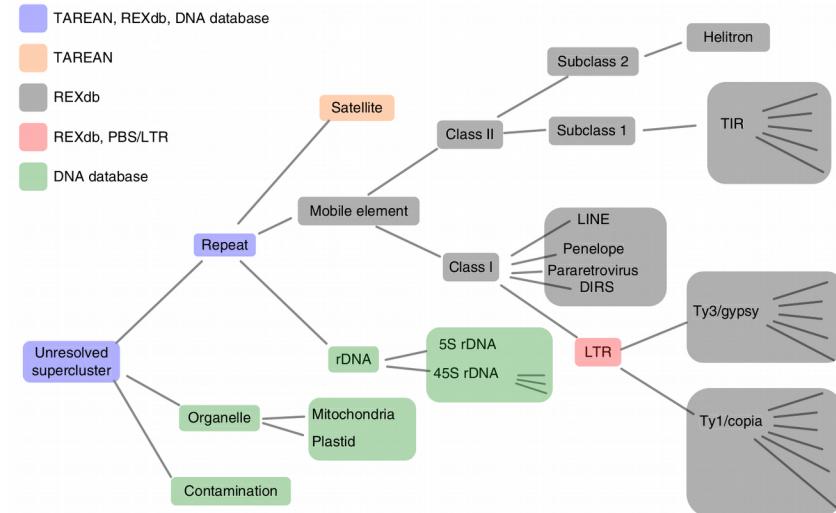
# Decision tree for automatic annotation

- TAREAN, REXdb, DNA database
- TAREAN
- REXdb
- REXdb, PBS/LTR
- DNA database



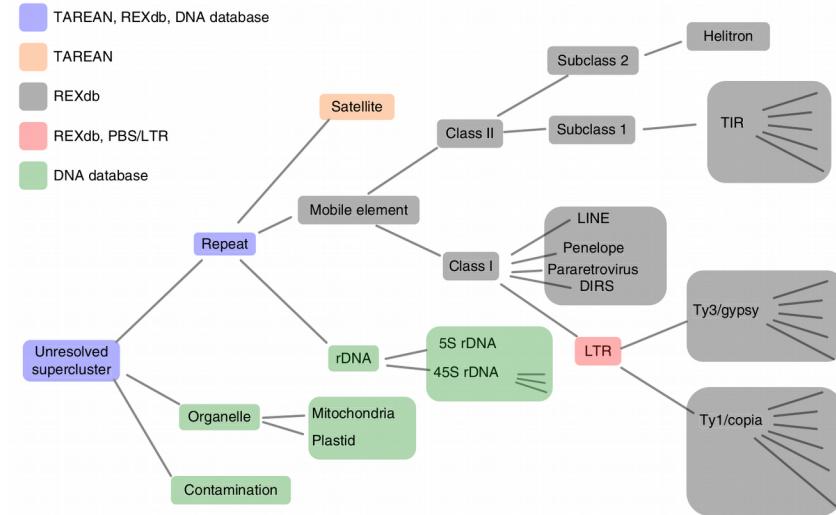
# How reliable is automatic annotation ?

- conserved repeats (rDNA)
- organelles, contamination
- satellite DNA (structure-based annotation)
- mobile elements (autonomous)
  - plants
  - other organisms
- non-autonomous mobile elements
  - ✓ LTR detection
    - insertion site detection (coming soon)



# How reliable is automatic annotation ?

- conserved repeats (rDNA)
- organelles, contamination
- satellite DNA (structure-based annotation)
- mobile elements (autonomous)
  - plants
  - other organisms
- non-autonomous mobile elements
  - ✓ LTR detection
    - insertion site detection (coming soon)

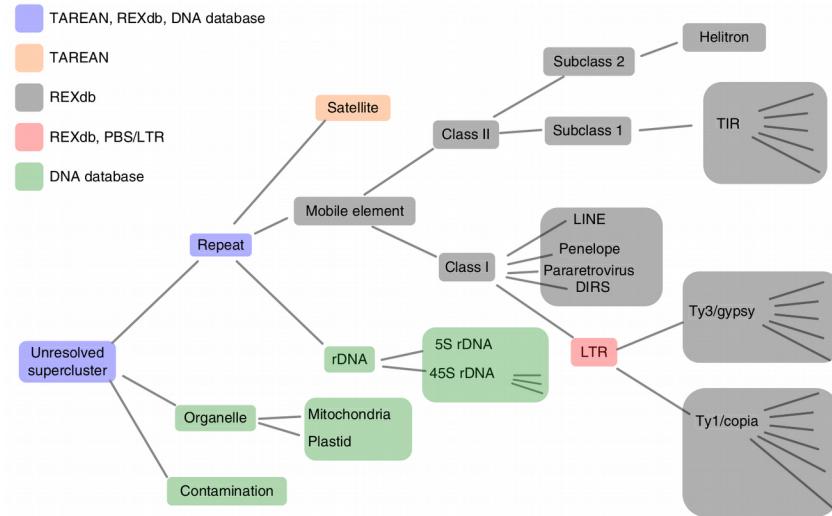


## Working with organisms not covered in REXdb ?

- use custom repeat database (and share it !)
- help us to improve REXdb ;-)

# How reliable is automatic annotation ?

- conserved repeats (rDNA)
- organelles, contamination
- satellite DNA (structure-based annotation)



- mobile elements (autonomous)
  - plants
  - other organisms
- non-autonomous mobile elements
  - LTR detection
    - insertion site detection (coming soon)

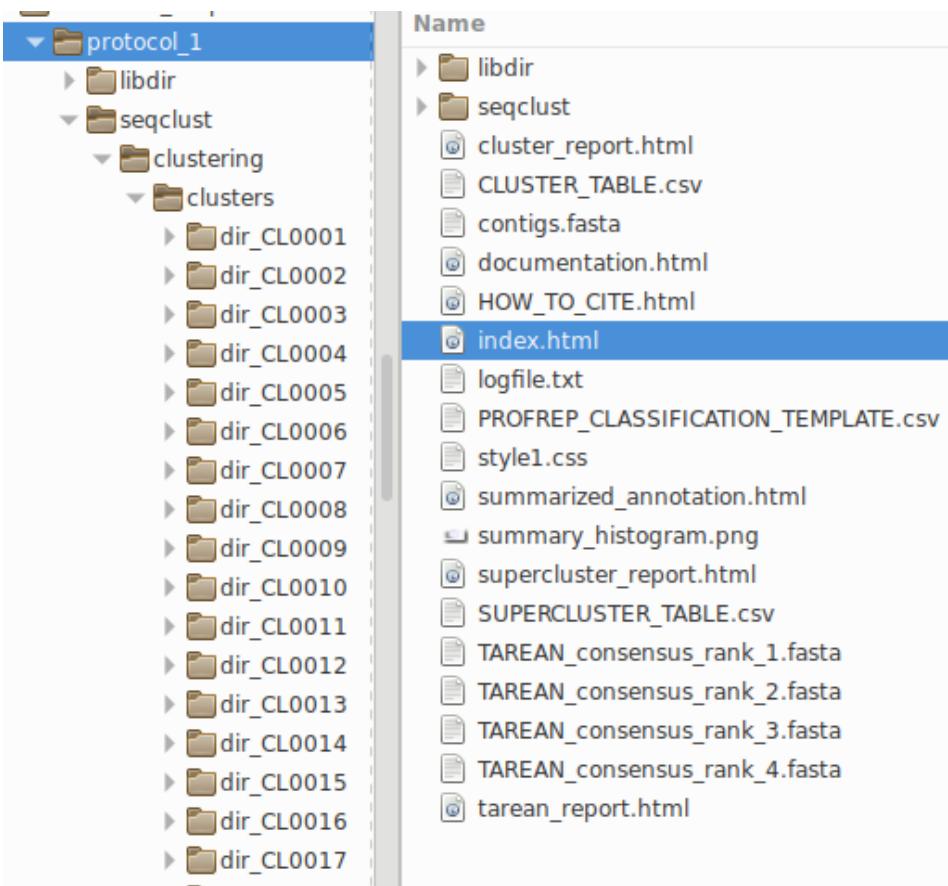
## Working with organisms not covered in REXdb ?

- use custom repeat database (and share it !)
- help us to improve REXdb ;-)

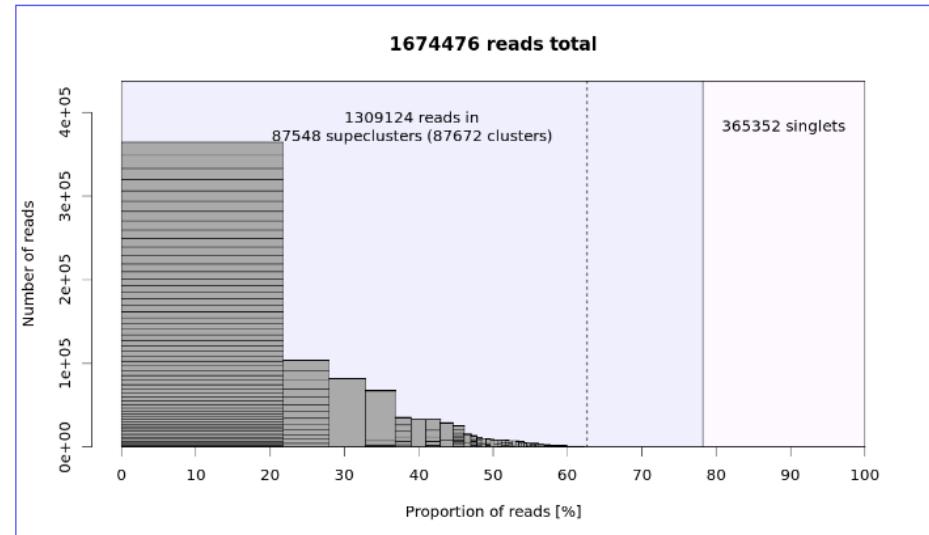
## Clusters without annotation

- The lack of annotation does not mean these are not repeats (all repeats are detected)
- More info available in cluster directories

# RepeatExplorer output files



## Clustering Summary



**Graphical summary of the clustering results.** Bars represent superclusters, with their heights and widths corresponding to the numbers of reads in the superclusters (y-axis) and to their proportions in all analyzed reads (x-axis), respectively. Rectangles inside the supercluster bars represent individual clusters. If the filtering of abundant satellites was performed, the affected clusters are shown in green, and their sizes correspond to the adjusted values. Blue and pink background panels show proportions of reads that were clustered and remained single, respectively. Top clusters are on the left of the dotted line.

## Run information:

Number of input reads: 2935772

Number of analyzed reads: 1674476

Proportion of reads in top clusters : 63 %

Cluster merging: No

Paired-end reads: Yes

## Available analyses:

[Tandem repeat analysis](#)

[Cluster annotation](#)

[Supercluster annotation](#)

[Repeat annotation summary](#)

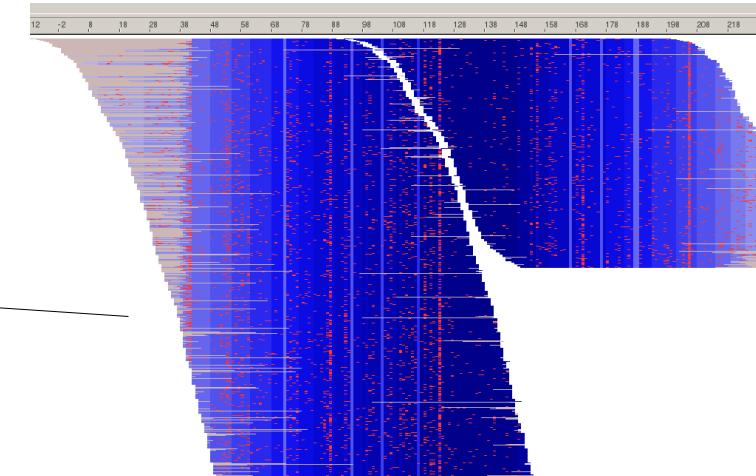
Always download and use archive with complete output of the analysis

- do not work with HTML only

# RepeatExplorer output files

## Cluster directories

Name
html_files
tarean
assembly.info
contigs.ace
contigs.aln
contigs.fasta
contigs.info.fasta
contigs.info.minRD5_sort-GR.fasta
contigs.info.minRD5_sort-length.fasta
contigs.info.minRD5_sort-RD.fasta
contigs.links
contigs.profile
contigs.qual
custom_db_extra_database_annotation.csv
dna_database_annotation.csv
graph_layout.GL
graph_layout.png
graph_layout_directed.RData
graph_layout_tmb.png
hitsort_part.csv
index.html
LTR_info.ADJ
LTR_info.LTR
LTR_info.with_PBS_blast.csv
protein_database_annotation.csv
reads.fasta
reads_selection.fasta
reads_selection_oriented.fasta
singlets.fasta



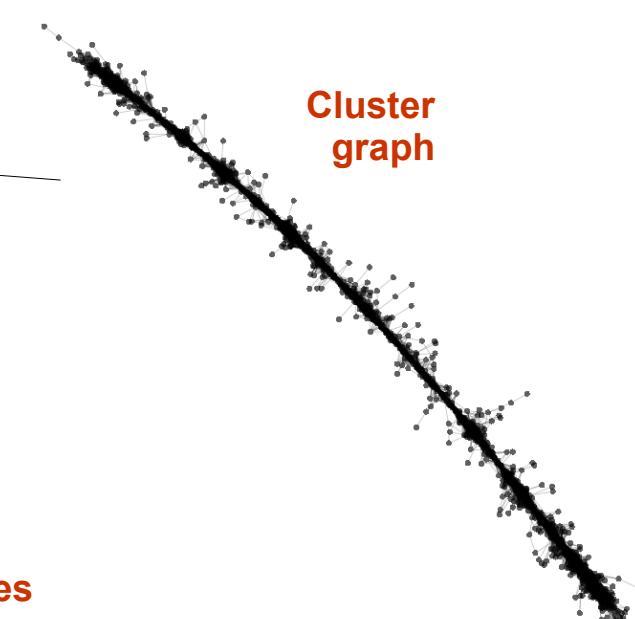
Contigs

assembly

Cluster graph

LTR-detection

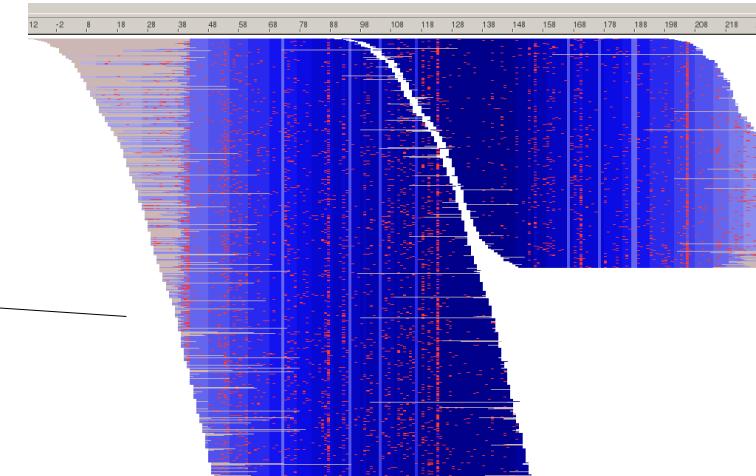
Read sequences



# RepeatExplorer output files

## Cluster directories

Name
html_files
tarean
assembly.info
contigs.ace
contigs.aln
contigs.fasta
contigs.info.fasta
contigs.info.minRD5_sort-GR.fasta
contigs.info.minRD5_sort-length.fasta
contigs.info.minRD5_sort-RD.fasta
contigs.links
contigs.profile
contigs.qual
custom_db_extra_database_annotation.csv
dna_database_annotation.csv
graph_layout.GL
graph_layout.png
graph_layout_directed.RData
graph_layout_tmb.png
hitsort_part.csv
index.html
LTR_info.ADJ
LTR_info.LTR
LTR_info.with_PBS_blast.csv
protein_database_annotation.csv
reads.fasta
reads_selection.fasta
reads_selection_oriented.fasta
singlets.fasta



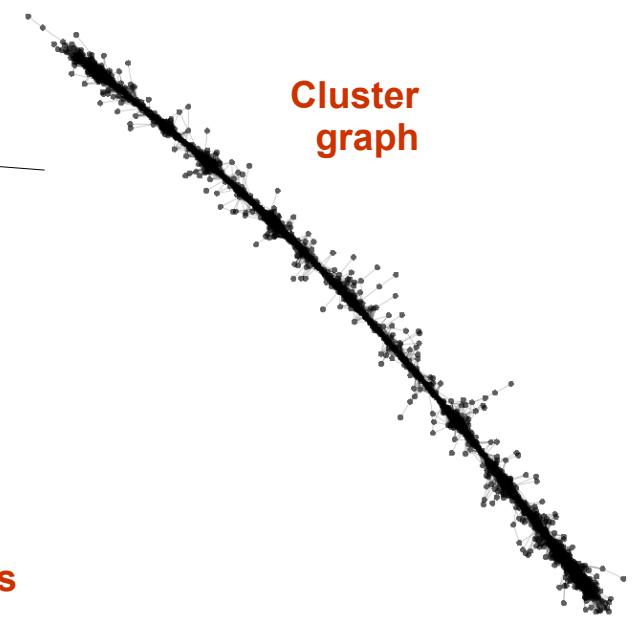
Contigs

assembly

Cluster graph

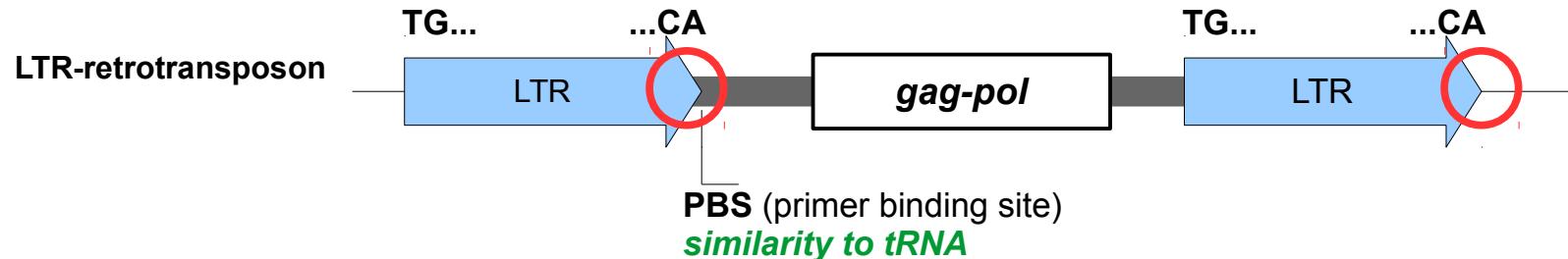
LTR-detection

Read sequences



# Insertion sites of mobile elements

## A tool for detection of LTR / PBS sites



### Program output:

CL	contig	pos.	site	site_depth	out_ma	masked	masked	region_in	region_out	blast to tRNA	%	length	site from	to	tRNA from	to	E-val	
19	400	364	TGCGACA	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)

↓

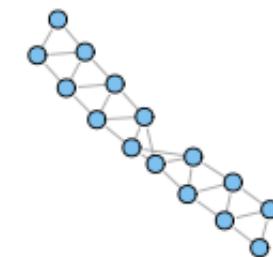
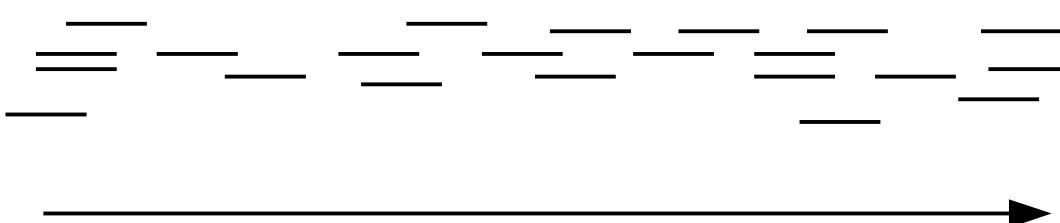
tRNA-Arg

TAAT\*TTTTTCCGCACCATCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAT\*TTTTTCCGCACCATCGCGA\*GGGATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAA\*TTTTTCCGCAACCATCATGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 CCAT\*TTTTTCCGAGACCATCGCGA\*GGGATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAT\*TTTTTCCGCACCATCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAT\*TTTTTCCGCACCATCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 AAATA\*TTTTCCGCACCATCGCGA\*GGGATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAT\*TTTTTCCGCGACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAT\*TTTTTCCGCGACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAT\*TTTTTCCGCGACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 ACAT\*TTTTTCCGCACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 ATTT\*TTTTTCCACGACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 ACAT\*TTTTTCCCGACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAT\*TTTTTCCGCACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAA\*TTTTTCCGCACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG  
 TAAA\*TTTTTCCGCCACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGGGAAGCCTAATTAG  
 TAAATA\*TTTTTCCGCACCACCGCGA\*GGAATCGTATT\*CGAGATGCGACAGATGGCGACTCTGCTGGGAC\*\*TA\*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTAG



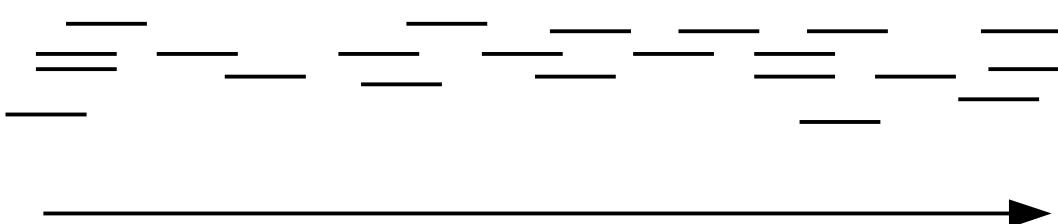
## Linear graphs

---

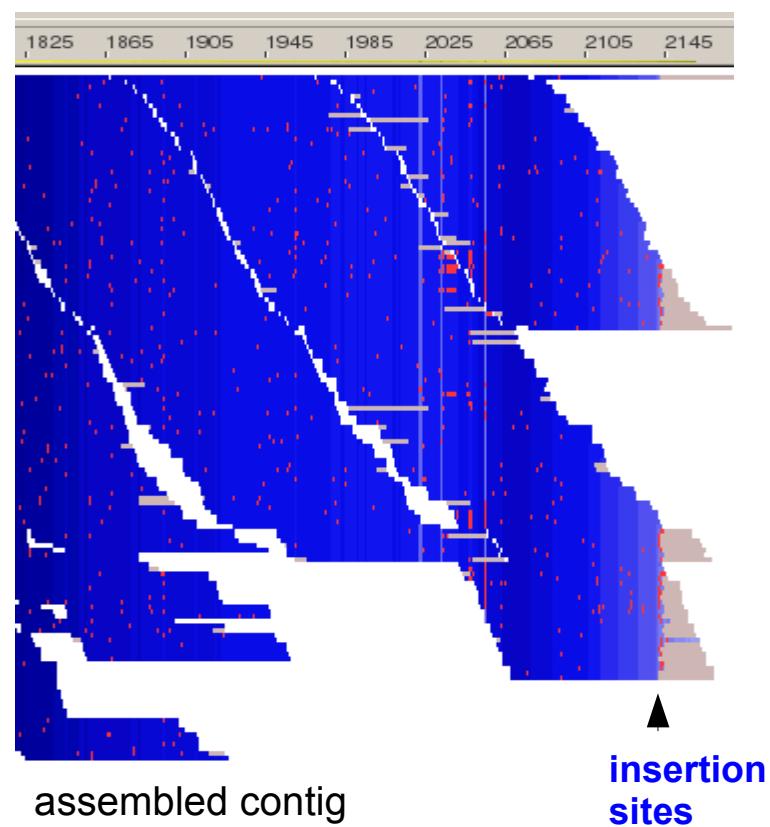
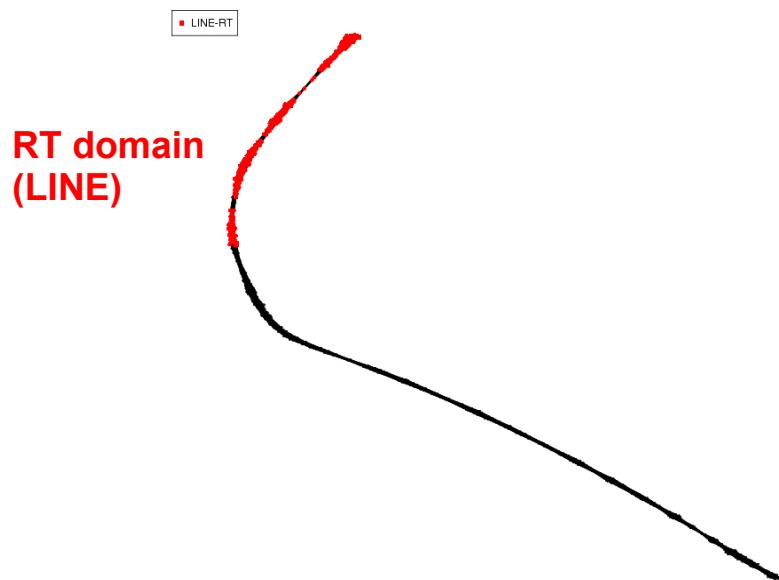
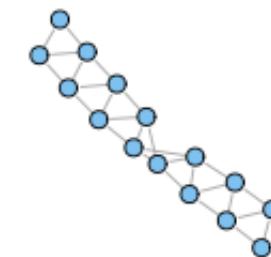


DNA transposons, LINEs, ...

# Linear graphs



DNA transposons, LINEs, ...

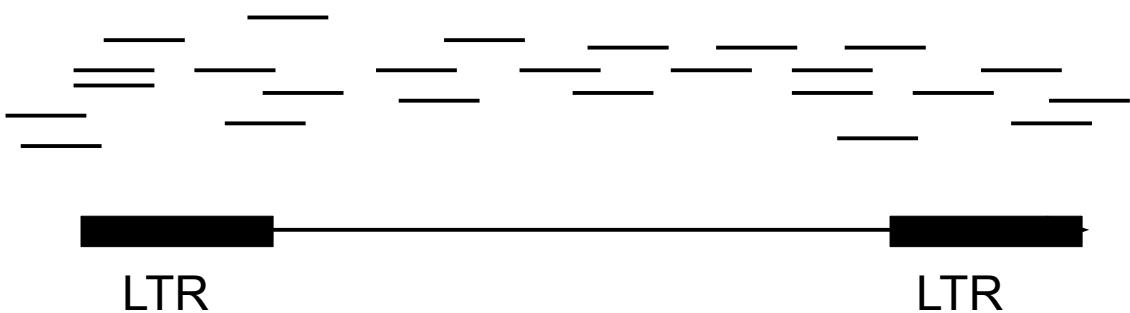


assembled contig

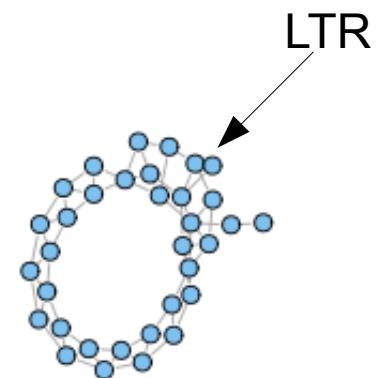
insertion  
sites

## Linear / large circular graphs

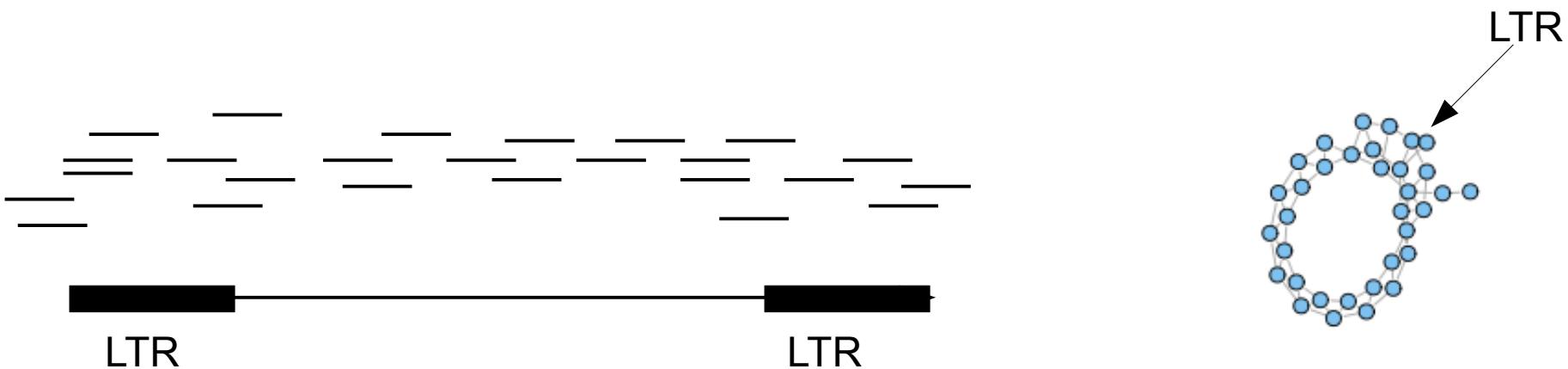
---



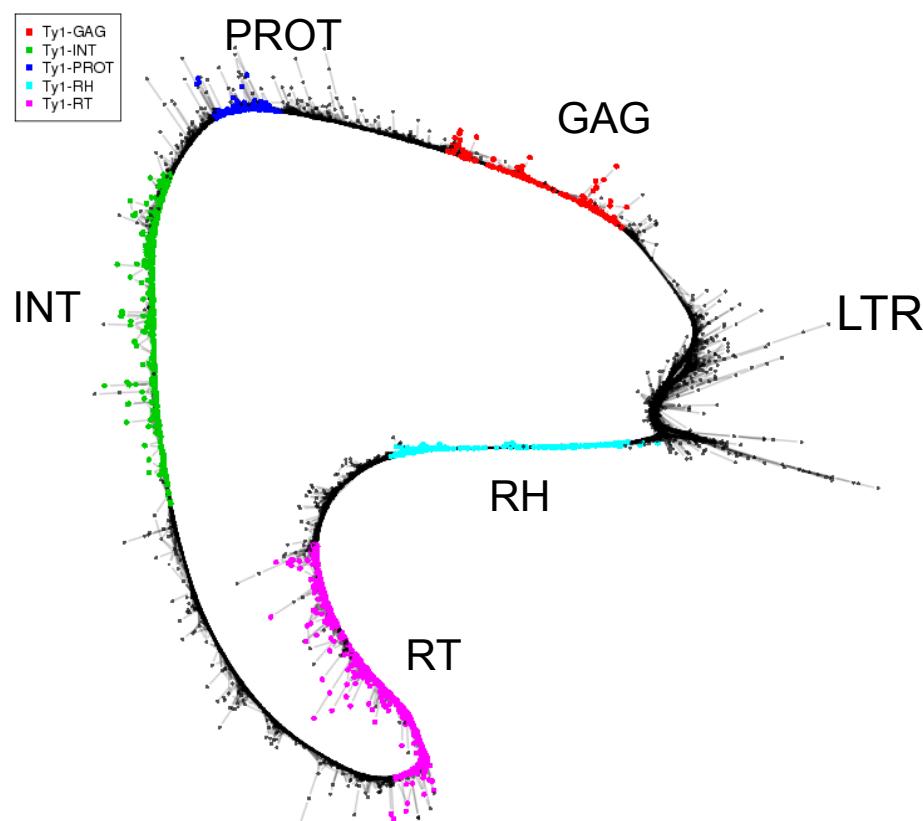
LTR-retrotransposons



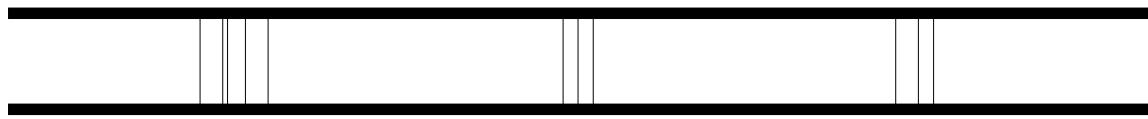
# Linear / large circular graphs



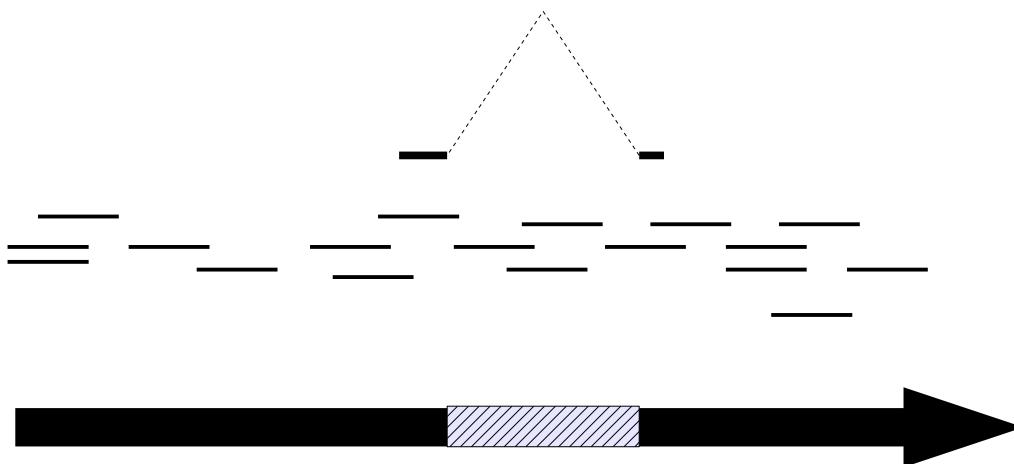
## LTR-retrotransposons



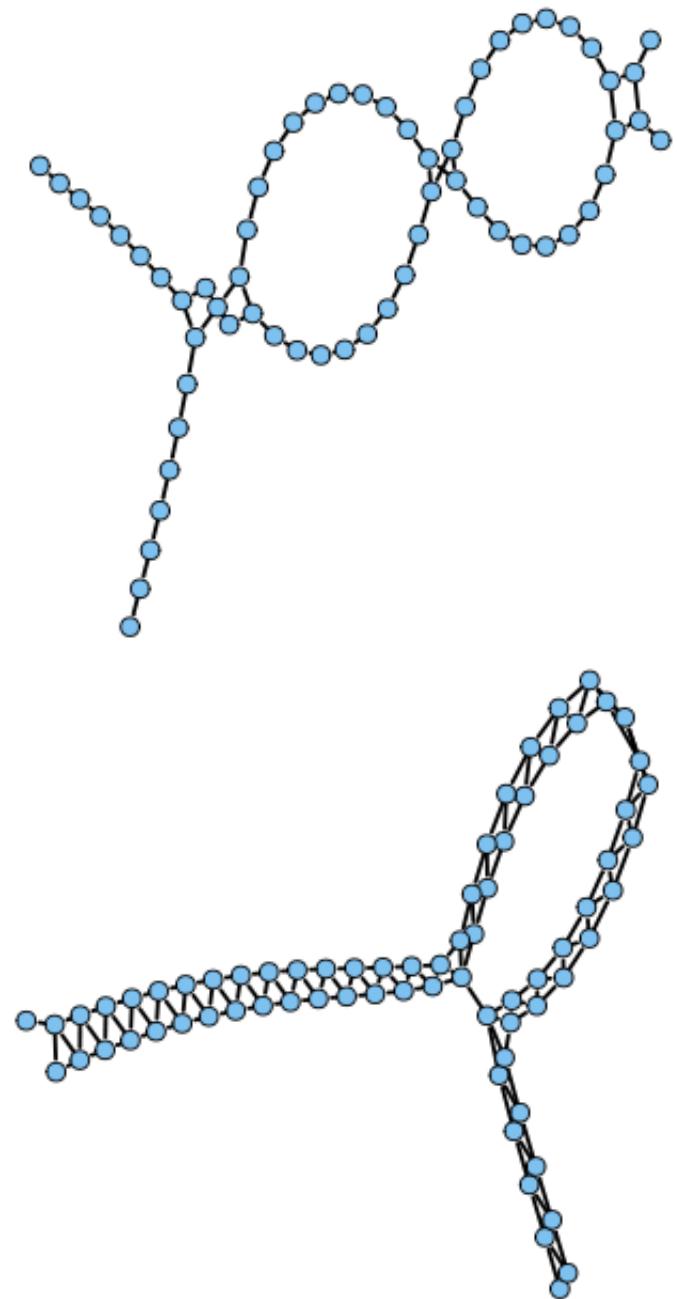
# It's getting complicated...



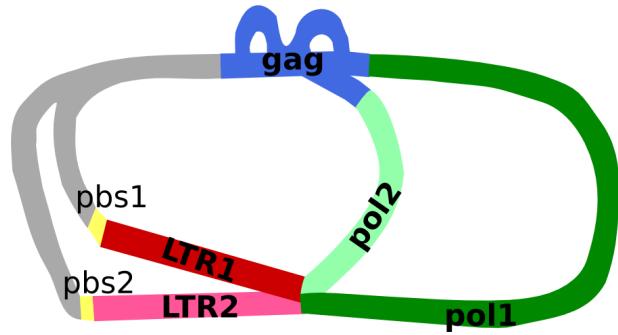
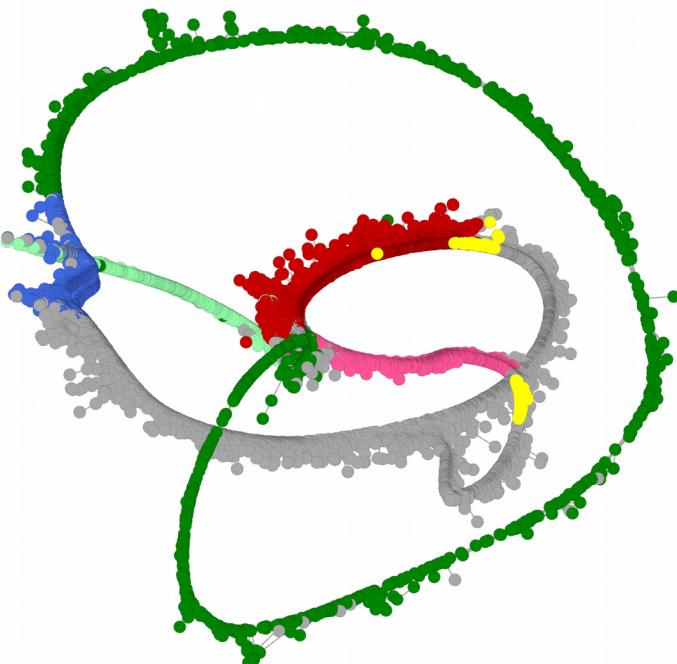
Repeat variants with partial similarity  
(depends on similarity threshold)



Variants differing in indel

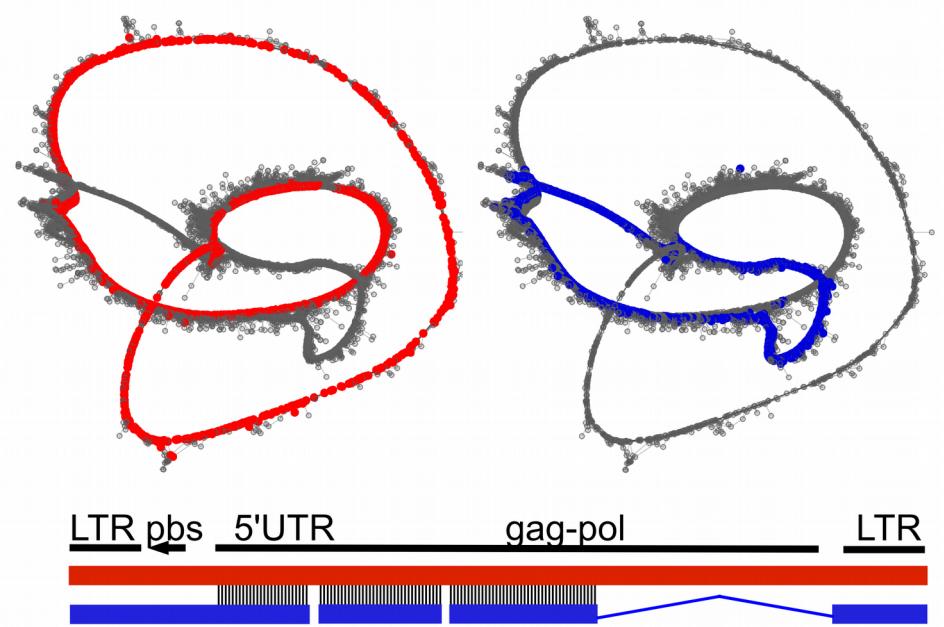


# It's getting complicated...



...but it has some meaning

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

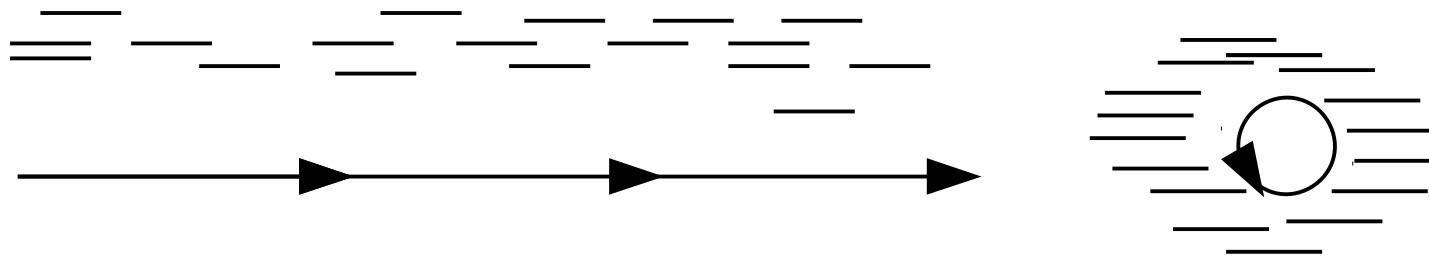


# Circular graphs

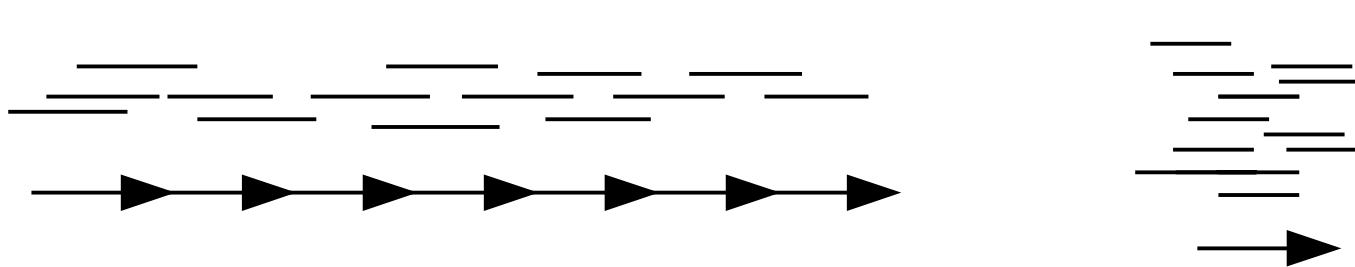
---

## TANDEM REPEATS

Read length << monomer

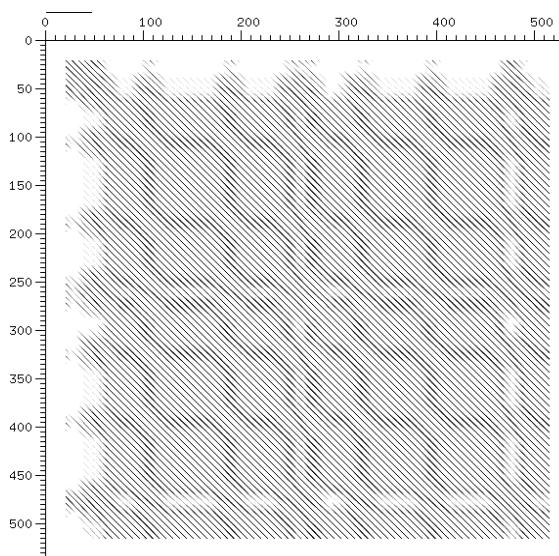
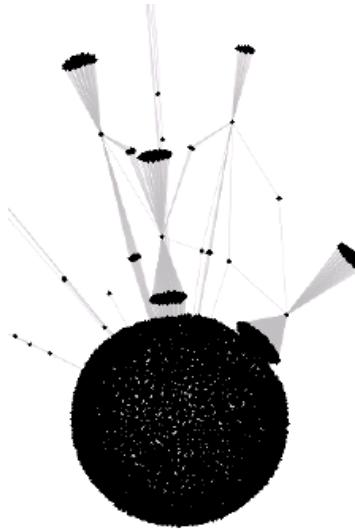


Read length  $\geq$  monomer

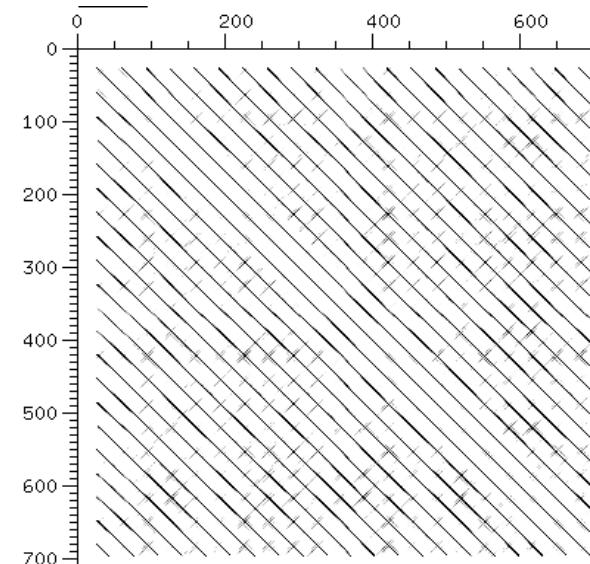
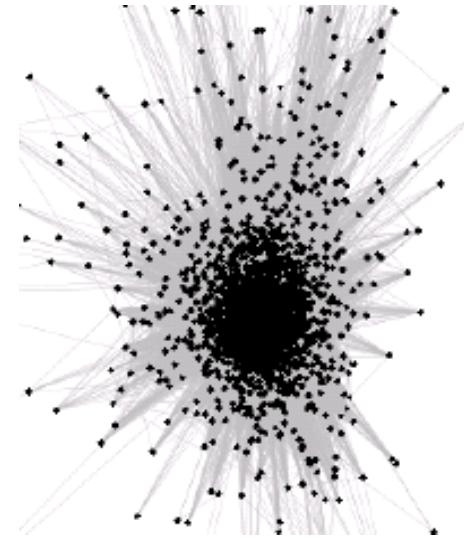


# Circular graphs

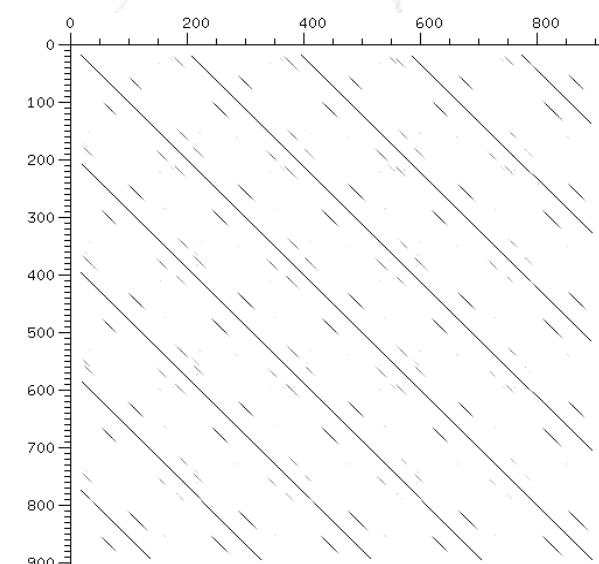
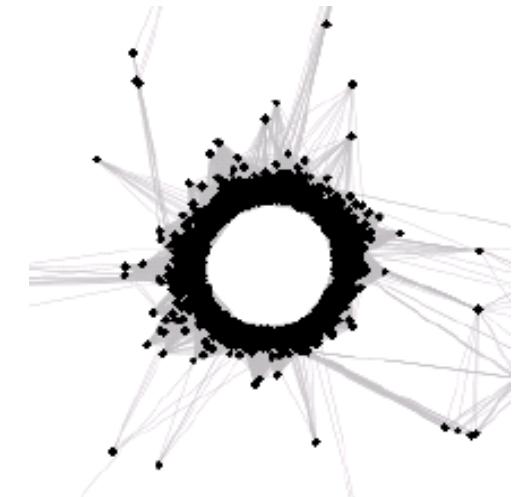
**microsatellite (6bp)  
(GAACCT)n**



**satellite 35 bp**

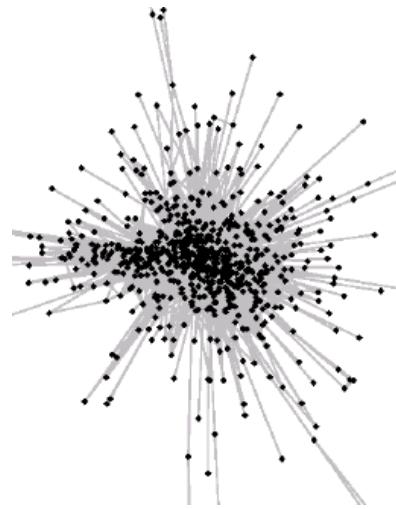


**satellite 190 bp**



# Circular graphs

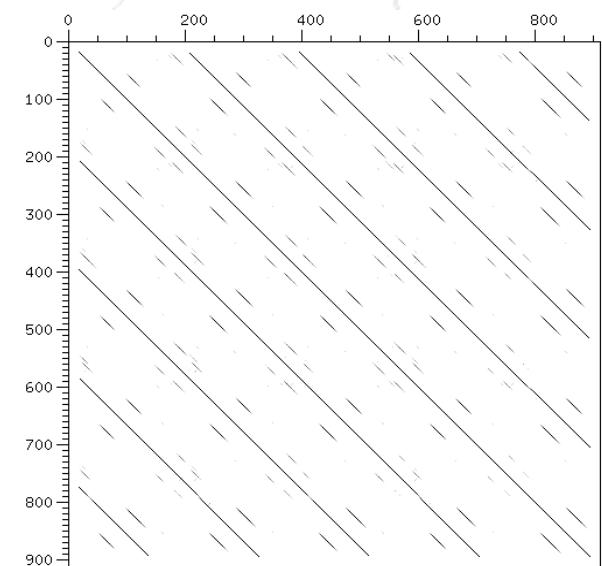
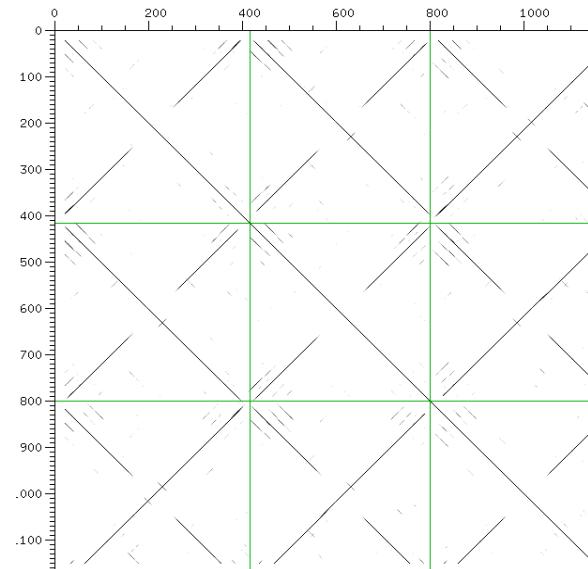
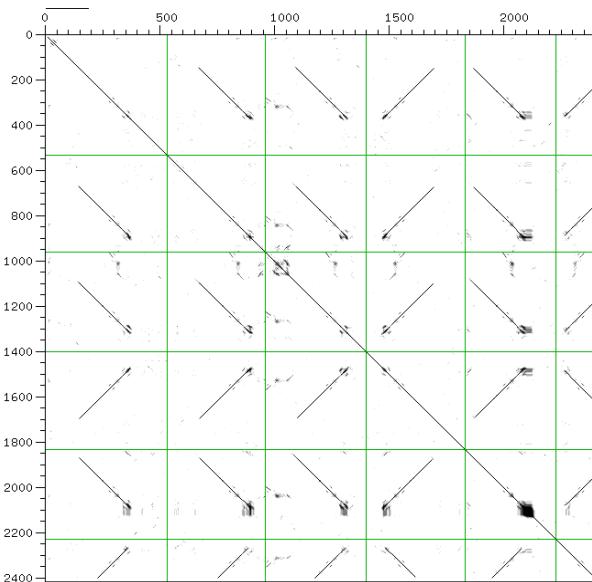
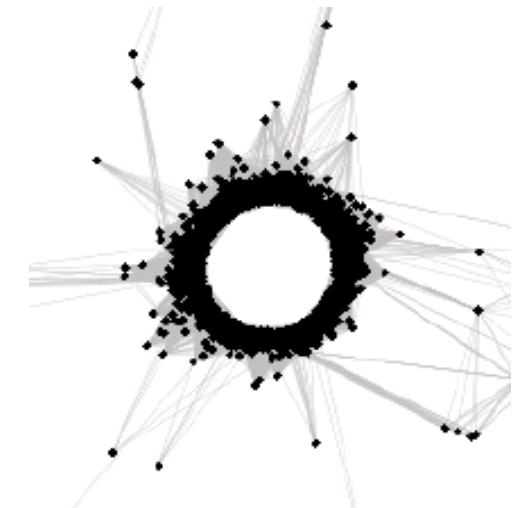
SINE



MITE  
(foldback)

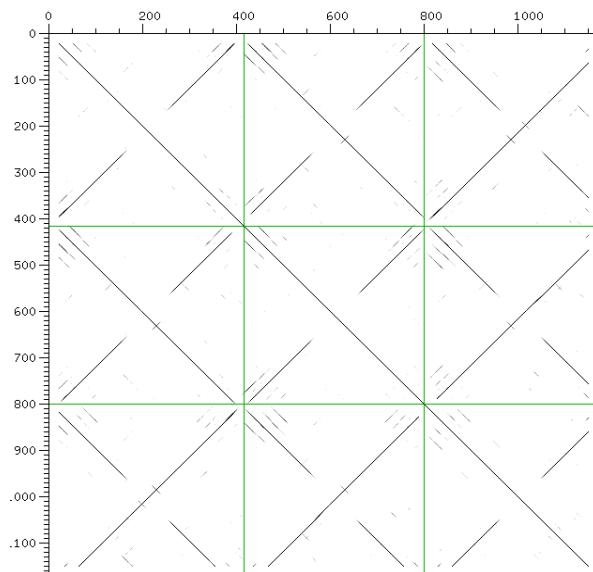


satellite 190 bp



# Insertion sites of mobile elements

# MITE (foldback)

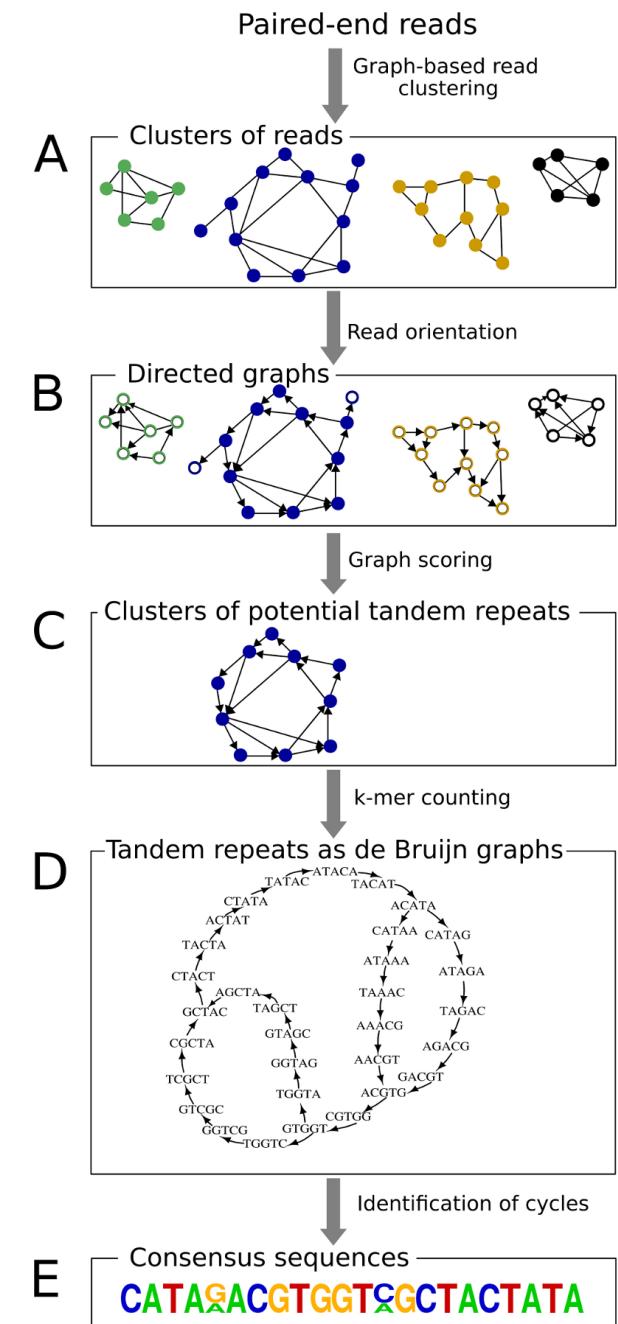


ATTCAATAATATAATTTTTTAGGGTTTCAACAAATTAGTCATATT  
TTTATTAAAAATTAAAATAAAGGGTTTGAACTATTAGTCATATT  
CTTAATTCAATTATAAAAGATTAAAGGGTTTAAACAAATTAGTCATATT  
CCTCGTAATAATAATTAAATTGGATTTCACACCATTAGTCATATT  
AAAAATTGTATTTTATTAGGTTTCAACAAATTAGTCATATT  
TTGAGCCAAACTAATAGTAAATTGGGTTTCACAAATTAGTCATATT  
ACACATGCCATTATGATAGAAAGGGTTTCAACAAATTAGTCATATT  
TAGTAGTAGAAATGGTTAGGGTTAGGGTTCAACAAATTAGTCATATT  
TTTATTATTGGTGTATTAAATAGGGTTTGAACTATTAGTCATATT  
CTTAATTCAATTATAAAAGATTAAAGGGTTTAAACAAATTAGTCATATT  
AAAAATTGTATTTTATTAGGTTTCAACAAATTAGTCATATT  
GCAGTTATGGTATAAAATTAAAGGGTTTCATAATTAGTCATATT  
TGTACTATTATTGCATATTAAAGGGTTTCATAATTGGTCATATT  
TTGAAATAAAATTGAGTATAAAATAGGGTTTCAACAAATTAACTCATATT  
TTAATGGACTAAATGTGTATTAAAGGGTTTCAACAAATTAGTCATATT  
TTTAACTAAATTGCATGTATAATTAGGGTTTCAACAAATTAGTCATATT  
TTCTATTCAAATCATGTATAAAAGGGTTTCAACAAATTAGTCATATT

TGAAACAAATATGACCAAAAAAGTTAAAACCCTTAATAAACATAAAGAGTATAAA  
TGAAACAAATATGACCAAAAAAGTTAAAACCCTTTAATTAAACAA  
TGAAACAAATATGACCAAAAAAGTTAAAATCCATAGATTTAATGTGAAAA  
TGAAACAAATATGACCAAAAAAGTTGAAAAACCCTATATATATATATGTA  
TGAAACAAATATACCAAAAAAGTTAAAACCCTTAATAATAATAGTTA  
TGAAACAAATATGACCAAAAAAGTTAAAAACTATTTTTTATAAGATTAT  
TGAAACAAATATAACCAAAAAAGTTAAAACCCTTAATTATACCCATT  
TGAAACAAATATGACCAAAAAAGTTAAAACCCTTTAATTAAACAACTTAT  
TGAAAATAAATATGACCAAAAAAGTTAAAACCCTGAAATATTGTTATAAGGG  
CGAAACAAATATGACCAACAAAATTAAAAACCCCTTCCTCTATATTTTTTA  
TGAAACAAATATGACCAAAAAAGTTAAAACCCTAAATGAATTACAAAATAGCGTG  
TGAAAACAATATGACCAAAAAAGTTAAAACCCTAAAATAAAACATAAAGAGTATAAA  
TGAAACAAATATGACCAAAAAAGTGAACCAATTGAAATCATACAAAAGAAG  
TGAAAATAAATATGACCAAAAAAGTTAAAACCCTGGTAATAAAAGAGTAAGCATATT  
TGAAACAAATATGACCAAAAAAGTTAAAATCCATAGATTTAATGTGAAAAATACGAT  
TGAAACAAATATGACCAAAAAAGTTAAAACCCTATTTAATGTGAAAAATACGCT  
TGAAACAAATATAACCAAAAAAGTTAAAACCCTATAAATTATACCCATTGTTT

# 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 a complementary analysis to RepeatExplorer*



# Using *RE* output for repeat annotation and quantification

---

- Check / correct automatic annotation
- Use all data available from the analysis archive (not just HTML output)
- Interpret the results considering your experimental setup
  - e.g., proportion of reads in clusters vs. singlet reads depends on coverage
  - efficiency of automatic annotation depends on organism (taxon, genome composition)
- *More info in our practical tutorials*