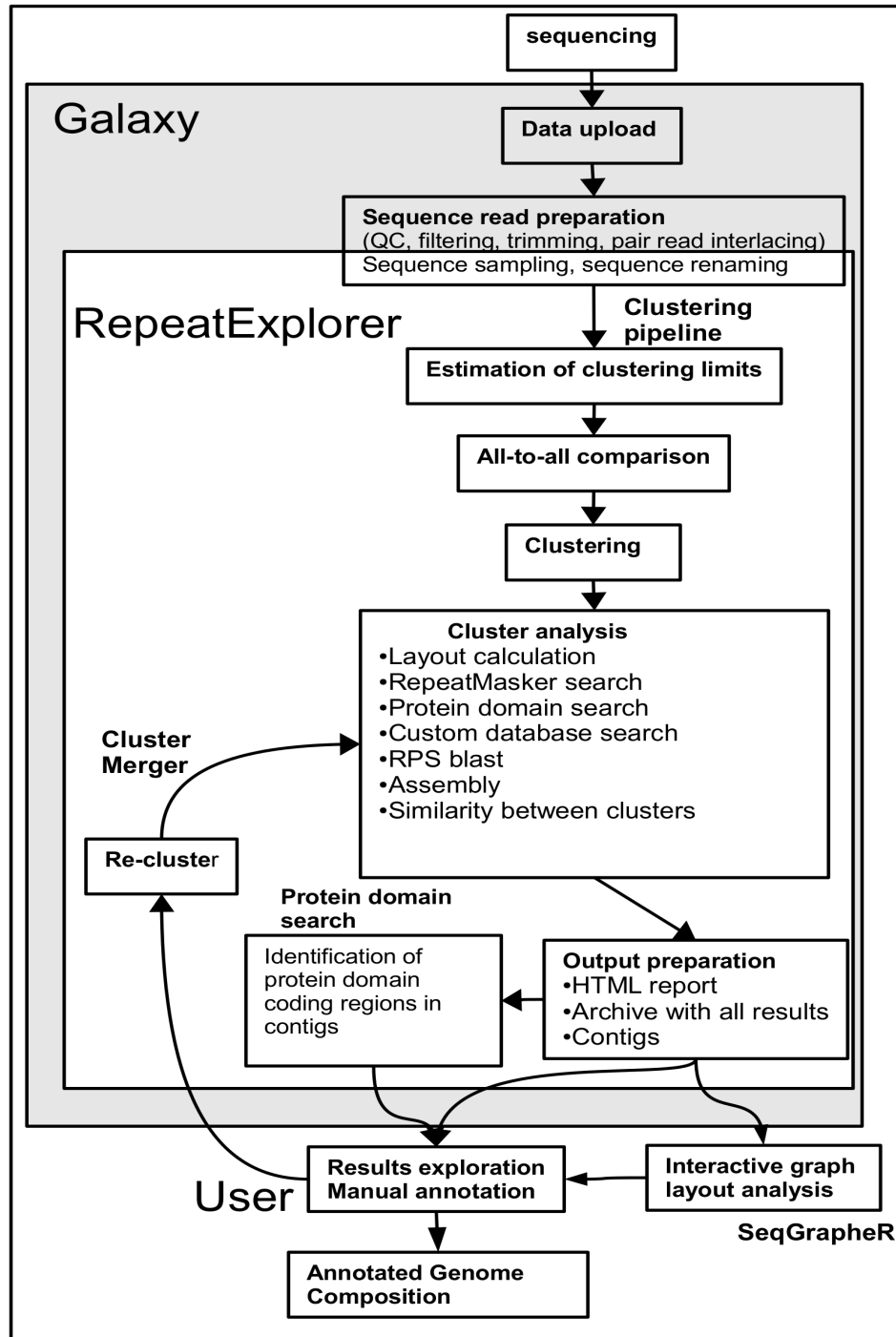


Using *RepeatExplorer* output for repeat
annotation and quantification

RepeatExplorer pipeline



The pipeline combines all principles of repeat identification:

De novo (“repetitiveness”)
(graph-based clustering can handle millions of reads)

Structure-based approaches

- shape of cluster graphs
- investigation of assembled contigs

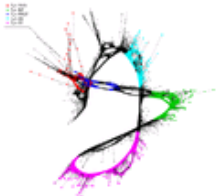
Similarity searches

- protein domains (blastx)
- CDD (RPS-BLAST)
- RepeatMasker

Repeat annotation is achieved by combining all three approaches

Summary table

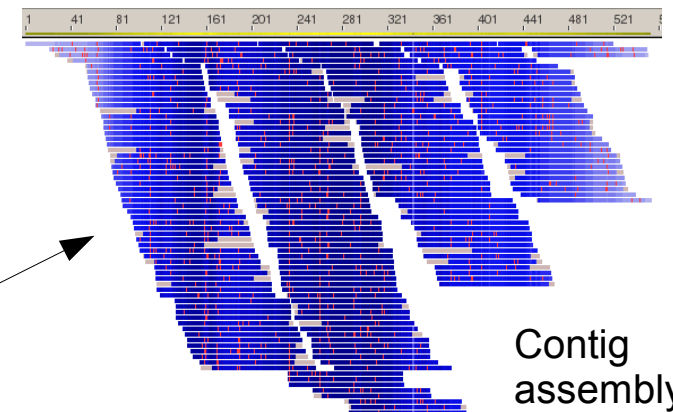
Top clusters

cluster	total length [bp]	number of reads	Genome proportion[%]	cumulative GP [%]	Repeat Masker	Domain hits	Layout	All missing mates [%]	Missing mates with no similarity hit [%]	Portion of similarity hits to other clusters[%]	Outside reads with similarity [%]
1 CL1	1611900	16119	18.00	18	rRNA (8531hits, 50.7%) DNA.MULE.MuDR. (368hits, 2.08%) LTR.Gypsy (7hits, 0.0161%) Satellite (1hits, 0.0023%)			4.746	4.746	0.00000	0.000
2 CL2	988500	9885	11.10	29	LTR.Copia (5818hits, 51.2%) Low_complexity (50hits, 0.202%) Simple_repeat (51hits, 0.168%) LTR.Gypsy (6hits, 0.0245%)	Ty1-RT Ty1/copia Ivana/Oryco (1346 hits 13.6%) Ty1-PROT Ty1.....		9.944	9.924	0.00000	0.000
3 CL3	922400	9224	10.30	39	LTR (2524hits, 19.1%) Satellite (312hits, 2.28%) LTR.Gypsy (54hits, 0.378%) DNA.CMC.EnSpm (63hits, 0.287%) DNA.hAT.Tip100 (3hits, 0.0194%)	Ty3-RT Ty3/gypsy chromovirus (2 hits 0.0217%)		25.520	25.520	0.00012	0.043
4 CL4	900800	9008	10.10	50	Low_complexity (340hits, 1.23%) LTR.Copia (100hits, 0.502%) Simple_repeat (1hits, 0.00422%) LTR.Gypsy (1hits, 0.00389%)	Ty3-RT Ty3/gypsy Ogre/Tat (1 hits 0.0111%)		6.083	6.083	0.00000	0.000

Detailed information for the largest clusters

RepeatExplorer output files

The screenshot shows the RepeatExplorer file explorer. The left pane displays a directory tree with 'PST_max' as the root. Under 'seqClust', there are subdirectories for 'assembly', 'blastx', 'clustering', and 'clusters'. The 'clusters' directory is expanded, showing subdirectories for 'dir_CL0001' through 'dir_CL0012'. 'dir_CL0004' is selected and highlighted in green. The right pane shows a list of files for 'dir_CL0004'. The files include 'blast' (20 items), 'CL4_html' (6 items), 'connections' (238 items), 'ACE_CL4.ace' (16,1 MiB), 'CL0004_blastx.csv' (510 B), 'CL0004_domains.csv' (446 B), 'CL4.GL' (42,8 MiB), 'contigs_CL4' (755,3 KiB), 'contigs_CL4.minRD5' (343,5 KiB), 'contigs_CL4.minRD5_sort-GR' (343,5 KiB), 'contigs_CL4.minRD5_sort-length' (343,5 KiB), 'contigs_CL4.prof' (1,5 MiB), 'contigs_CL4.prof.pdf' (137,3 KiB), 'reads.fas' (8,6 MiB), 'reads.ids' (1016,6 KiB), 'RM-custom_output_table' (1,4 KiB), 'RM-custom_reads.fas.out' (2,1 MiB), 'RM-custom_reads.fas.tbl' (1,9 KiB), 'RM-reads.fas.out' (2,1 MiB), 'RM-reads.fas.tbl' (2,4 KiB), and 'RM_output_table' (1,2 KiB).



Protein domains

Domain	ID	Type	Lineage	Hits	MeanScore
Ty3-INT	MedT2_ID89	Ty3/gypsy	chromovirus	96	49.7
Ty3-INT	PopT1_ID21	Ty3/gypsy	chromovirus	74	46.7
Ty3-RT	PetI1_ID72	Ty3/gypsy	chromovirus	71	60.4
Ty3-RT	CRR3_ID17	Ty3/gypsy	chromovirus	56	48.4
Ty3-GAG	PetI1_ID72	Ty3/gypsy	chromovirus	49	35.5
Ty3-INT	MedT1_ID93	Ty3/gypsy	chromovirus	42	58.8
Ty3-INT	PopT2_ID20	Ty3/gypsy	chromovirus	41	45.3
Ty3-PROT	lpoB1_ID87	Ty3/gypsy	chromovirus	37	45.8
Ty3-PROT	PetI1_ID72	Ty3/gypsy	chromovirus	36	45.9
Ty3-RH	BraR5_ID117	Ty3/gypsy	chromovirus	31	57.7
Ty3-INT	CRA2a_ID5	Ty3/gypsy	chromovirus	31	53.4
Ty3-GAG	PetI2_ID76	Ty3/gypsy	chromovirus	30	35.0

Similarity searches

SW score	perc div.	perc del.	perc ins.	query sequence	position begin	position end	query (left)	matching repeat	repeat class/family	position begin	position end	repeat (left)	ID
602	17.0	0.0	0.0	SJM_1000031f	1	100	(0)	C Gypsy-33_ST-I-int	LTR/Gypsy	(4879)	1441	1342	1
256	14.3	0.0	0.0	SJM_1000031r	1	42	(58)	+ Gypsy-33_ST-I-int	LTR/Gypsy	1224	1265	(5055)	2 *
581	6.8	0.0	0.0	SJM_1000031r	28	100	(0)	+ Gypsy-33_ST-I-int	LTR/Gypsy	1197	1269	(5051)	2
885	3.0	0.0	0.0	SJM_1000270f	1	100	(0)	+ Gypsy-33_ST-I-int	LTR/Gypsy	3002	3101	(3219)	3
802	5.0	5.0	0.0	SJM_1000270r	1	100	(0)	C Gypsy-33_ST-I-int	LTR/Gypsy	(2983)	3337	3233	4
680	11.0	0.0	0.0	SJM_1002242f	1	100	(0)	+ Gypsy-33_ST-I-int	LTR/Gypsy	2328	2427	(3893)	5

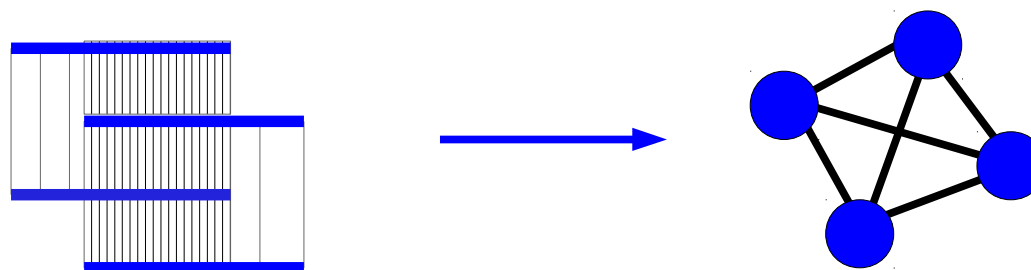
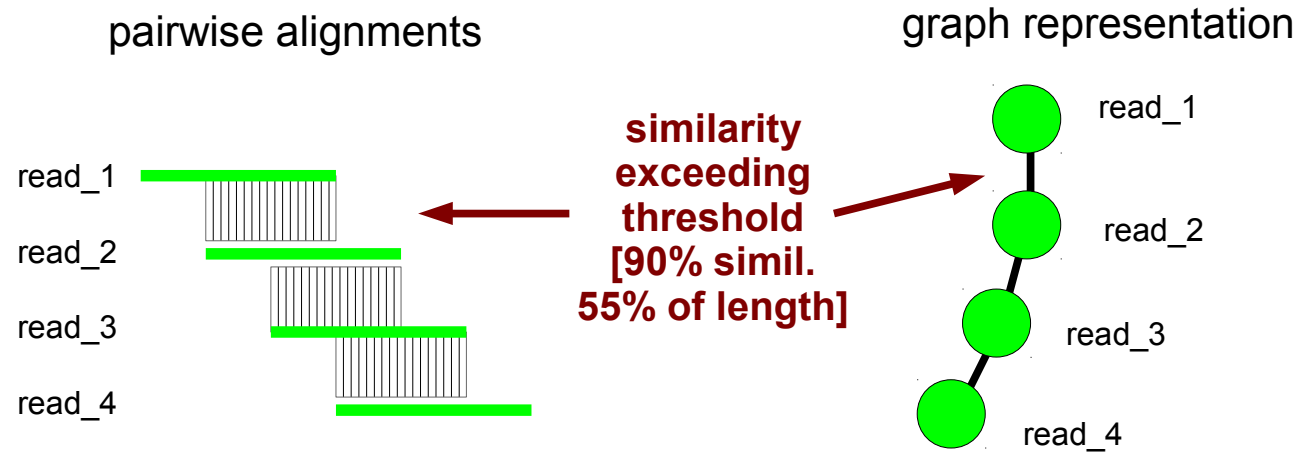
Repeat annotation

For repeat annotation in individual clusters, consider:

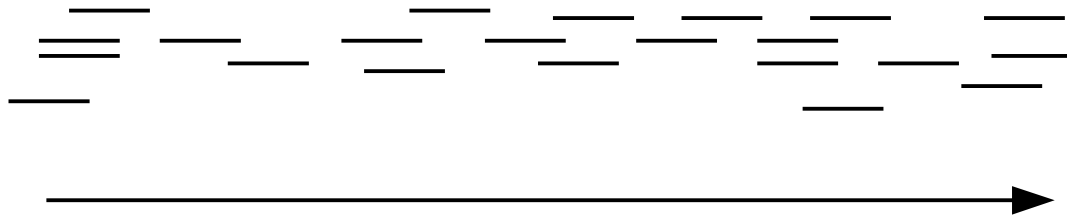
- Shape of cluster graph
- Significant similarity to conserved types of repeats (45S and 5S rDNA, *cpDNA*)
- Similarity to conserved protein domains
- Structural features of assembled contigs (dot-plot)
- Variability and masked regions of reads in contig assembly
- (blastn and blastx to GenBank)
- **New !** Cluster connections via paired-end reads
- **New !** LTR + primer binding site (PBS) detection

Be prepared to leave some clusters without annotation !

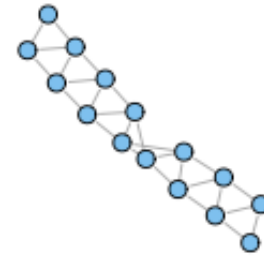
Graph shapes



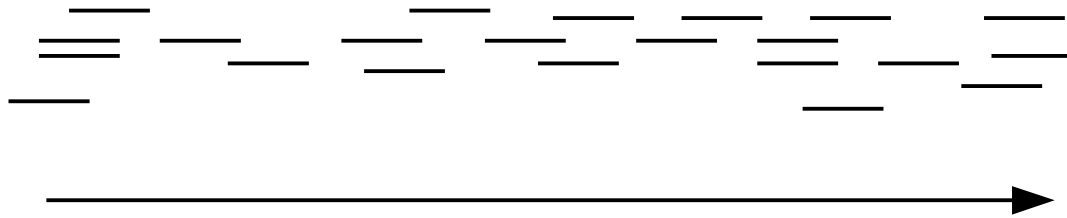
Linear graphs



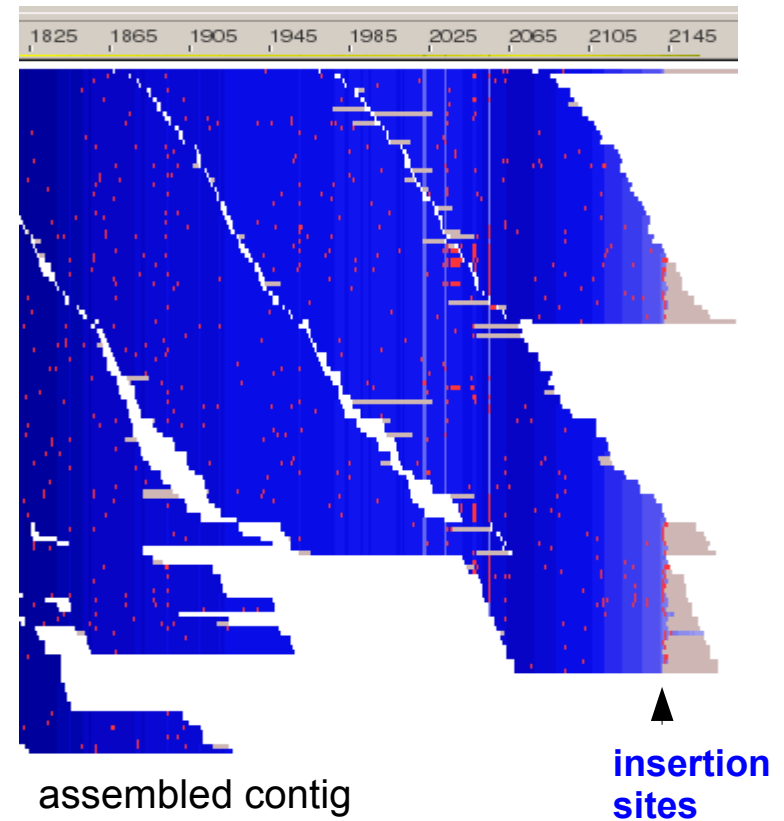
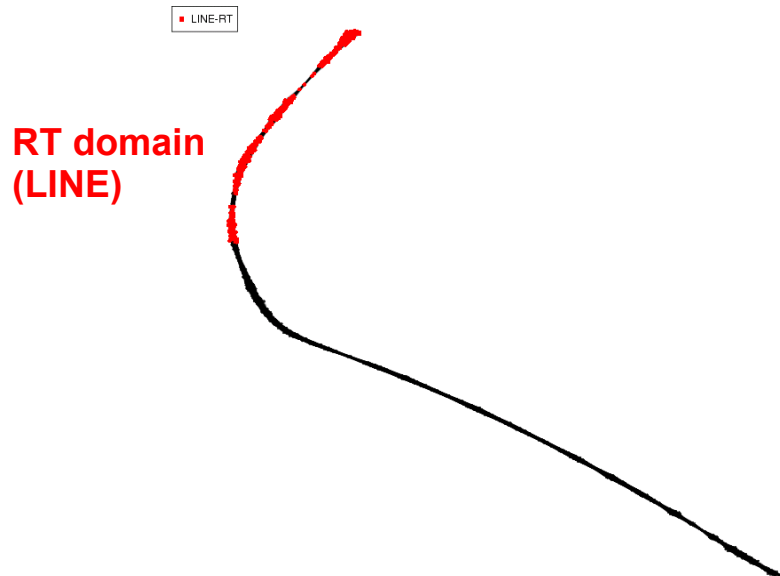
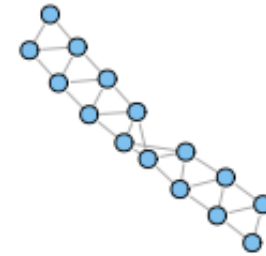
DNA transposons, LINEs, ...



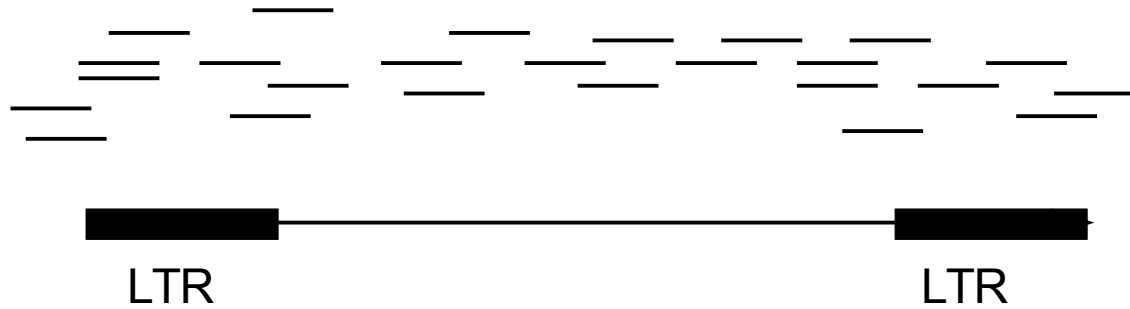
Linear graphs



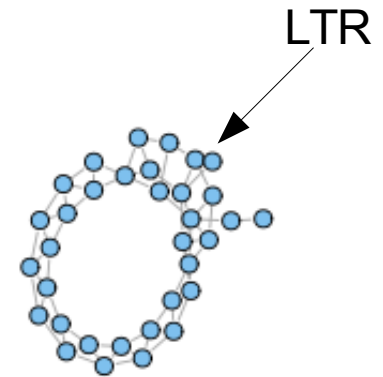
DNA transposons, LINEs, ...



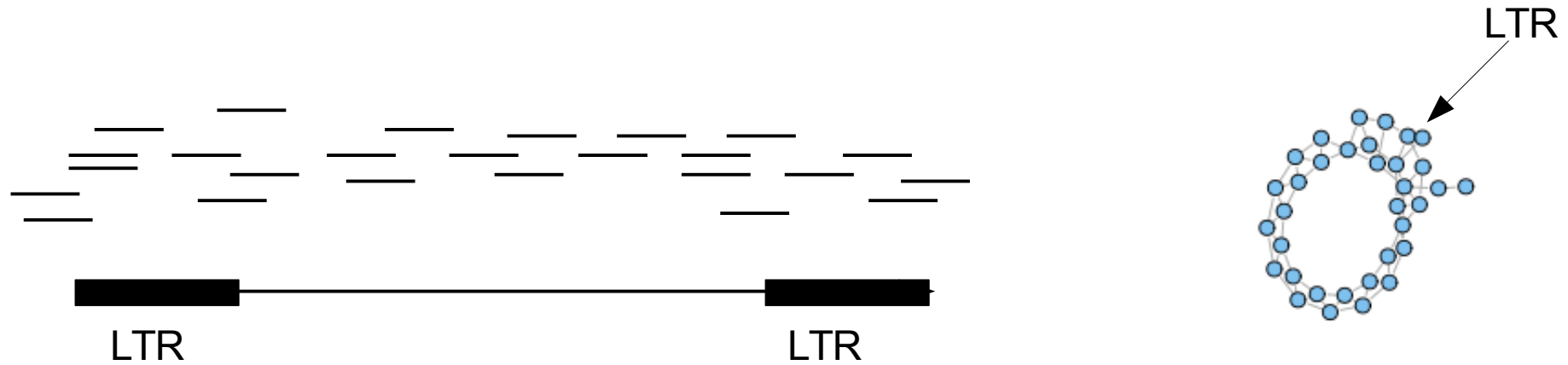
Linear / large circular graphs



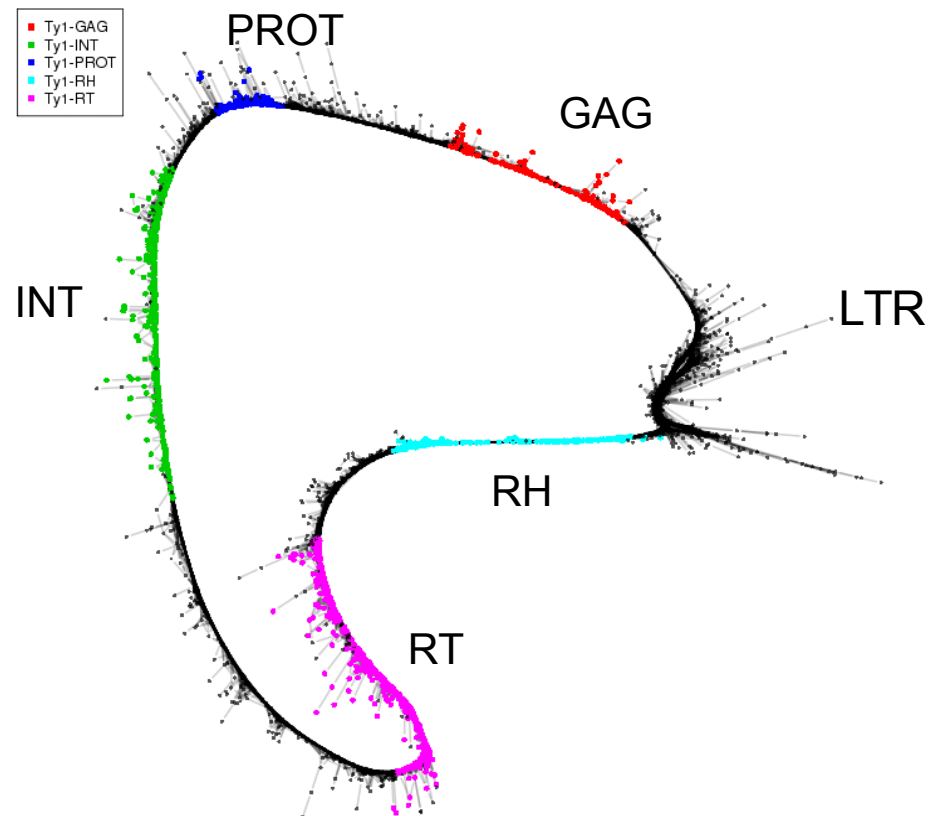
LTR-retrotransposons



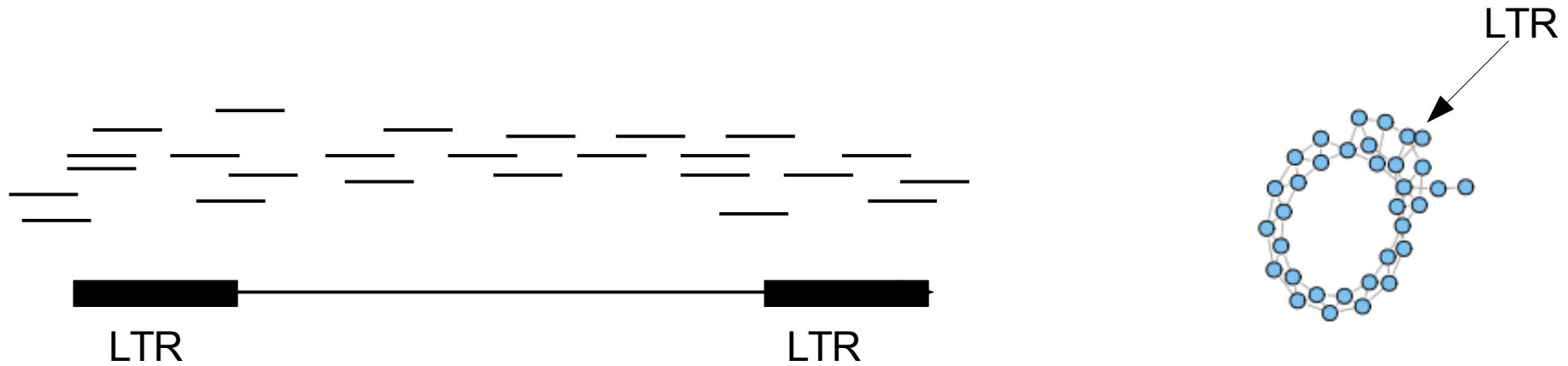
Linear / large circular graphs



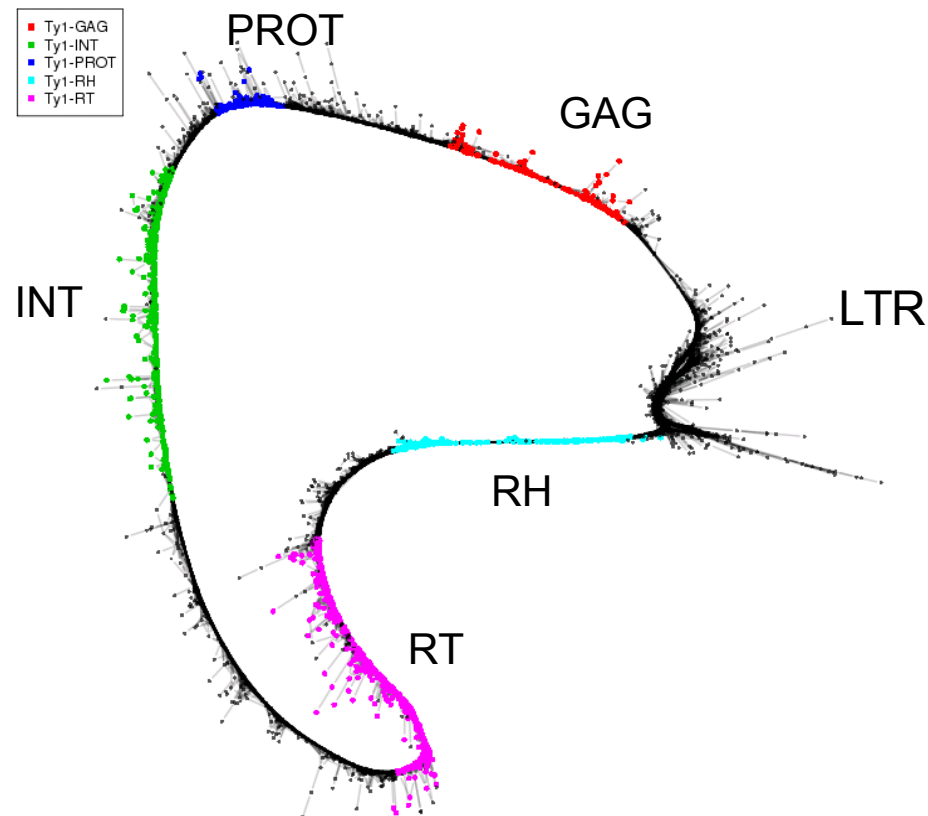
LTR-retrotransposons



Linear / large circular graphs



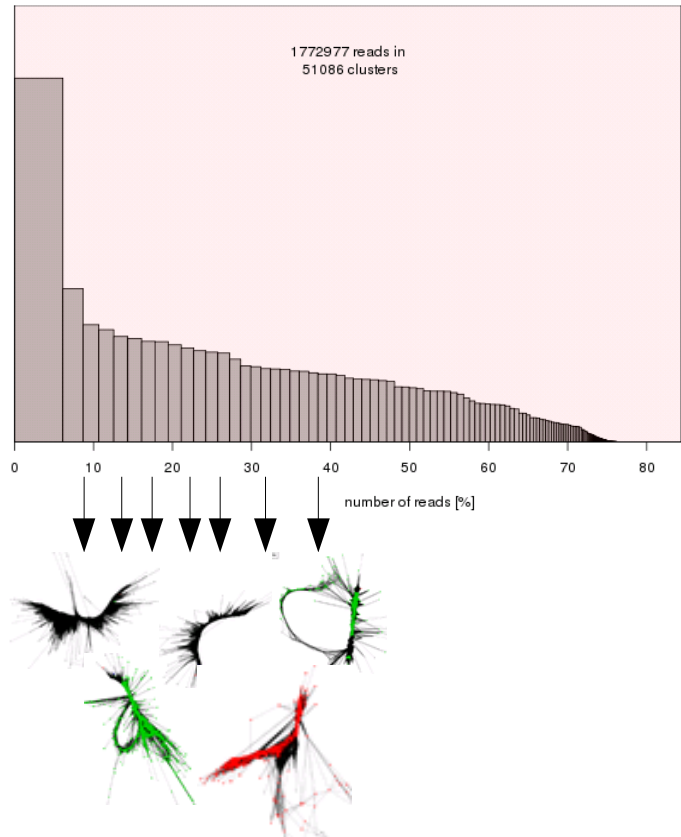
LTR-retrotransposons



Cluster fragmentation
leads to linear graphs !

Linear / large circular graphs

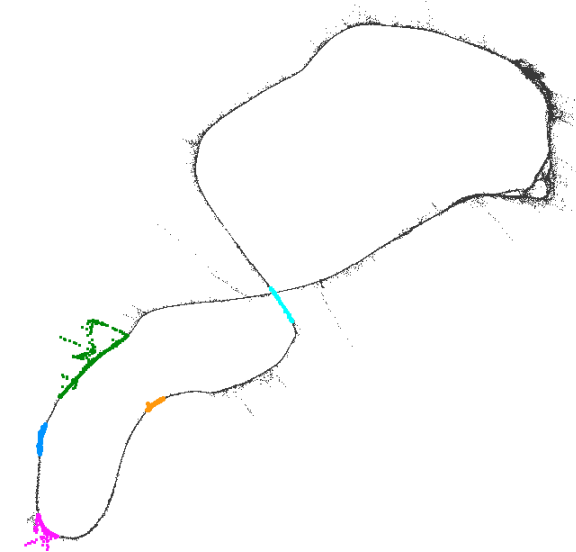
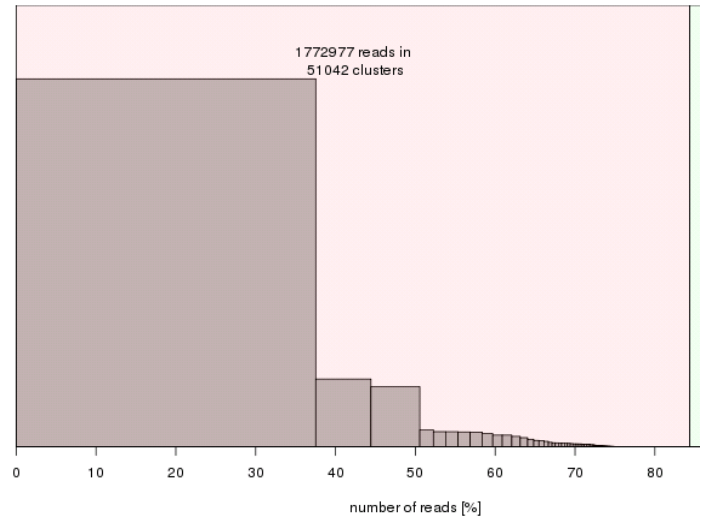
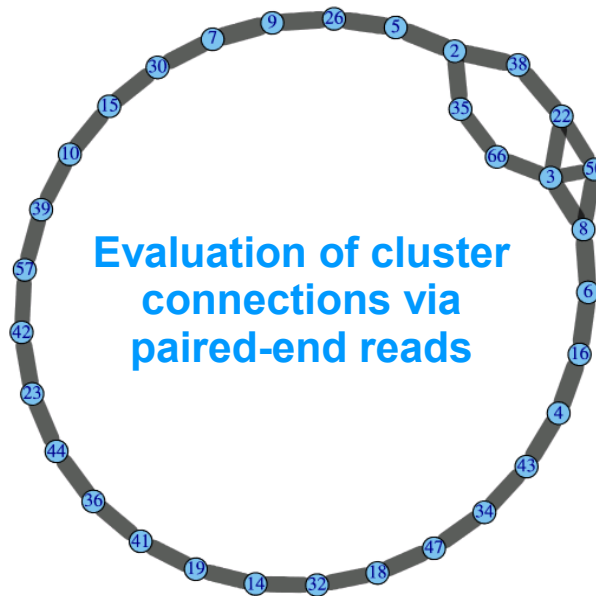
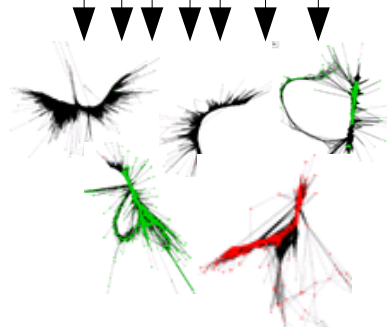
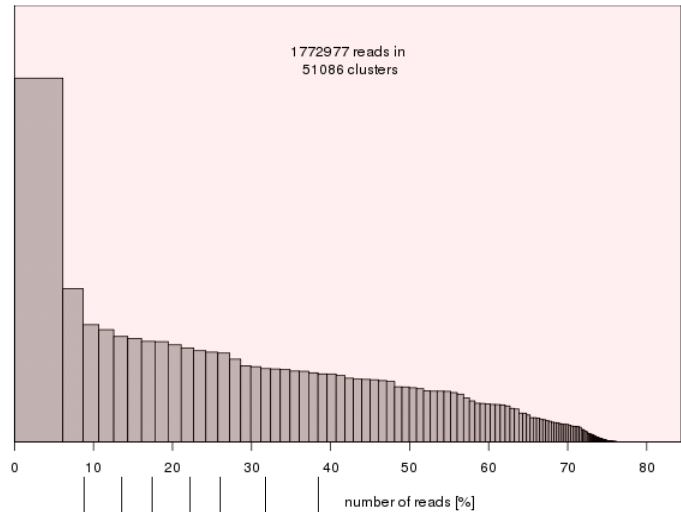
Cluster fragmentation leads to linear graphs !



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

Linear / large circular graphs

Cluster fragmentation leads to linear graphs !



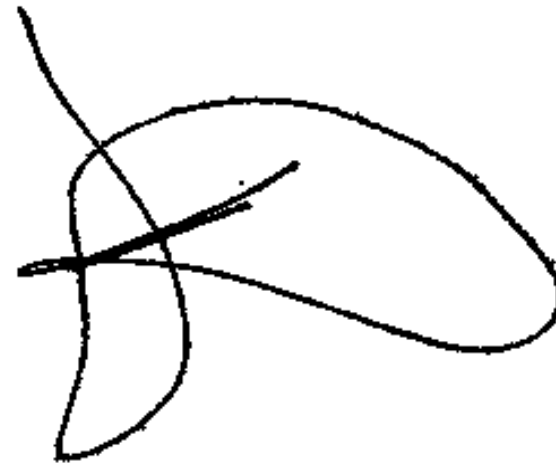
LTR-retrotransposon Ogre in *Vicia pannonica* (~ 40% of the genome)

Linear / large circular graphs

Cluster fragmentation leads to linear graphs !

rDNA (2045hits, 80.8%)
rDNA.18S (363hits, 14.4%)
X25SrDNA (84hits, 3.31%)
rDNA.26S (5hits, 0.198%)

45S rDNA

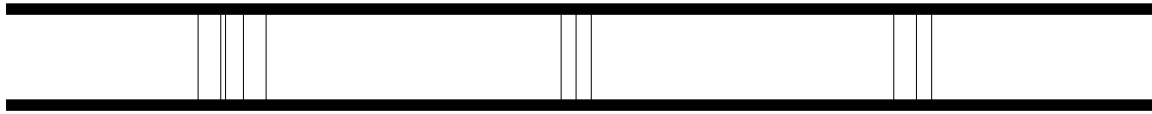


plastid DNA
(contamination)

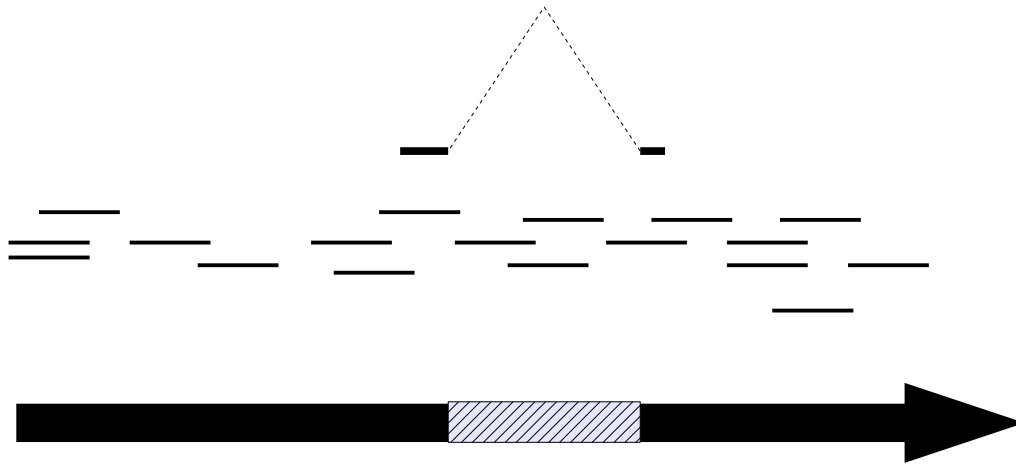
Organelle.plastid (15931hits, 89.5%)
Low_complexity (77hits, 0.225%)
LTR.Gypsy (1hits, 0.00215%)

Both can be easily recognized by similarity searches

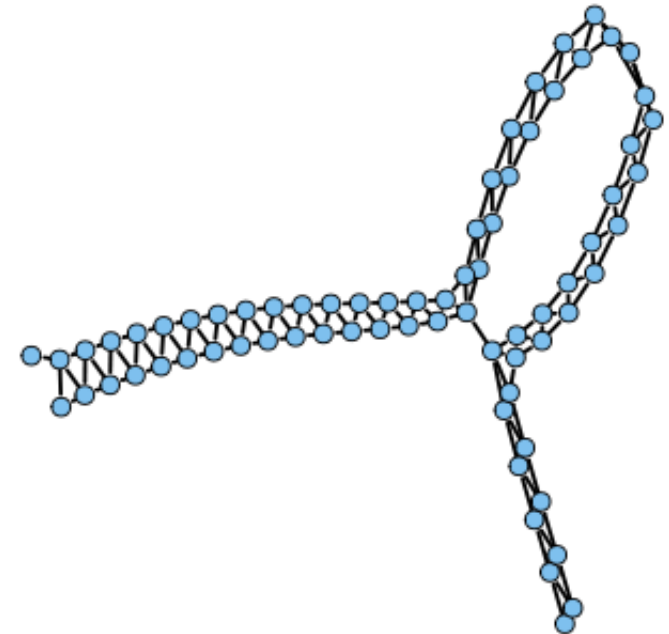
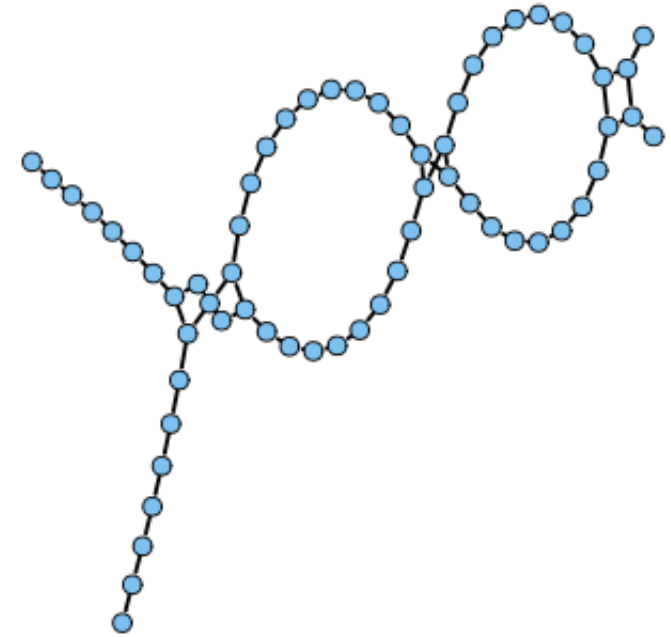
It's getting complicated...



Repeat variants with partial similarity
(depends on similarity threshold)



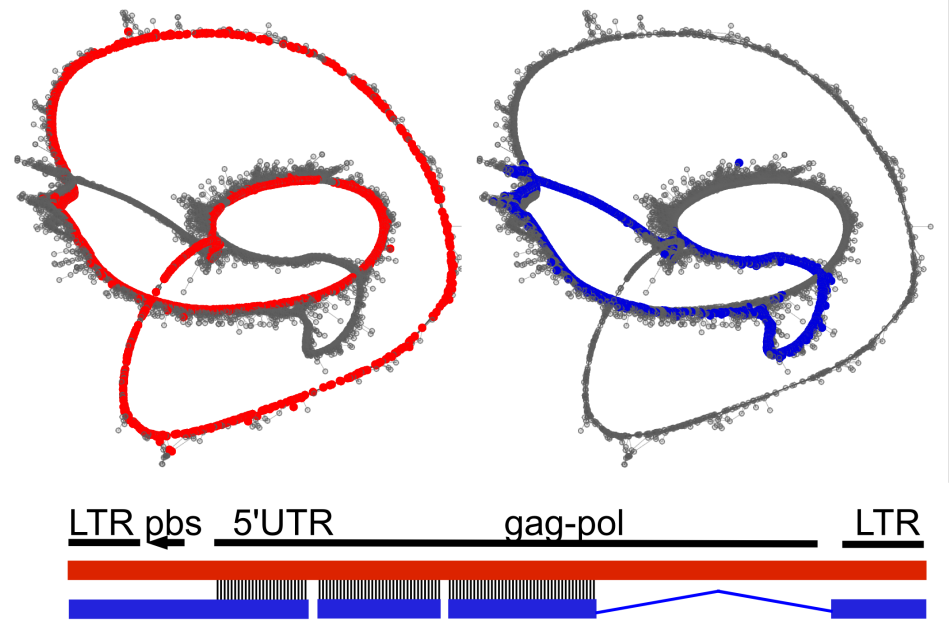
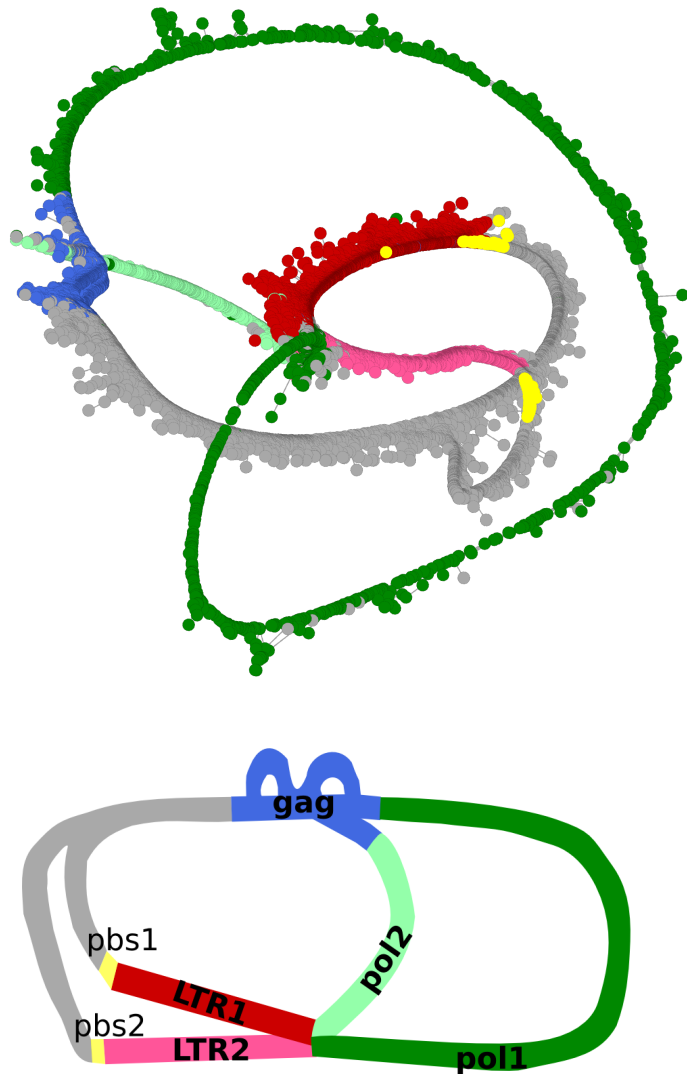
Variants differing in indel



It's getting complicated...

...but it has some meaning

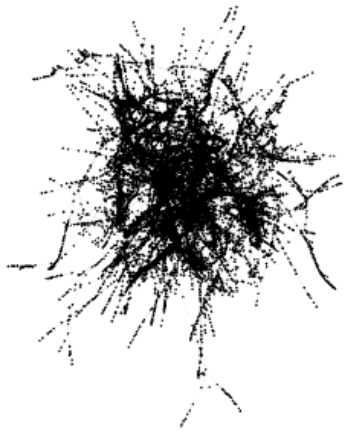
(in this case, presence of element variants differing in LTR sequence and in deletion within gag-pol region)



It's getting complicated...

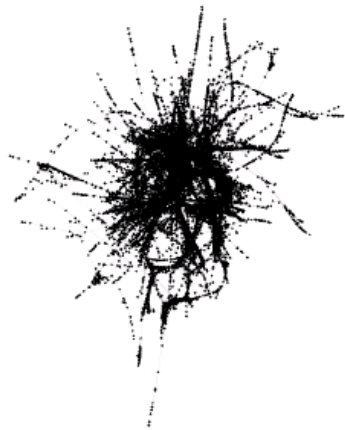
...but it has some meaning

OR NOT ? ;-)



CL1

Number of reads: 42851
Number of pairs: 717667
Density: 0.0007817
Diameter: NA
Mean edge weight: 130.31
Max. degree: 257



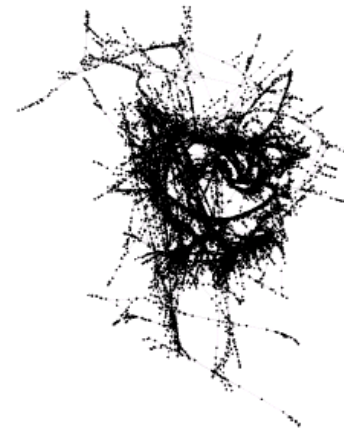
CL2

Number of reads: 47505
Number of pairs: 1460012
Density: 0.001294
Diameter: NA
Mean edge weight: 129.94
Max. degree: 647



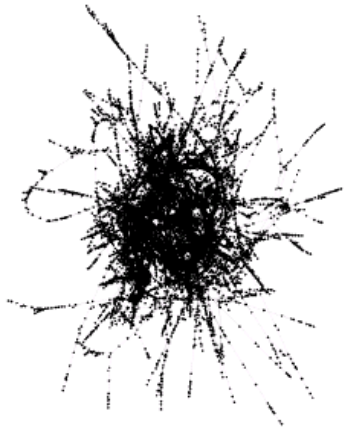
CL3

Number of reads: 44258
Number of pairs: 473407
Density: 0.0004834
Diameter: NA
Mean edge weight: 129.42
Max. degree: 186

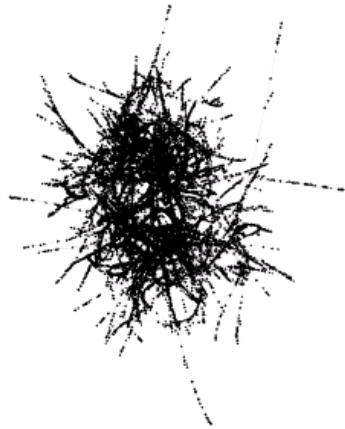


CL4

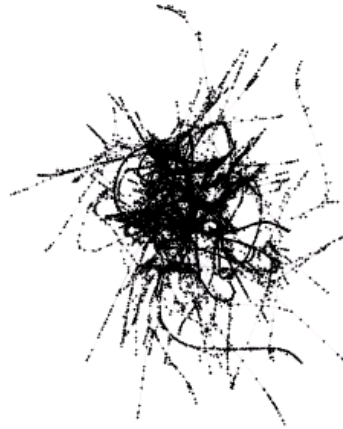
Number of reads: 46876
Number of pairs: 1022899
Density: 0.000931
Diameter: NA
Mean edge weight: 132.18
Max. degree: 290



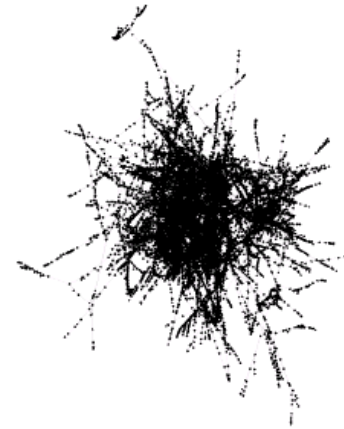
CL5



CL6



CL7

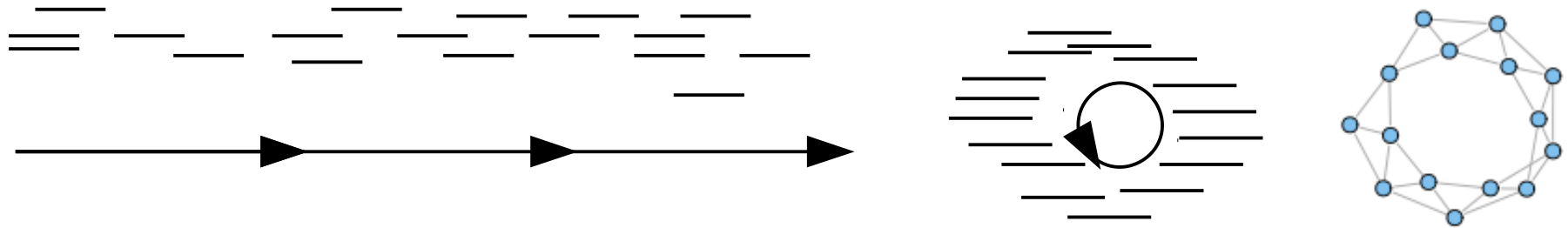


CL8

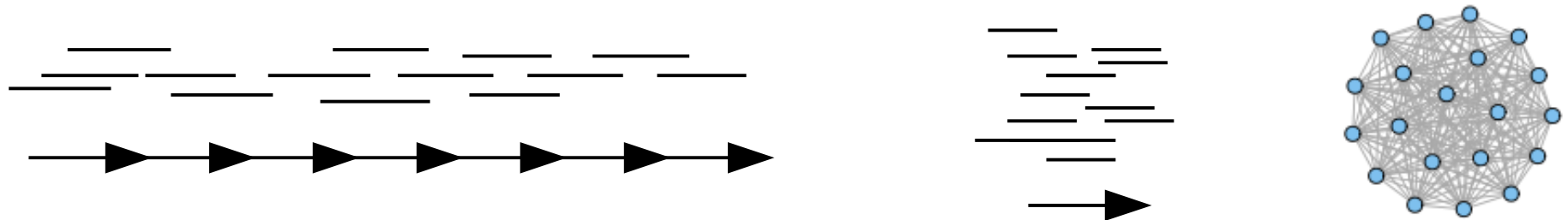
Circular graphs

TANDEM REPEATS

Read length $\ll$ monomer

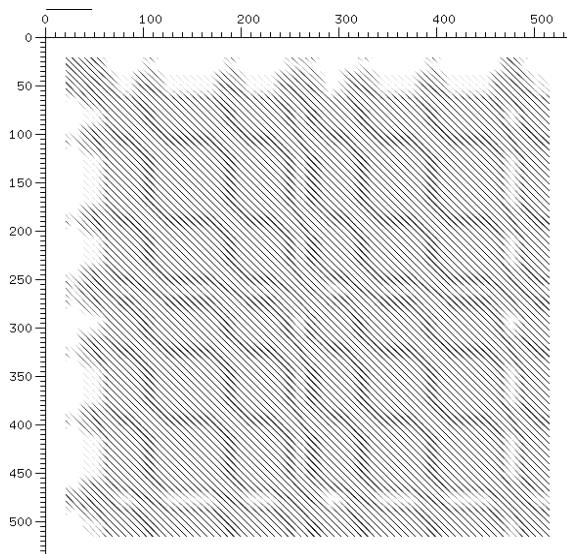
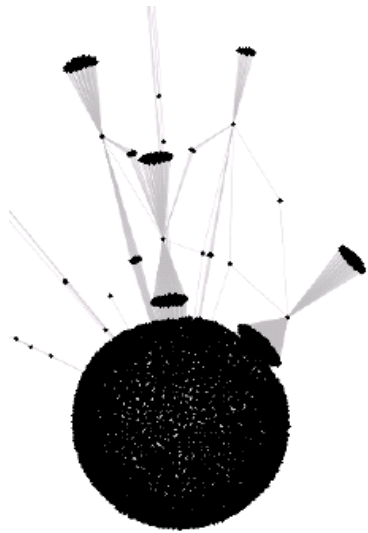


Read length $\geq$ monomer

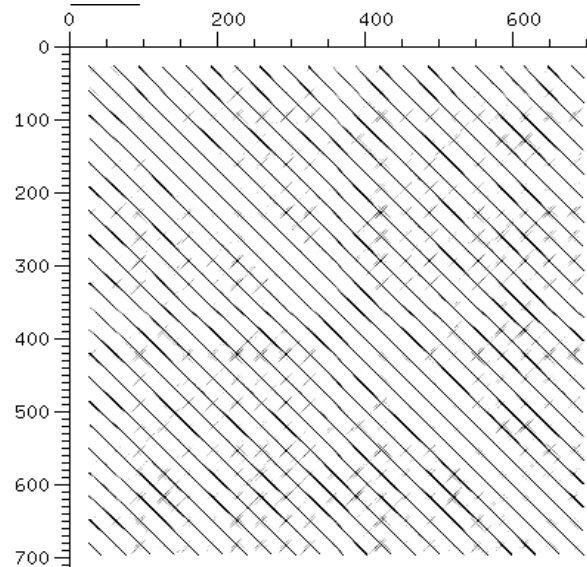
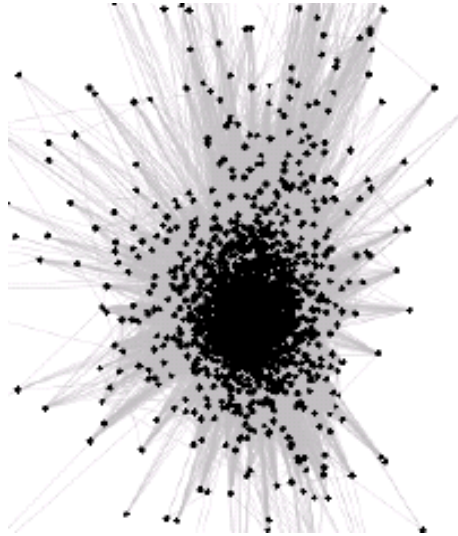


Circular graphs

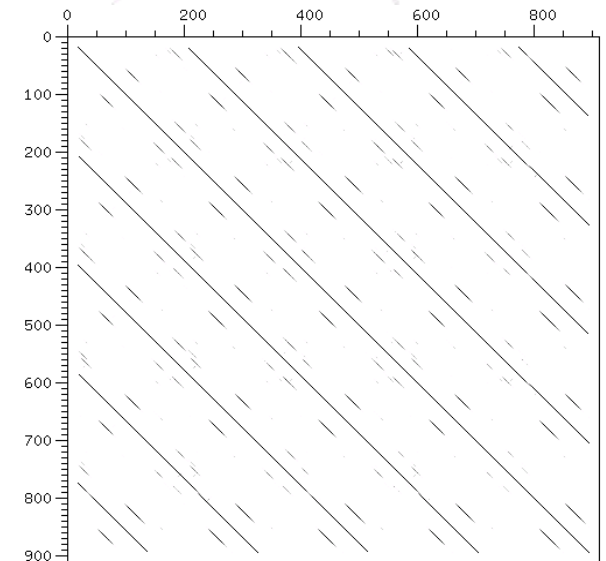
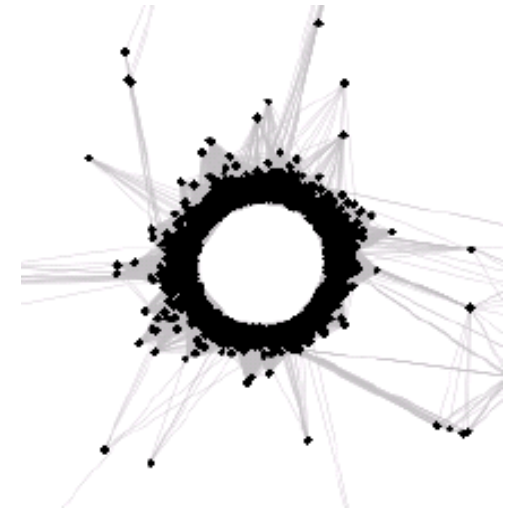
microsatellite (6bp)
(GAACCT)_n



satellite 35 bp

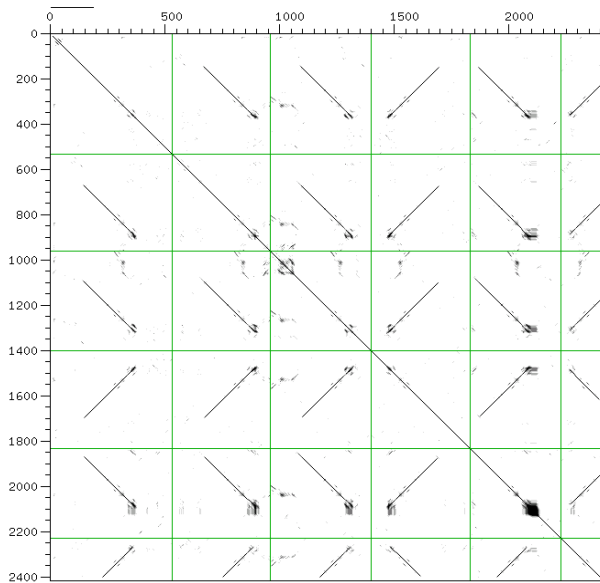
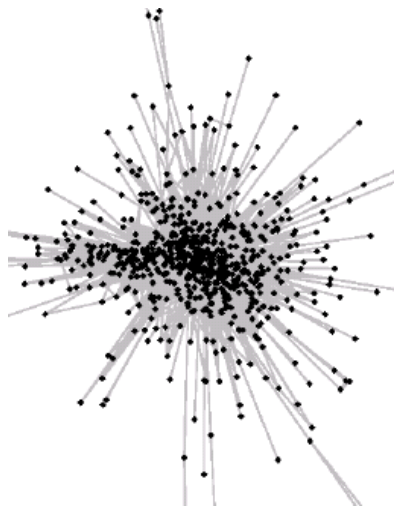


satellite 190 bp

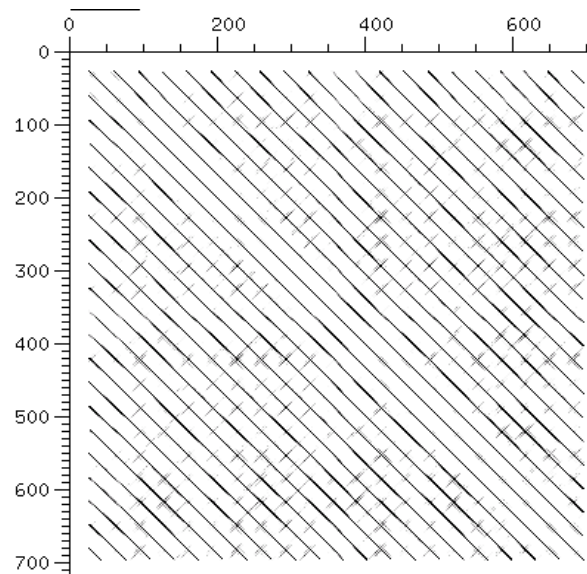
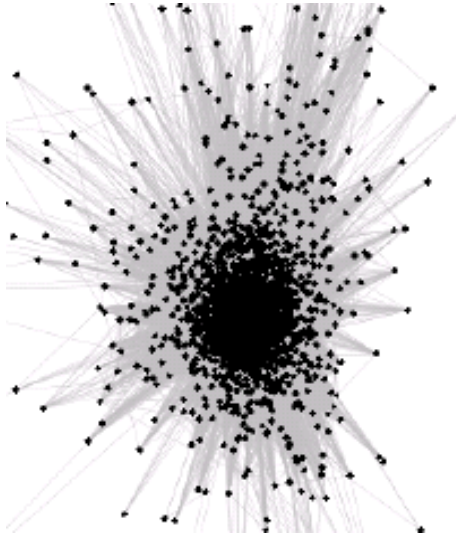


Circular graphs

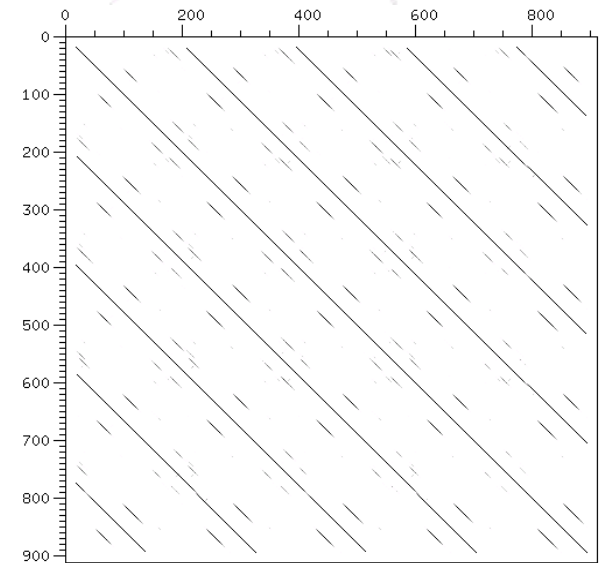
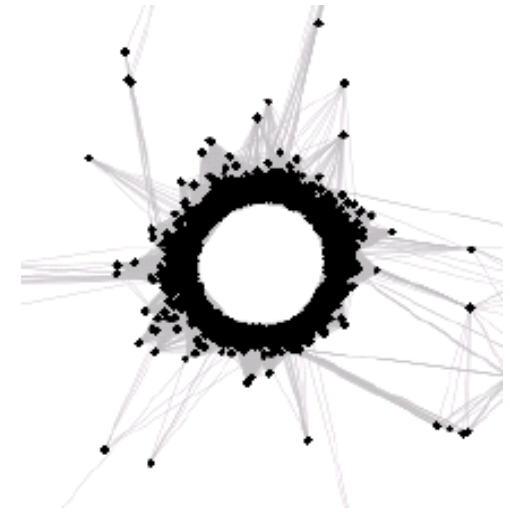
SINE



satellite 35 bp

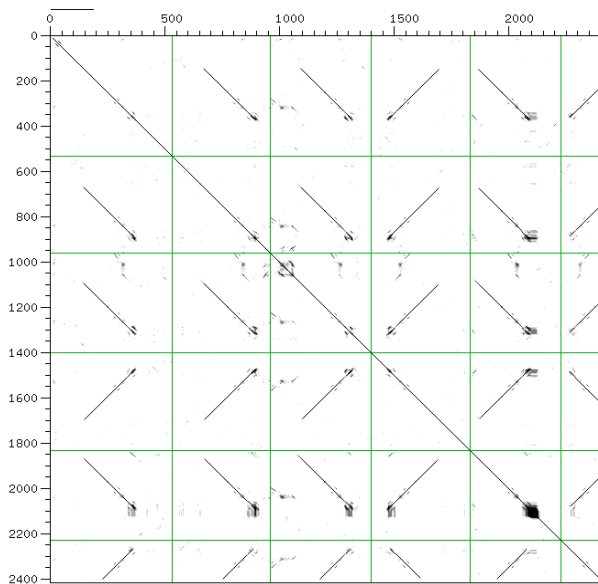
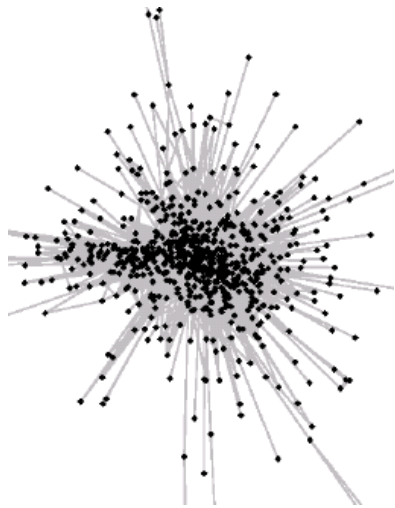


satellite 190 bp

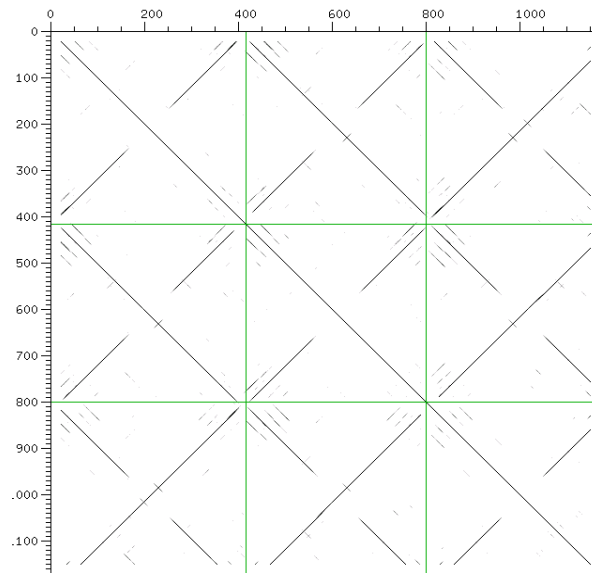


Circular graphs

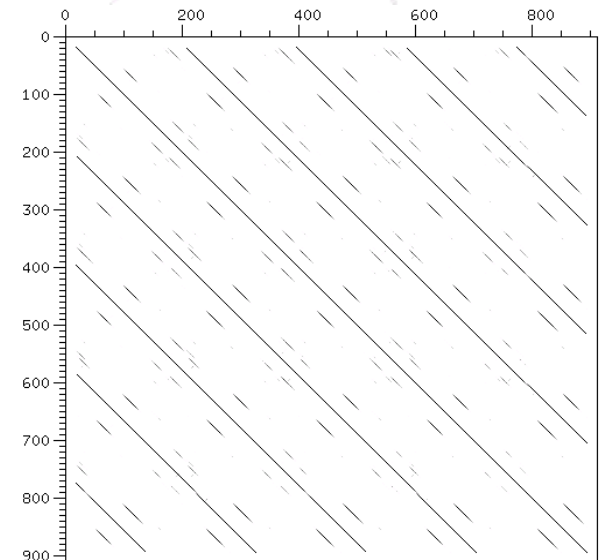
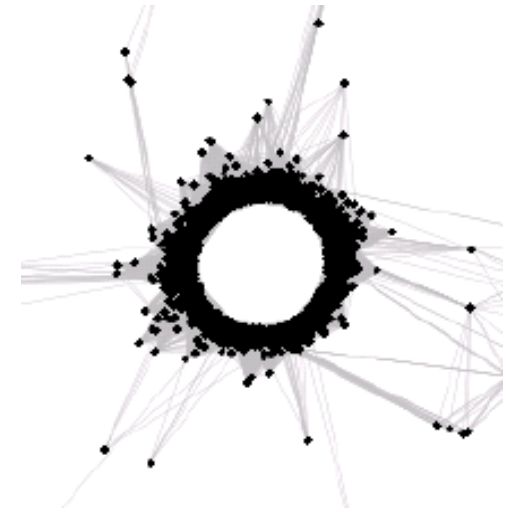
SINE



MITE
(foldback)

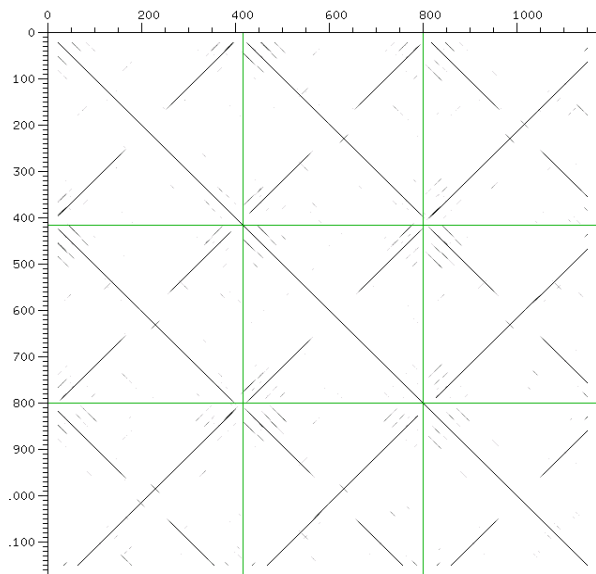


satellite 190 bp



Insertion sites of mobile elements

MITE (foldback)

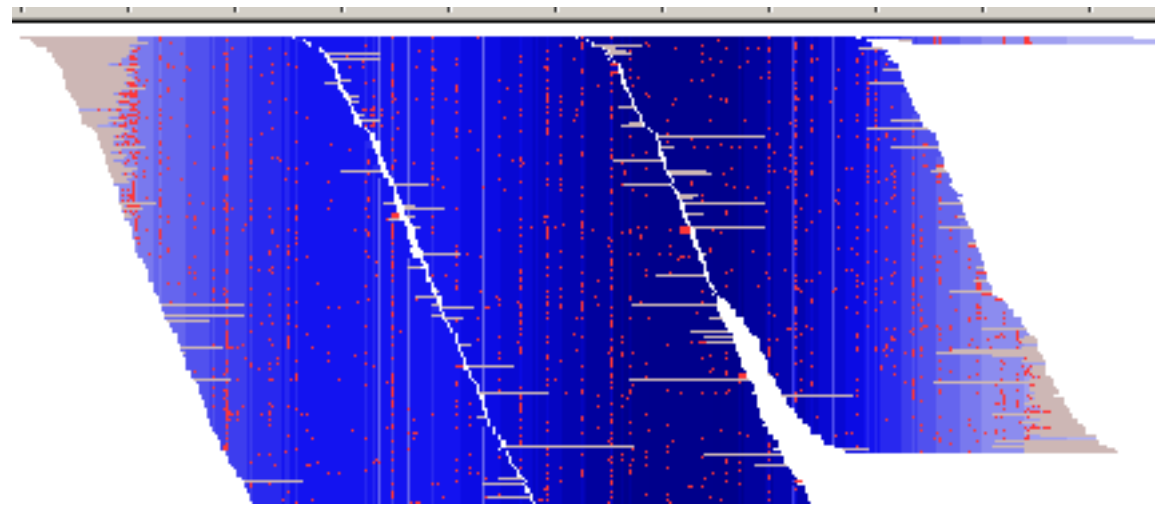


```
ATTCAATAATATAATTTTTTTTAGGGTTTTCAACAATTTTAGTCATATT  
TTTATTAATAATTTAAATAAATAAGGGTTTTGAACAATTTTAGTCATATT  
CTTAATTCAATTATAAAAGATTTAAAGGGTTTTTAACAATTTTAGTCATATT  
CCTCGTAATAATAATTAATAAATTTTTGGATTTTCAACCATTTTAGTCATATT  
AAAATTGTATTTTAAATTTTTTTTAGGGTTTTCAACAATTTTAGTCATATT  
TTGAGCCAAACTAATAGTTAAATTTTAGGGTTTTCAACAATTTTAGTCATATT  
ACACATGCCATTTTATGATAGAAAAAGGGTTTTCAACAATTTTAGTCATATT  
TAGTAGTAGAAATGGTTTTAGGGTTTAGGGTTTTCAACAATTTTAGTCATATT  
TTTTATTATTTTGTGATTTTAAATAGGGTTTTGAACAATTTTAGTCATATT  
CTTAATTCAATTATAAAAGATTTAAAGGGTTTTTAACAATTTTAGTCATATT  
AAAATTGTATTTTAAATTTTTTTTAGGGTTTTCAACAATTTTAGTCATATT  
GCAGTTTATGGTATAAAATTTTAAAGGGTTTTCAATAATTTTAGTCATATT  
TGTAATTTTATTTGCATATTTTAAAGGGTTTTCAATCTTTTGGTCATATT  
TTTGAAATAAAATAGAGTATAAATAGGGTTTTCAACAATTTTAAATCATATT  
TTAATGGACTAAATGTGTATTTTAAAGGGTTTTCAACAATTTTAGTCATATT  
TTTTAACTAAATTCATGTATTTTAGGGTTTTCAACAATTTTAGTCATATT  
TTCTATTTCAAATCATGTATAAATTAAGGGTTTTCAACAATTTTAGTCATATT
```

← TIR

```
TGAAACAAATATGACCAAAAAAGTTAAAAACCCCTTAAATAAAACATAAAGAGTATAAA  
TGAAACAAATATGACCAAAAAAGTTAAAAACCCATTTTAAATTTAAACA  
TGAAACAAATATGACCAAAAAAGTTAAAAATCCATAGATTTAATGTGAAAA  
TAAACAAATATGACCAAAAAAGTTAAAAACCCATATATATATATGTGA  
TGAAACAAATATACCAAAAAAGTTAAAAACCCCTTAAATAAAATATAGTTA  
TGAAACAAATATGACCAAAAAAGTTAAAAACCTATTTTTTTTATAAGATTAT  
TGAAACAAATATACCAAAAAAGTTAAAAACCCCTTAAATTATACCCATTTT  
TGAAACAAATATGACCAAAAAAGTTAAAAACCCATTTTAAATTTAAACAATTAT  
TGAAATAAATATGACCAAAAAAGTTAAAAACCCCTTAAATATTTGTTATAAGGG  
CGAAACAAATATGACCAACAAATTTAAAAACCCCTTCTCTCTATATTTTTTA  
TGAAACAAATATGACCAAAAAAGTTAAAAACCCCTTAAATGAATTACAAATATGCGTG  
TAAACAAATATGACCAAAAAAGTTAAAAACCCCTTAAATAAAACATAAAGAGTATAAA  
TAAACAAATATGACCAAAAAAGTTAAAAACCCCTTAAATCATACAAAAAAGAAG  
TGAAATAAATATGACCAAAAAAGTTAAAAACCCGGTAATAAATAGAGTAAGCATATT  
TGAAACAAATATGACCAAAAAAGTTAAAAATCCATAGATTTAATGTGAAAAATACGATAT  
TGAAACAAATATGACCAAAAAAGTTAAAAACCCGTATATTTAATGTGAAAAATACGCT  
TGAAACAAATATACCAAAAAAGTTAAAAACCCCTTAAATTATACCCATTTTGTTTT
```

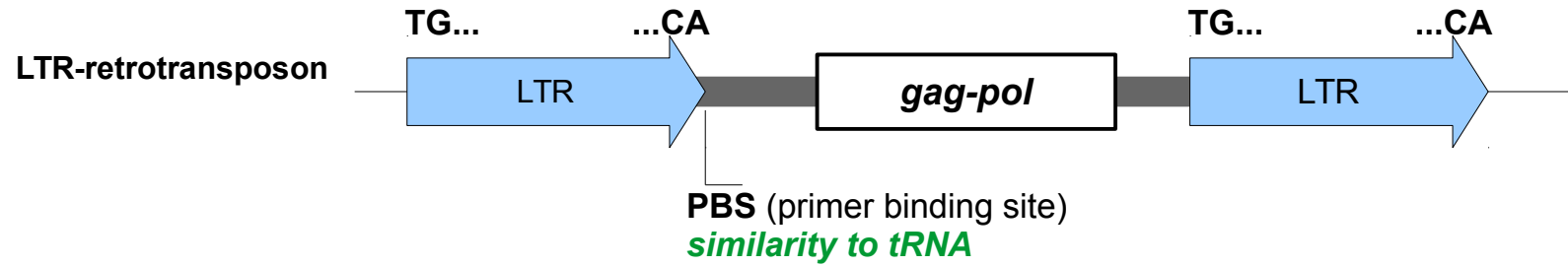
TIR →



contig

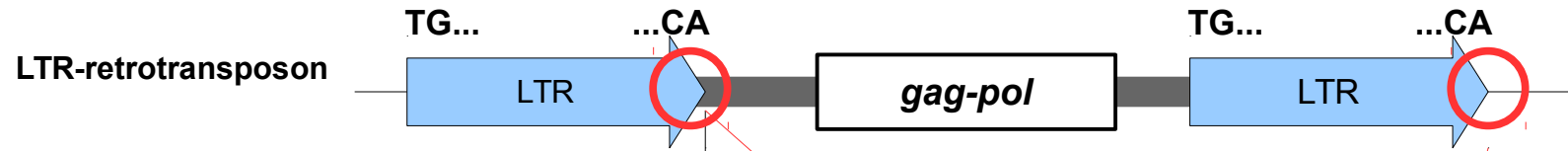
Insertion sites of mobile elements

A new tool for detection of LTR / PBS sites

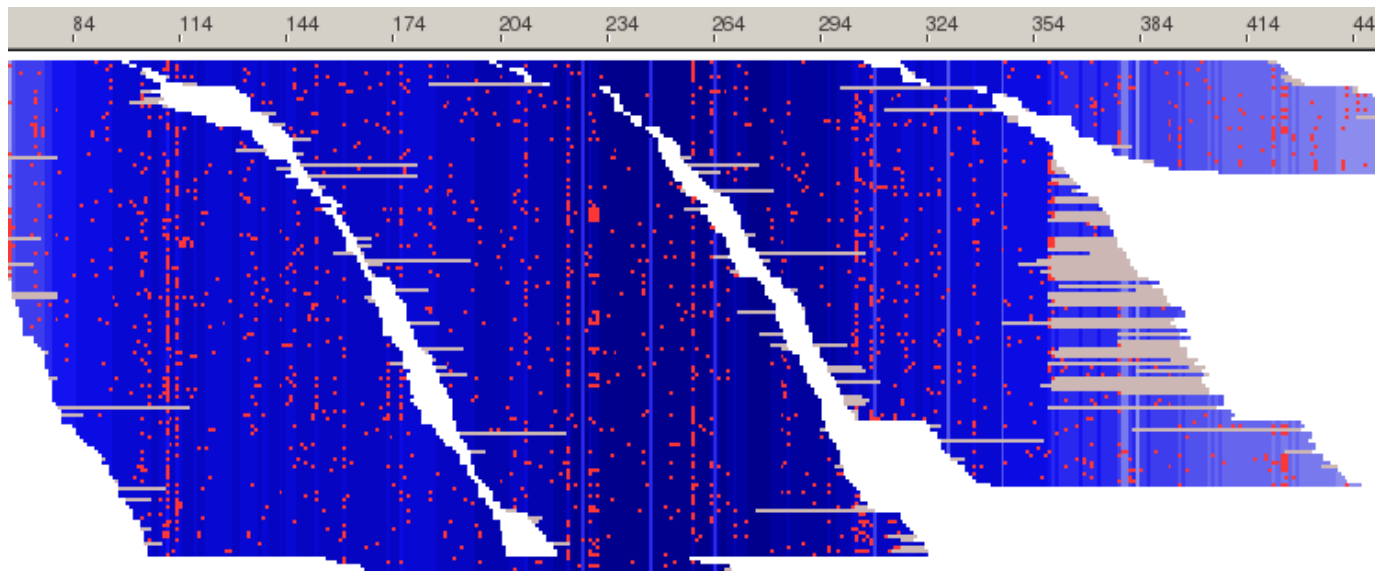


Insertion sites of mobile elements

A new tool for detection of LTR / PBS sites

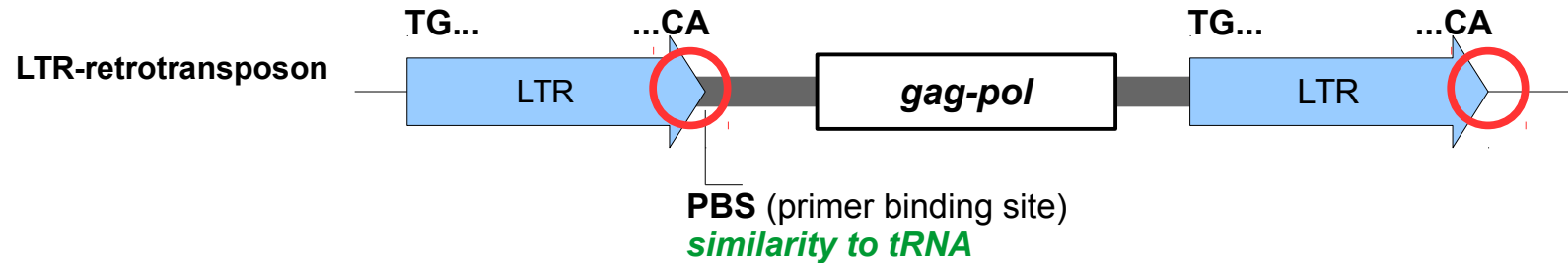


PBS (primer binding site)
similarity to tRNA



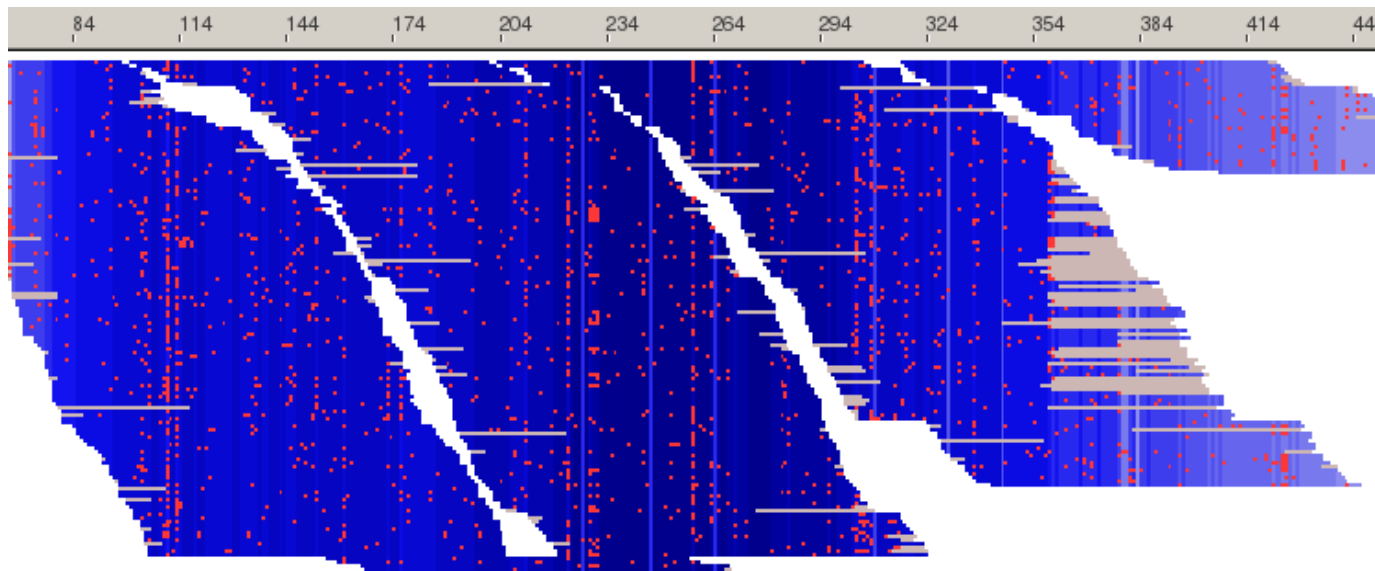
Insertion sites of mobile elements

A new tool for detection of LTR / PBS sites



The tool searches assembled contigs (ACE files) for **differences in proportions of masked reads** accompanied by **occurrence of TG/CA**

↓
...and **similarity to tRNA**

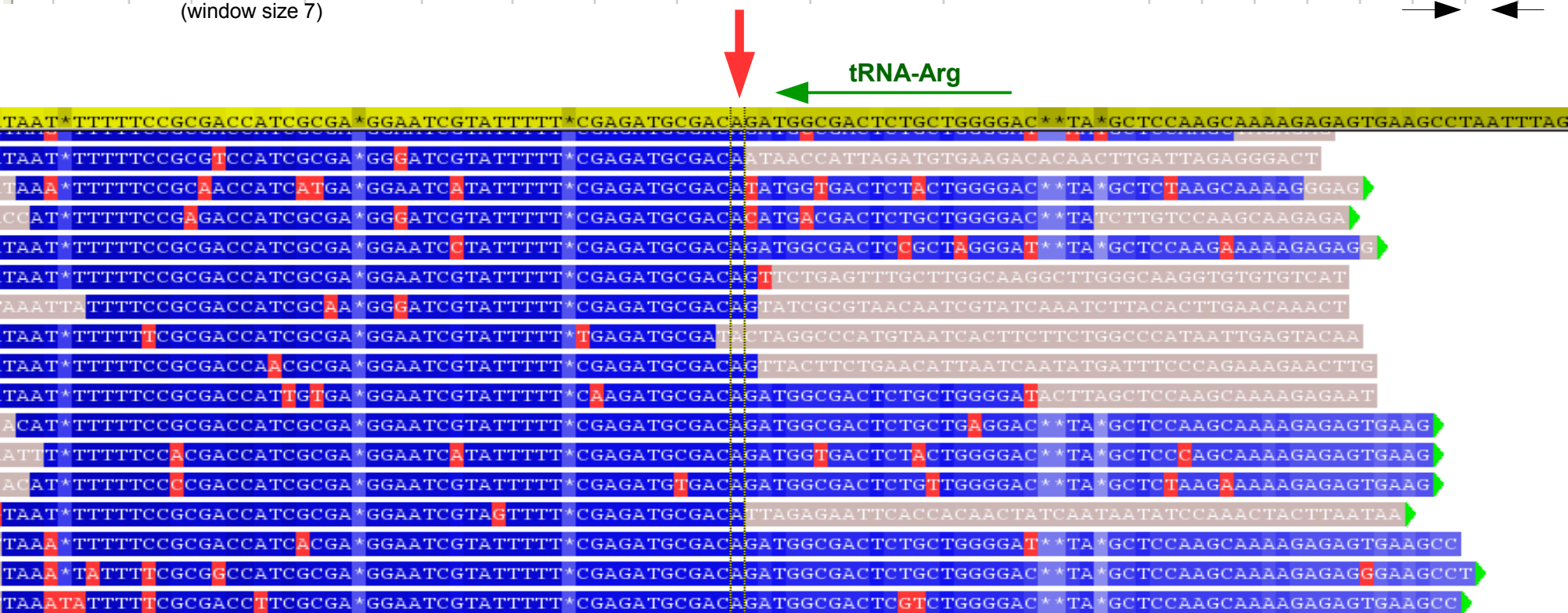


A new tool for detection of LTR / PBS sites



CL	contig	pos.	site	site_depth	out_ma	maske	maske	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)

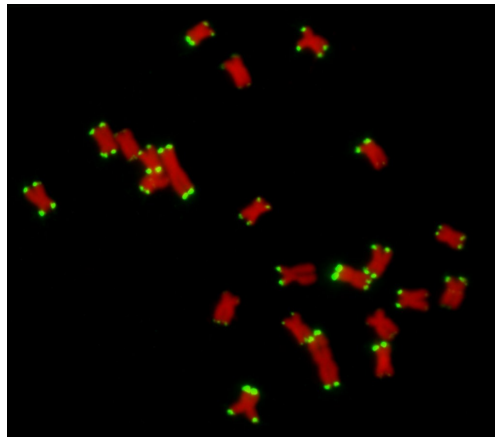


“Chimeric” clusters ?

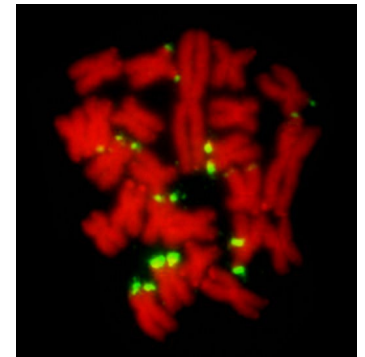
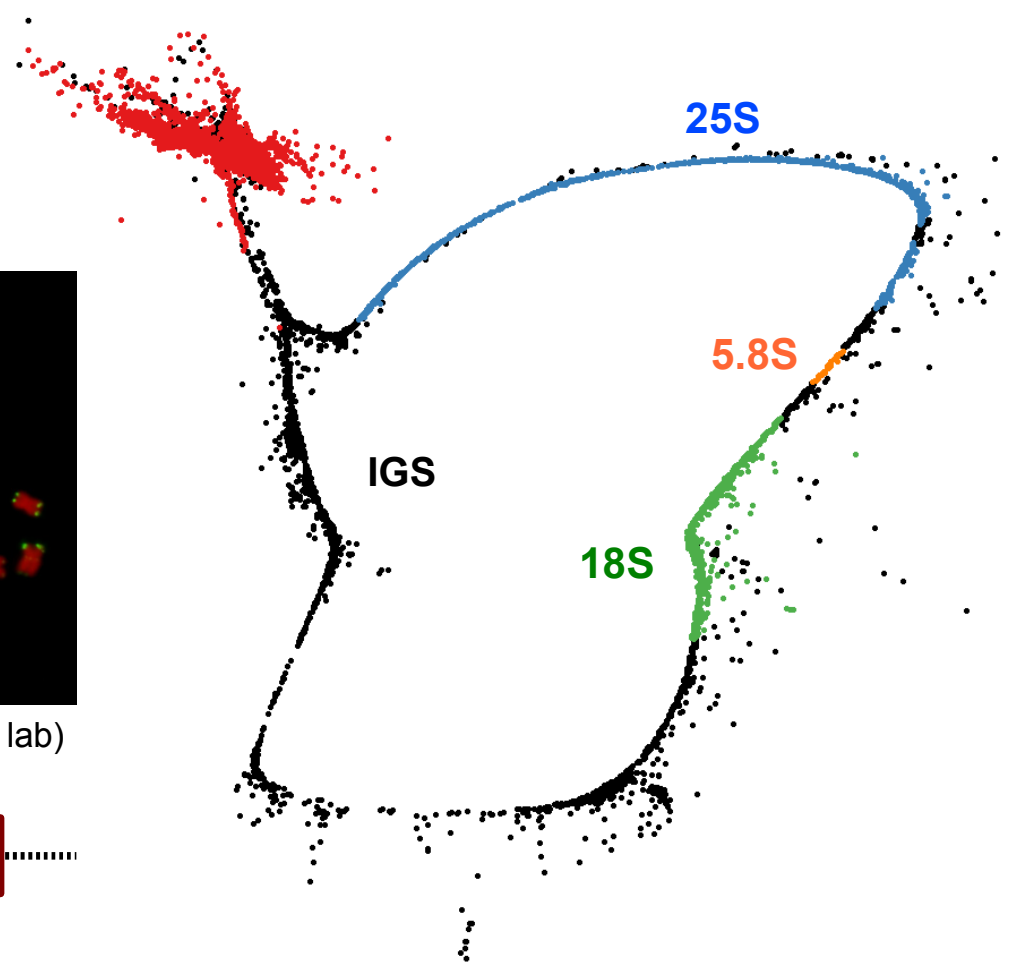
Reconstruction of rDNA gene cluster in *Silene latifolia*

TANDEM REPEAT

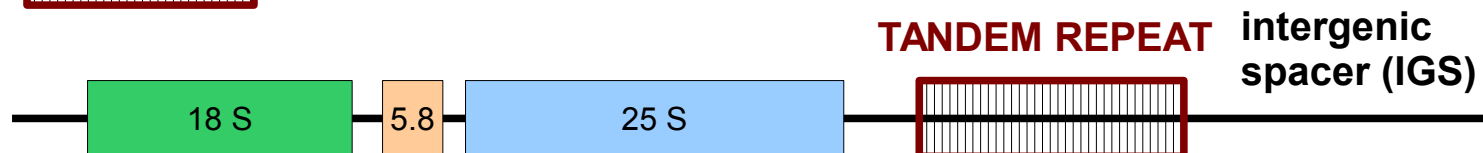
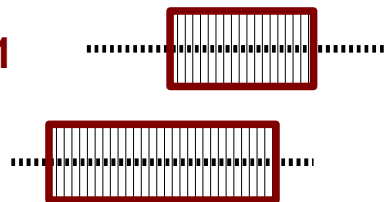
Satellite repeat X43.1
(Buzek et al. 1997)



(Vyskot lab)



Additional
loci of X43.1
occur
outside
rDNA IGS



Repeat quantification

	7192836 (= 100%)	64.1			
CL	reads	genome %	class	type	note
1	304159	4.229	gypsy	Tat	PROT-RT/RH-INT
2	234749	3.264			?, PE->24,18
3	216307	3.007	gypsy	chromo	ALL domains
4	202822	2.820	copla	Maximus	ALL domains
5	149693	2.081	gypsy	Athila	ALL domains
6	145911	2.029	gypsy	Tat	ALL domains
7	143766	1.999	gypsy	chromo	ALL domains
8	142608	1.983	copla	Maximus	ALL domains
9	141836	1.972	LINE		RT
10	123886	1.722	gypsy	chromo	GAG
11	79345	1.103			?,PE->21,95
12	72781	1.012	copla	Angela	ALL domains
13	67096	0.933	gypsy	Tat	ALL domains
14	65455	0.910	gypsy	Athila	GAG, PE->1(!!!),36
15	62334	0.867	gypsy	Tat	ALL domains
16	53845	0.749	copla	Ivana/Oryco	ALL domains
17	49341	0.686			? DNA transp ?? + TR
18	45062	0.626			?,PE->2,63
19	44762	0.622			?,PE->28
20	43332	0.602	tandem		monom ?? ~400 (~1200)
21	42344	0.589	gypsy	chromo	ALL domains
22	40125	0.558	gypsy	Tat	PE->15
23	39923	0.555			?,PE->7,73
24	36353	0.505	gypsy	chromo	(GAG), PE->2,3,28
25	35977	0.500			?,PE->2,81
26	35674	0.496			?
27	34829	0.484	rDNA	5S	
28	34534	0.480	gypsy	chromo	PE->29,19,24
29	34302	0.477	gypsy	chromo	PE->28
30	33114	0.460			?
31	32930	0.458			?
32	32122	0.451	LINE		RT

Based on numbers of reads
in individual clusters

(more info during practical training)

Using paired-end read
information for improving
cluster annotation