

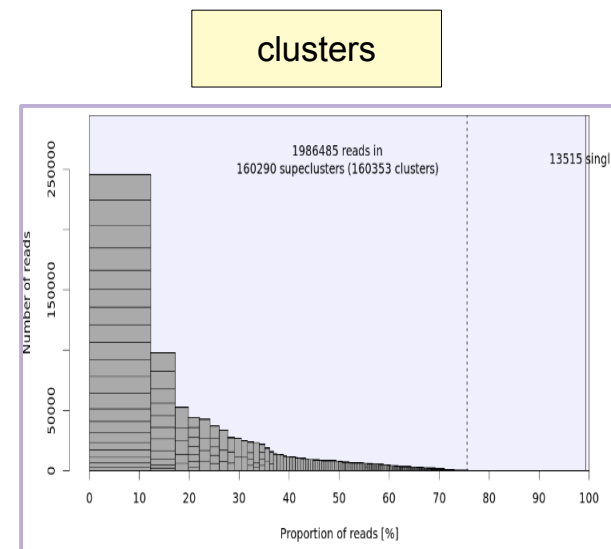
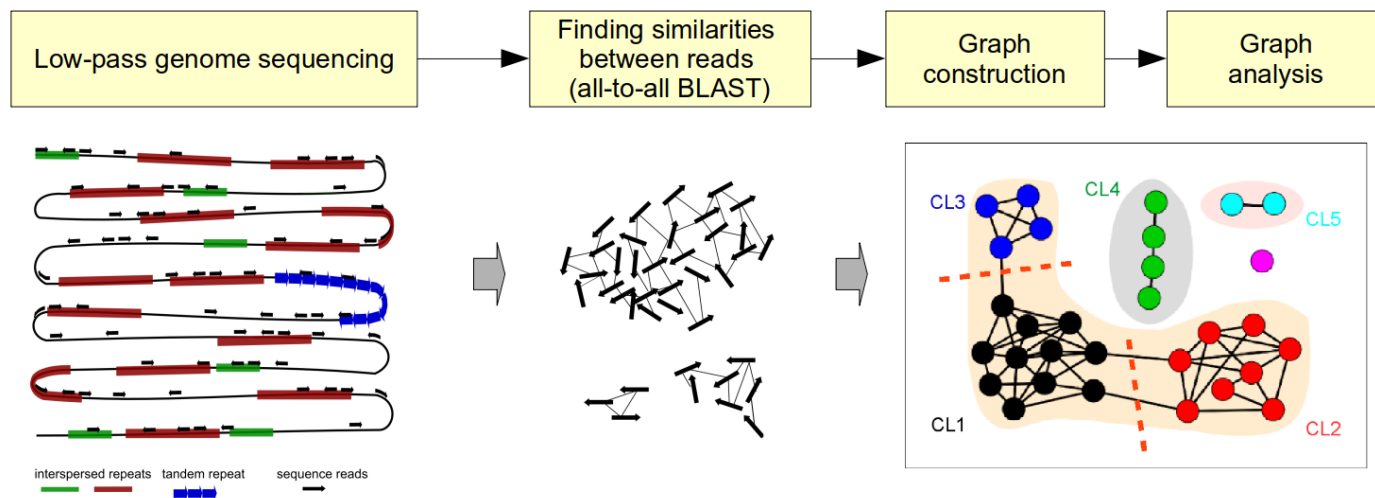
# RepeatExplorer applications

---

# RepeatExplorer applications

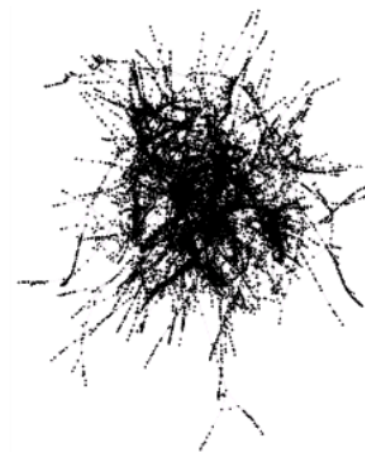
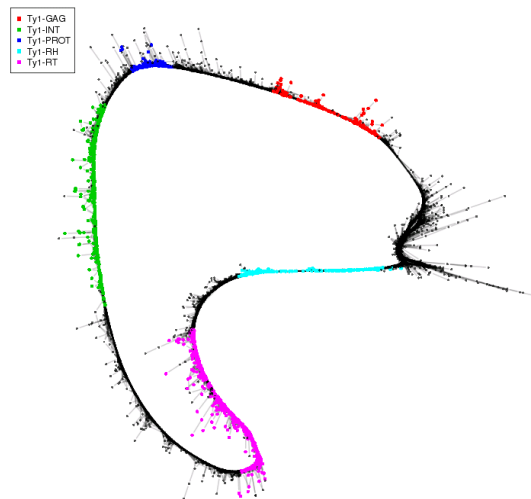
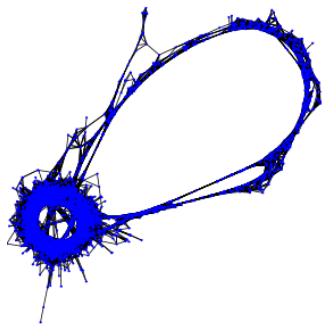
## What RepeatExplorer can do for you:

- Identify all repeated sequences with certain number of copies
- Repeat quantification



# RepeatExplorer applications

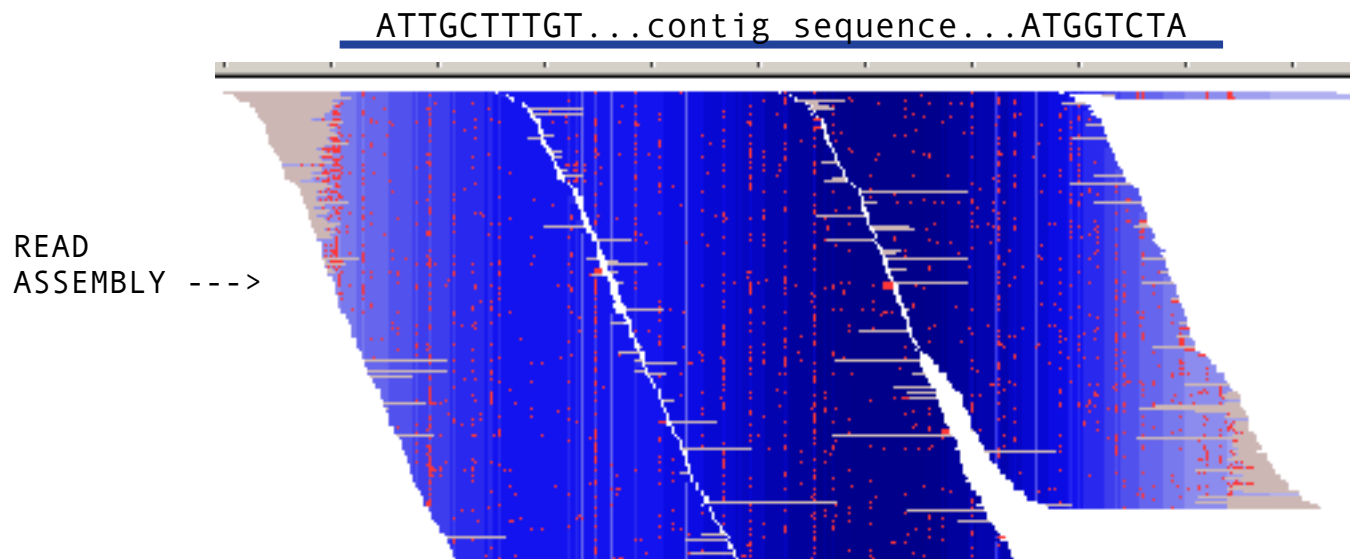
- What RepeatExplorer can do for you:
  - Identify all repeated sequences with certain number of copies
  - Repeat quantification
  - Provide models of repeat populations (sequence variability)



# RepeatExplorer applications

## What RepeatExplorer can do for you:

- Identify all repeated sequences with certain number of copies
- Repeat quantification
- Provide models of repeat populations (sequence variability, consensus sequences)



# RepeatExplorer applications

## What RepeatExplorer can do for you:

- Identify all repeated sequences with certain number of copies
- Repeat quantification
- Provide models of repeat populations (sequence variability, consensus sequences)
- Help in repeat classification/annotation

	Proportion[%]	Nsuperclusters	Nclusters	Nreads
Unclassified_repeat (conflicting evidences)	0	0	0	0
--rDNA	0	0	0	0
--45S_rDNA	2.22	1	7	44414
--18S_rDNA	1.2	1	4	23926
--25S_rDNA	0	0	0	0
--5.8S_rDNA	0	0	0	0
--5S_rDNA	0.07	1	1	1320
--satellite	0.47	4	4	9320
--mobile element	0	0	0	0
--Class_I	0	0	0	0
--SINE	0	0	0	0
--LTR	4.91	17	20	98226
--Ty1_copia	1.87	7	7	37461
--Ale	0	0	0	0
--Alesia	0	0	0	0
--Angela	0.23	2	2	4606
--Bianca	0.02	1	1	447
--Bryco	0	0	0	0
--Lyco	0	0	0	0
--Gymco-III	0	0	0	0
--Gymco-I	0	0	0	0
--Gymco-II	0	0	0	0
--Ikeros	0.17	1	1	3374
--Ivana	0.38	3	3	7674
--Gymco-IV	0	0	0	0
--Osser	0	0	0	0
--SIRE	1.51	13	13	30151
--TAR	0.11	2	2	2196
--Tork	0	0	0	0
--Ty1-outgroup	0	0	0	0
--Ty3_gypsy	0	0	0	0
--non-chromovirus	0	0	0	0
--non-chromo-outgroup	0	0	0	0
--Phygy	0	0	0	0
--Seigy	0	0	0	0
--OTA	0	0	0	0
--Athila	2.18	12	12	43690
--Tat	0	0	0	0

# RepeatExplorer applications

---

## What RepeatExplorer can do for you:

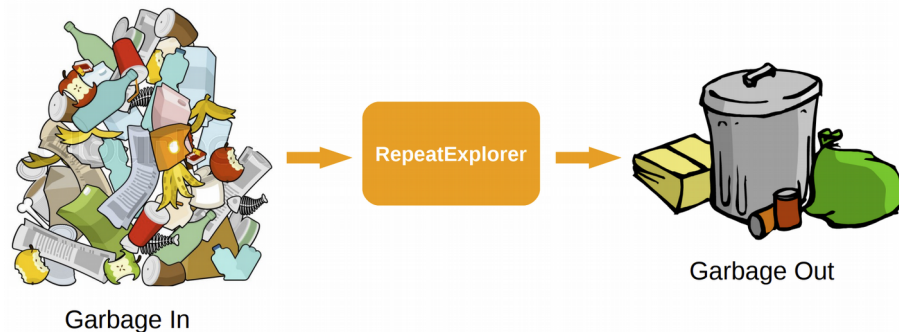
- Identify all repeated sequences with certain number of copies
- Repeat quantification
- Provide models of repeat populations (sequence variability, consensus sequences)
- Help in repeat classification/annotation

## What it cannot do:

- Genome assembly or reconstruction of individual repeat copies
- Use assembled genomes as input
- Analyze RNA / transcriptomic data
- Use long NGS reads (PacBio, Oxford Nanopore) as input (*but work in progress*)

# RepeatExplorer applications

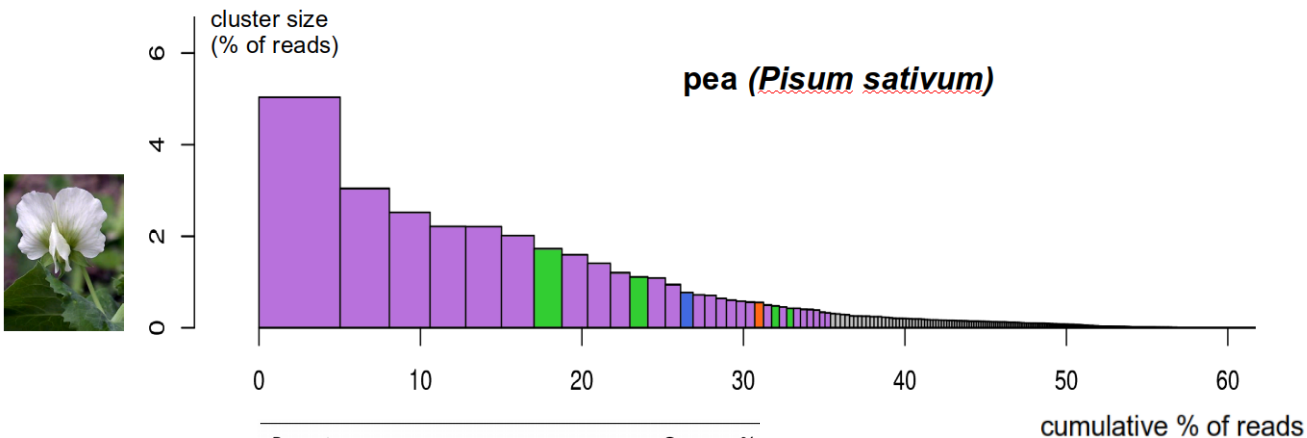
---



## What it cannot do:

- Genome assembly or reconstruction of individual repeat copies
- Use assembled genomes as input
- Analyze RNA / transcriptomic data
- Use long NGS reads (PacBio, Oxford Nanopore) as input (*but work in progress*)

# Repeat analysis in a single species



Repeat type:

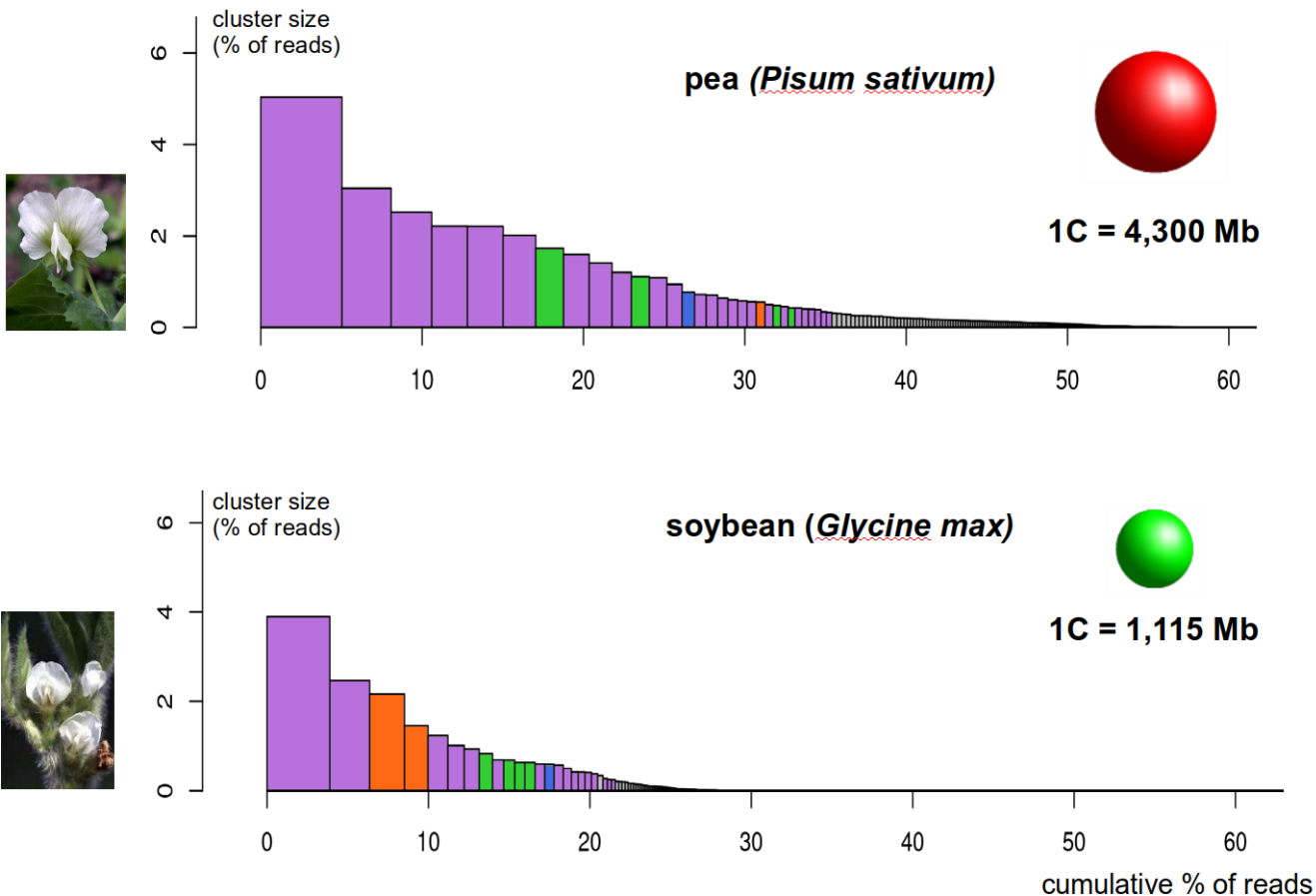
Ty3/gypsy  
Ty1/copia  
Satellite DNA  
rDNA  
other/unknown

Repeats	Genome %
<b>LTR/gypsy</b>	<b>55.61</b>
TatIV_Ogre	46.68
Athila	5.42
Tekay	3.44
TatV	0.05
CRM	0.02
<b>LTR/copia</b>	<b>11.72</b>
SIRE	7.43
Angela	2.45
Ivana	1.03
TAR	0.22
Tork	0.15
Ale	0.11
Ikeros	0.04
LTR/TRIM	0.29
<b>DNA transposon</b>	<b>1.94</b>

.....



# Repeat analysis in a single species



Repeat type:

Ty3/gypsy

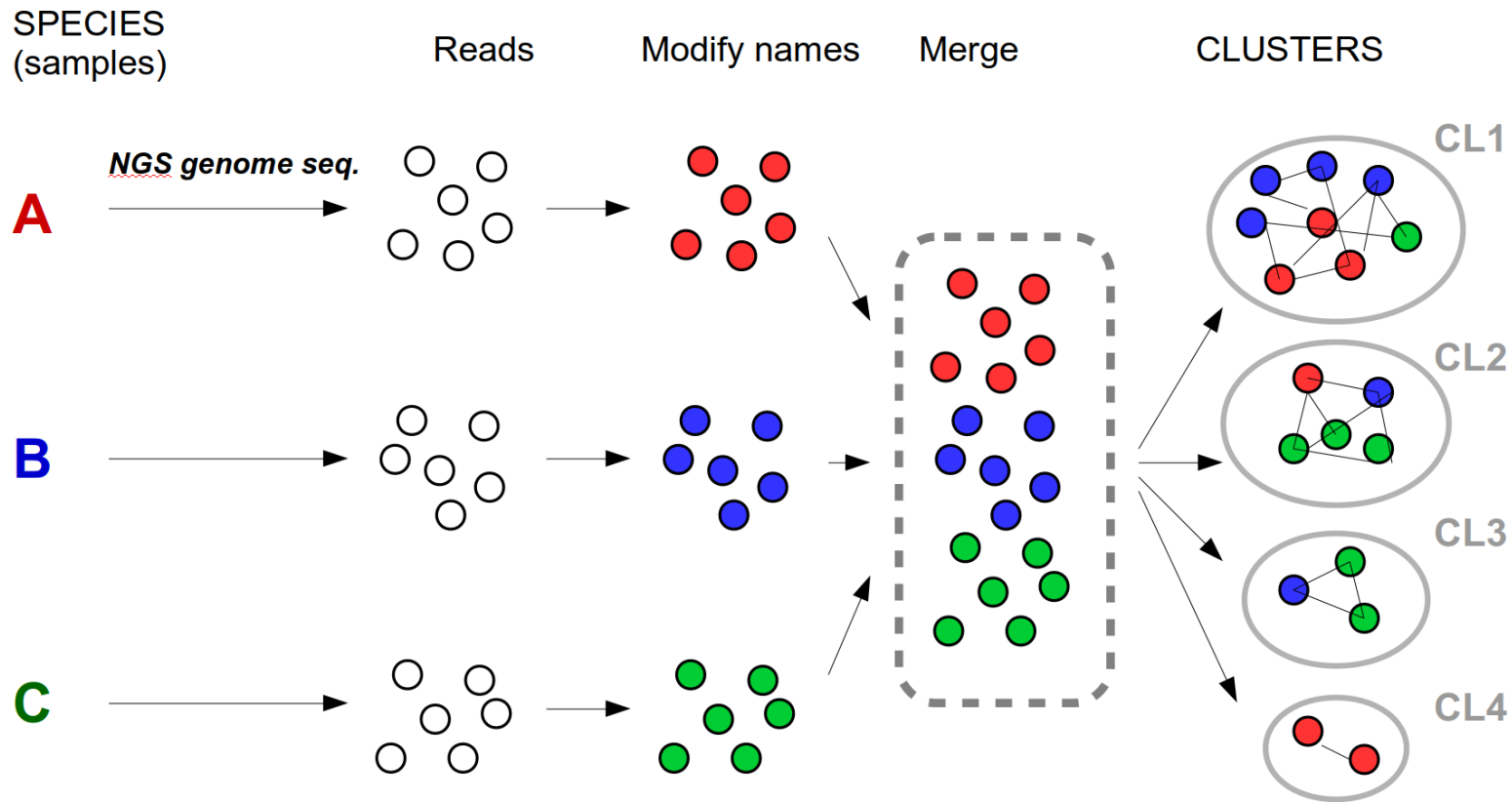
Ty1/copia

Satellite DNA

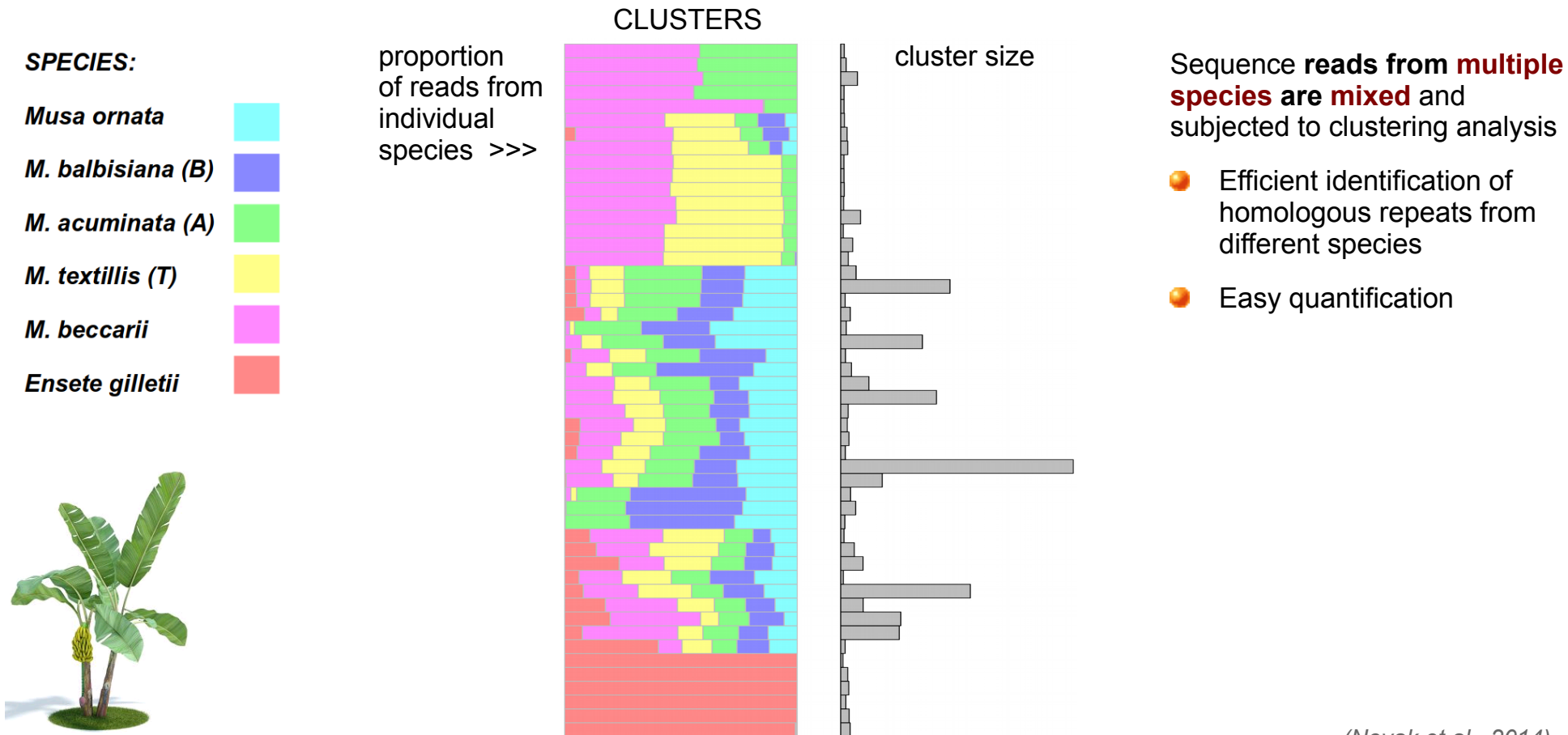
rDNA

other/unknown

# Comparative analysis

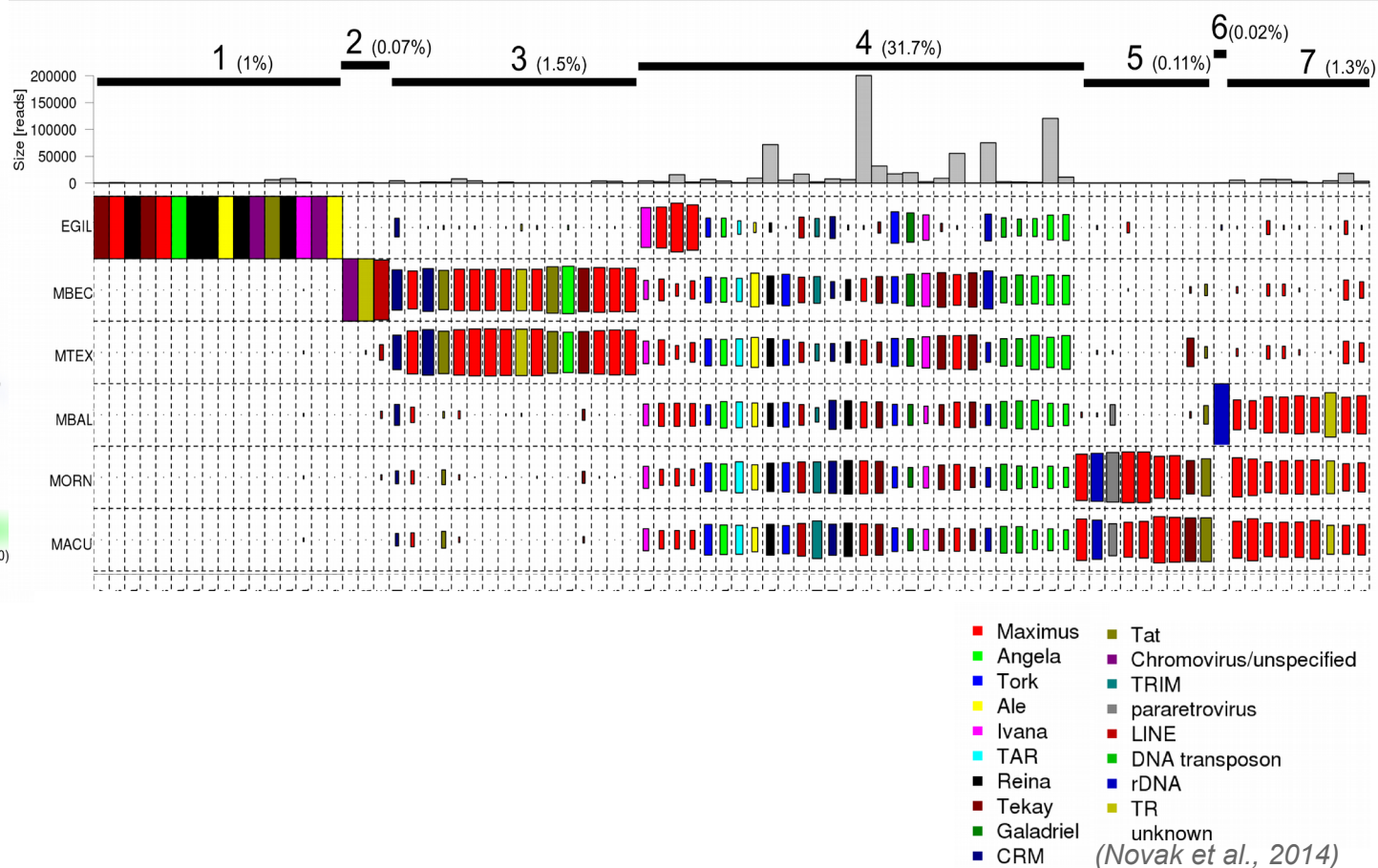
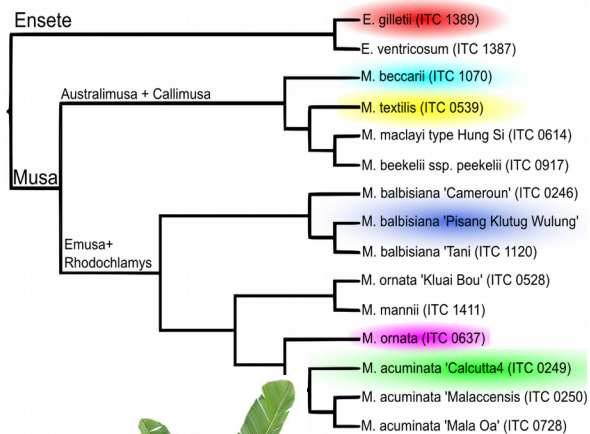


# Comparative analysis

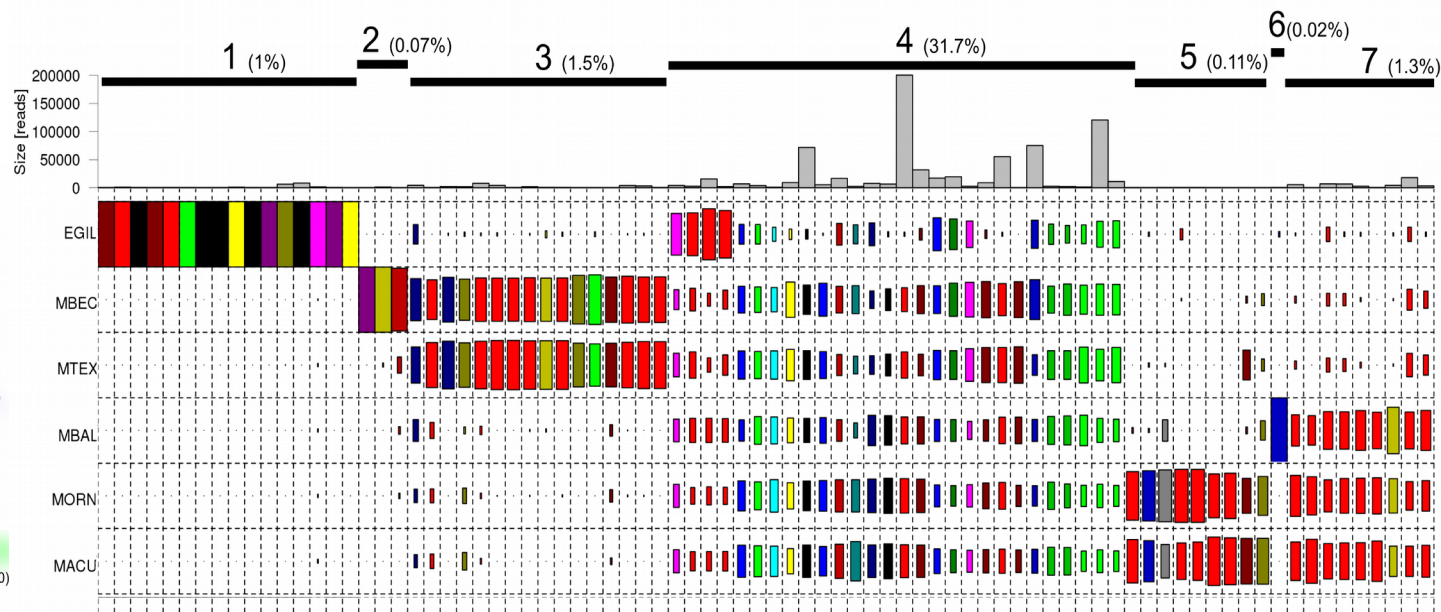
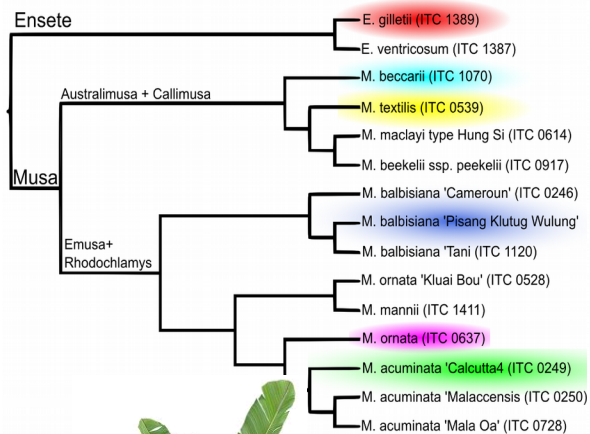


(Novak et al., 2014)

# Comparative analysis



# Comparative analysis



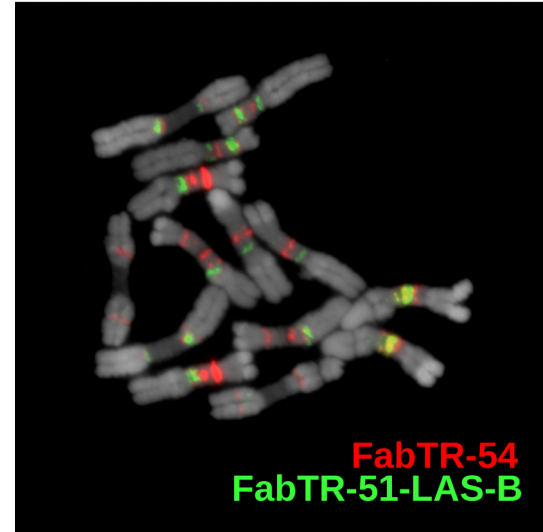
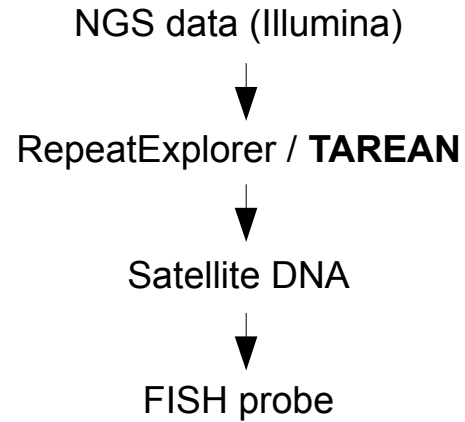
Genomic repeat abundances  
contain phylogenetic signal

Dodsworth et al. (2015)  
*Syst. Biol.* 64(1): 112-126

■ Maximus ■ Tat  
 ■ Angela ■ Chromovirus/unspecified  
 ■ Tork ■ TRIM  
 ■ Ale ■ pararetrovirus  
 ■ Ivana ■ LINE  
 ■ TAR ■ DNA transposon  
 ■ Reina ■ rDNA  
 ■ Tekay ■ TR  
 ■ Galadriel ■ unknown  
 ■ CRM (Novak et al., 2014)

# Development of repeat-based markers

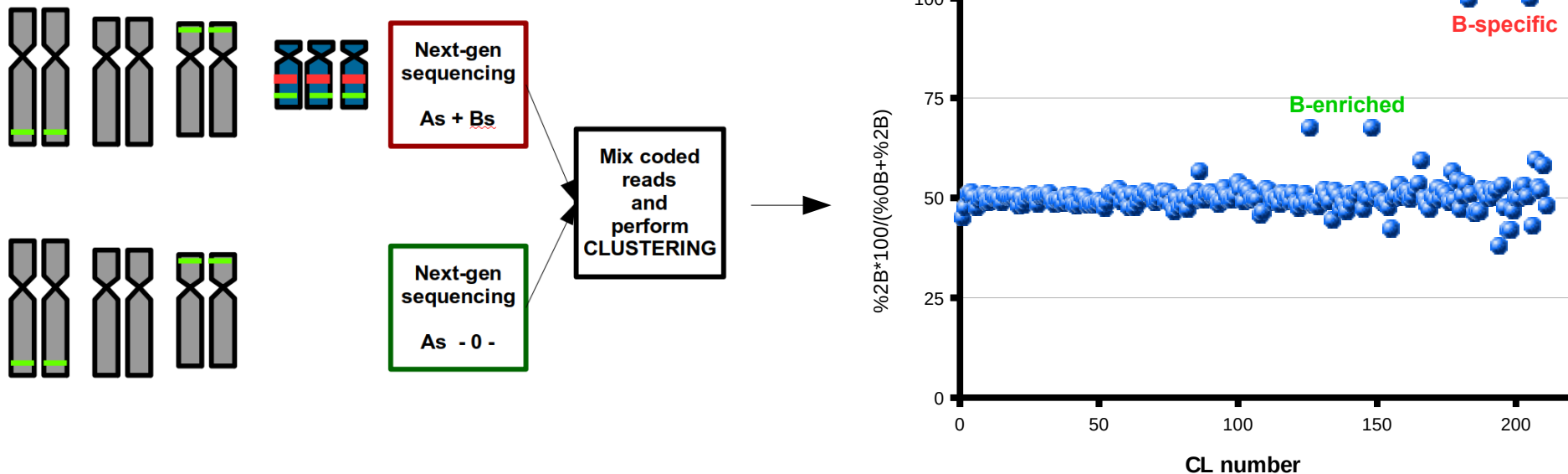
## Cytogenetic markers (FISH probes)



*Lathyrus sativus*  
(FISH by L. Avila Robledillo)

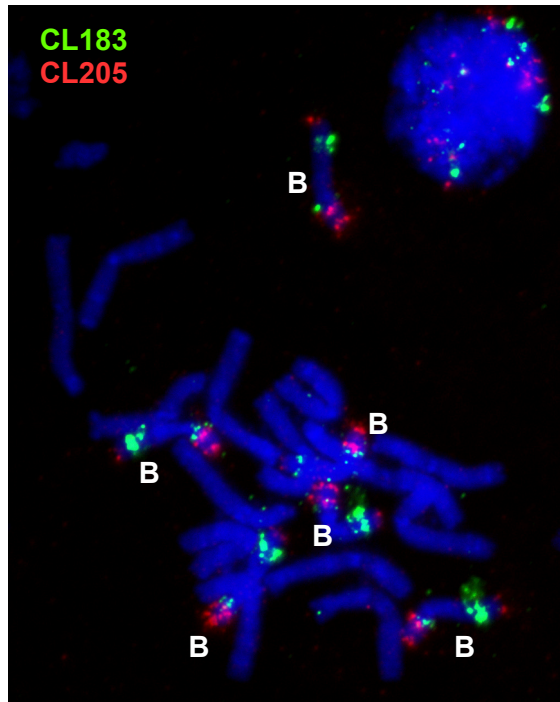
# Identification of chromosome B-specific repeats

## Comparative analysis of B+/- plants



# Identification of chromosome B-specific repeats

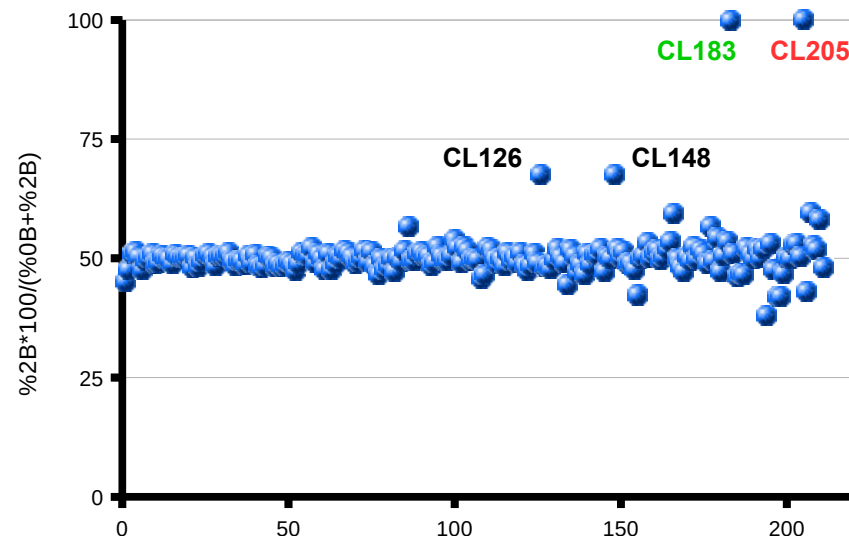
## Comparative analysis of B+/- plants



*FISH by Alevtina Ruban*

*Wu et al. (2019) New Phytol.*

### Detection of clusters enriched on Bs



### Repeats significantly enriched on Bs

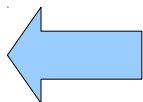
CL	reads	[%]	0B	%	2B	%	$\%2B*100/(\%0B+\%2B)$	Repeat type
126	5216	0.191	1688	0.120	3528	0.252	68	satellite (86 bp)
148	2877	0.105	934	0.067	1943	0.139	68	tandem (?)
183	731	0.027	1	0.000	730	0.052	100	satellite (~1.1 kb)
205	358	0.013	0	0.000	358	0.026	100	satellite (185 bp)



# Cluster-centered downstream applications

## Cluster annotation and 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(III),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			?



## CLUSTERS

- Represent specific repeat families/variants or their parts
- Can be used as a reference for similarity-based mapping of various sequencing data

## Using clusters as reference for:

- RNA-seq reads (mRNA, smRNAs,...)
  - Detection of transcribed repeats
  - Comparative analysis (tissues,...)
- Assessing completeness of genome assemblies
  - Mapping simulated reads from the assembly
- ChIP-seq reads
  - Association of repeats with specific types of chromatin

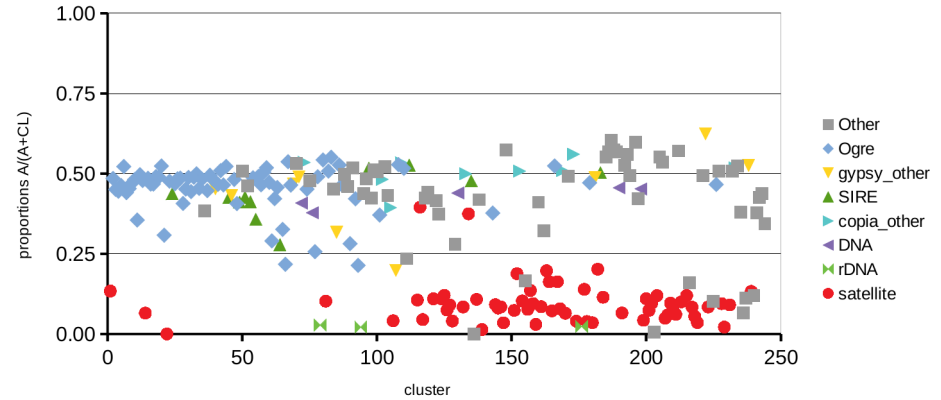
# Assessing completeness of genome assemblies

Illumina  
WGS  
reads  
(cluster  
size)

Mapped  
fragments  
from  
assembly

comparison

			sum of hits		blasted fragments	CL/(CL+BLAST)
			133232913	207900483		
Size_adjusted	Proportion [%]	Final_annotation	CL	hits		
85804	4.187	All/repeat/satellite/FabTR_FokI	0	6171084	2.968	0.13
38954	1.901	All/repeat/mobile_element/Class_I	1	1343009	0.646	0.49
37146	1.813	All/repeat/mobile_element/Class_I	2	3723833	1.791	0.45
37051	1.808	All/repeat/mobile_element/Class_I	3	3091188	1.487	0.45
36830	1.797	All/repeat/mobile_element/Class_I	4	3014725	1.450	0.46
33547	1.637	All/repeat/mobile_element/Class_I	5	3221042	1.549	0.52
32220	1.572	All/repeat/mobile_element/Class_I	6	3719091	1.789	0.44
31149	1.520	All/repeat/mobile_element/Class_I	7	2574963	1.239	0.45
30972	1.511	All/repeat/mobile_element/Class_I	8	2619726	1.260	0.47
30361	1.482	All/repeat/mobile_element/Class_I	9	2791506	1.343	0.48
29548	1.442	All/repeat/mobile_element/Class_I	10	2857979	1.375	0.35
29227	1.426	All/repeat/mobile_element/Class_I	11	1649483	0.793	0.50
27341	1.334	All/repeat/mobile_element/Class_I	12	2924386	1.407	0.48
26964	1.316	All/repeat/mobile_element/Class_I	13	2544864	1.224	0.06
26082	1.273	All/repeat/satellite/FabTR_53	14	189991	0.091	0.49
25651	1.252	All/repeat/mobile_element/Class_I	15	2496449	1.201	0.47
25147	1.227	All/repeat/mobile_element/Class_I	16	2280990	1.097	0.47
23553	1.149	All/repeat/mobile_element/Class_I	17	2239097	1.077	0.49
22914	1.118	All/repeat/mobile_element/Class_I	18	2304645	1.109	0.49
22829	1.114	All/repeat/mobile_element/Class_I	19	2227830	1.072	0.52
22696	1.108	All/repeat/mobile_element/Class_I	20	2547541	1.225	0.31
22481	1.097	All/repeat/mobile_element/Class_I	21	1024482	0.493	0.00
21364	1.043	All/repeat/satellite/FabTR_83	22	6	0.000	0.48
20442	0.998	All/repeat/mobile_element/Class_I	23	1971842	0.948	0.44
19605	0.957	All/repeat/mobile_element/Class_I	24	1610084	0.774	0.47
19289	0.941	All/repeat/mobile_element/Class_I	25	1753755	0.844	0.48
19234	0.939	All/repeat/mobile_element/Class_I	26	1835955	0.883	0.49
18932	0.924	All/repeat/mobile_element/Class_I	27	1842540	0.886	0.41
18345	0.895	All/repeat/mobile_element/Class_I	28	1318285	0.634	0.45
18112	0.884	All/repeat/mobile_element/Class_I	29	1533938	0.738	0.49
17991	0.878	All/repeat/mobile_element/Class_I	30	1753161	0.843	0.44
17167	0.838	All/repeat/mobile_element/Class_I	31	1457392	0.701	0.49
16726	0.816	All/repeat/mobile_element/Class_I	32	1663803	0.800	0.50



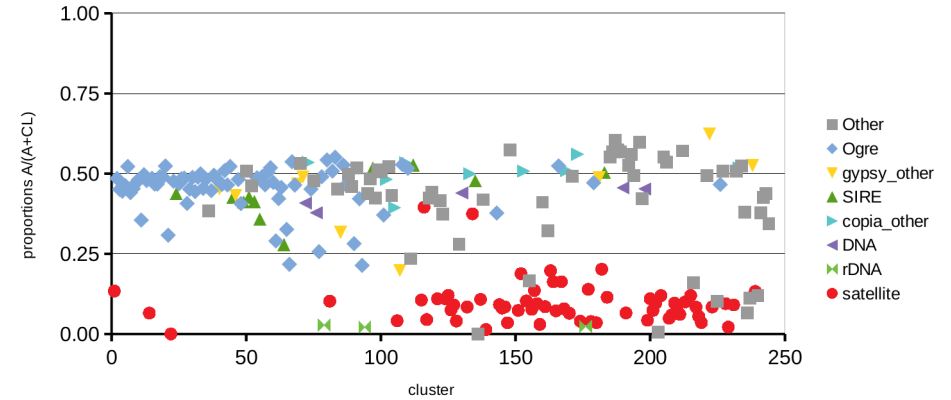
# Assessing completeness of genome assemblies

Illumina  
WGS  
reads  
(cluster  
size)

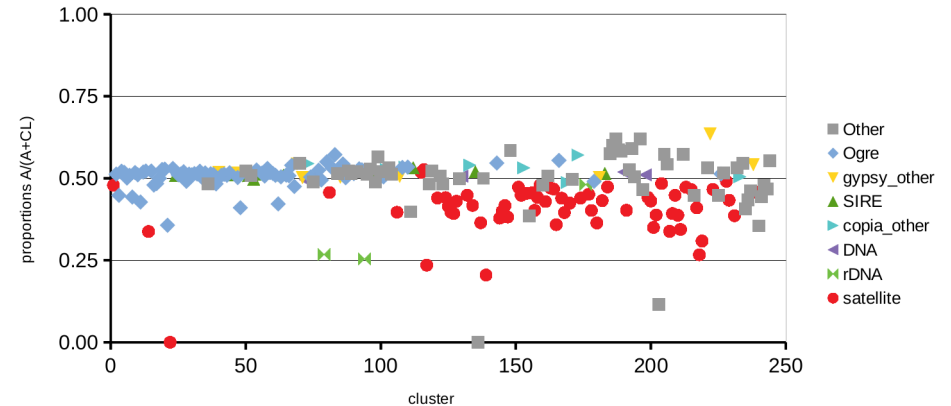
Mapped  
fragments  
from  
assembly

				sum of hits	blasted fragments	CL/(CL+BLAST)
				133232913	207900483	
Size_adjusted	Proportion [%]	Final_annotation	CL	hits		
85804	4.187	All/repeat/satellite/FabTR_FokI	0	6171084	2.968	
38954	1.901	All/repeat/mobile_element/Class_I	1	1343009	0.646	0.13
37146	1.813	All/repeat/mobile_element/Class_I	2	3723833	1.791	0.49
37051	1.808	All/repeat/mobile_element/Class_I	3	3091188	1.487	0.45
36830	1.797	All/repeat/mobile_element/Class_I	4	3014725	1.450	0.45
33547	1.637	All/repeat/mobile_element/Class_I	5	3221042	1.549	0.46
32220	1.572	All/repeat/mobile_element/Class_I	6	3719091	1.789	0.52
31149	1.520	All/repeat/mobile_element/Class_I	7	2574963	1.239	0.44
30972	1.511	All/repeat/mobile_element/Class_I	8	2619726	1.260	0.45
30361	1.482	All/repeat/mobile_element/Class_I	9	2791506	1.343	0.47
29548	1.442	All/repeat/mobile_element/Class_I	10	2857979	1.375	0.48
29227	1.426	All/repeat/mobile_element/Class_I	11	1649483	0.793	0.35
27341	1.334	All/repeat/mobile_element/Class_I	12	2924386	1.407	0.50
26964	1.316	All/repeat/mobile_element/Class_I	13	2544864	1.224	0.48
26082	1.273	All/repeat/satellite/FabTR_53	14	189991	0.091	0.06
25651	1.252	All/repeat/mobile_element/Class_I	15	2496449	1.201	0.49
25147	1.227	All/repeat/mobile_element/Class_I	16	2280990	1.097	0.47
23553	1.149	All/repeat/mobile_element/Class_I	17	2239097	1.077	0.47
22914	1.118	All/repeat/mobile_element/Class_I	18	2304645	1.109	0.49
22829	1.114	All/repeat/mobile_element/Class_I	19	2227830	1.072	0.49
22696	1.108	All/repeat/mobile_element/Class_I	20	2547541	1.225	0.52
22481	1.097	All/repeat/mobile_element/Class_I	21	1024482	0.493	0.31
21364	1.043	All/repeat/satellite/FabTR_83	22	6	0.000	0.00
20442	0.998	All/repeat/mobile_element/Class_I	23	1971842	0.948	0.48
19605	0.957	All/repeat/mobile_element/Class_I	24	1610084	0.774	0.44
19289	0.941	All/repeat/mobile_element/Class_I	25	1753755	0.844	0.47
19234	0.939	All/repeat/mobile_element/Class_I	26	1835955	0.883	0.48
18932	0.924	All/repeat/mobile_element/Class_I	27	1842540	0.886	0.49
18345	0.895	All/repeat/mobile_element/Class_I	28	1318285	0.634	0.41
18112	0.884	All/repeat/mobile_element/Class_I	29	1533938	0.738	0.45
17991	0.878	All/repeat/mobile_element/Class_I	30	1753161	0.843	0.49
17167	0.838	All/repeat/mobile_element/Class_I	31	1457392	0.701	0.44
16726	0.816	All/repeat/mobile_element/Class_I	32	1663803	0.800	0.49
			33	1603448	0.816	0.50

Old assembly (March 2020)

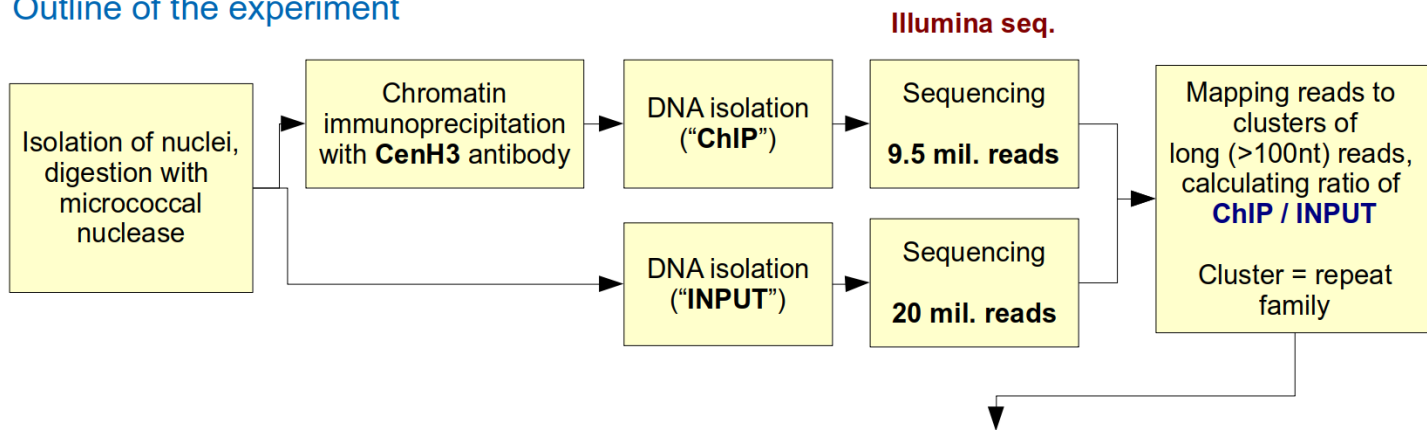


New assembly (Jan 2021)



# Identification of centromeric repeats by ChIP-seq

## Outline of the experiment



**ChIP-seq Mapper** >>>  
now available as a tool  
on our Galaxy server

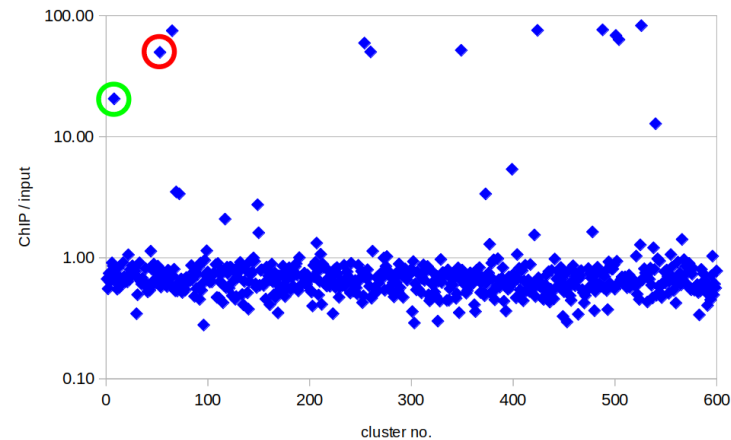
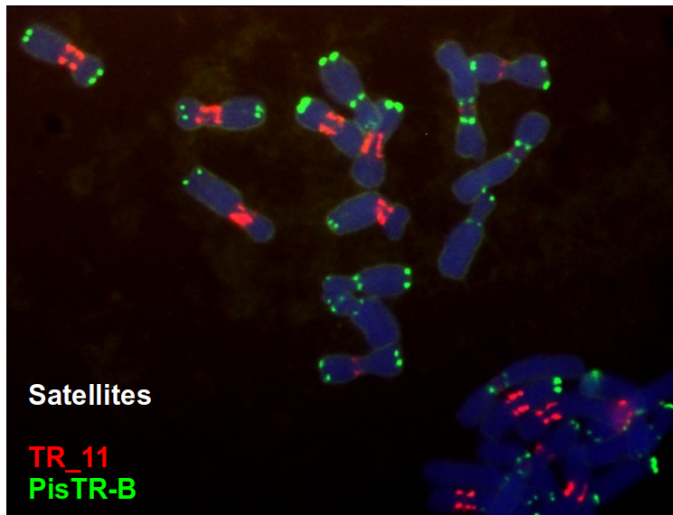
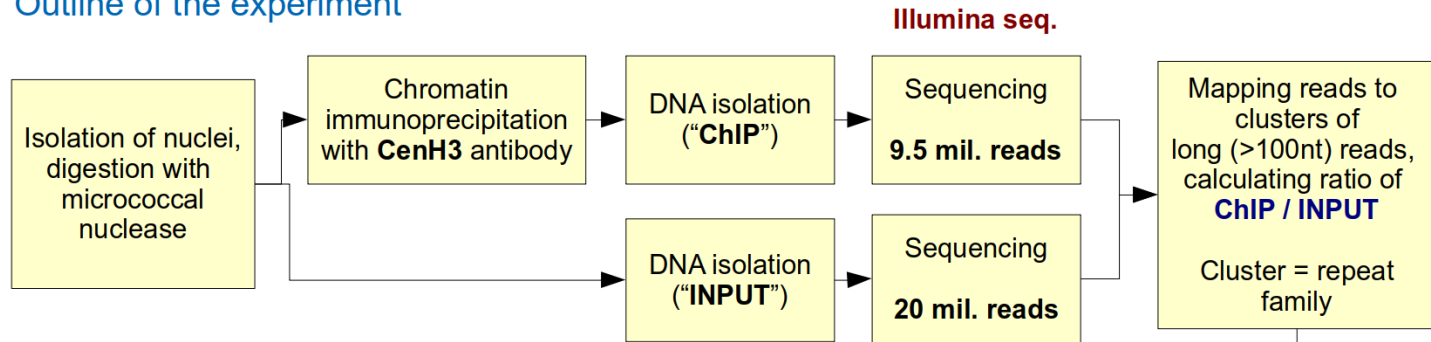
The screenshot shows the Galaxy web interface for the ChIP-seq Mapper tool (version 0.1.1). The tool is located under the 'Tools' section, specifically under 'Utilities' and 'ChIP-seq Mapper'. The main configuration panel includes the following fields:

- Chip Sequences:** 36: PFCC (NGS data in fasta format)
- Input Sequences:** 35: PFCI (NGS data in fasta format)
- Reference - Contig Sequences:** 34: PFL\_max contigs.info (Contigs obtained from RepeatExplorer clustering pipeline in fasta file)
- Number of clusters to be shown in graph:** 600
- Minimum bit score threshold:** 30 (Note: All similarity hits with lower bit score will not be considered for ChIP/Input ratio calculation)

The 'Execute' button is visible at the bottom of the configuration panel.

# Identification of centromeric repeats by ChIP-seq

## Outline of the experiment



Neumann et al. (2012) *PLoS Genetics* 8:e1002777

# RepeatExplorer applications...

---

... in our practical training:

- Repeat analysis in a single species
- Comparative analysis
- Satellite DNA analysis and design of FISH probes
- ChIP-seq analysis