

5th Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis in Plants

Welcome !



*Institute of Plant Molecular Biology,
Biology Centre ASCR
České Budějovice, Czech Rep.*



Programme

Tuesday (May 24)

9:30 - Principles of repeat identification in plant genomes and introduction to graph-based clustering (J. Macas)

10:10 - RepeatExplorer pipeline - algorithms and implementation (P. Novák)

10:30 - 11:00 *Coffee break*

11:00 - Using RepeatExplorer output for repeat annotation and quantification (J. Macas)

11:20 - Transposon protein databases – recent updates to cover all Viridiplantae (P. Neumann)

11:40 - New tools – ChIP-seq and satellite DNA analysis (P. Novák)

12:00 - 13:30 *Lunch*

13:30 - (18:00) Practical training I

- introduction to Galaxy environment
- overview of RepeatExplorer tools
- pre-processing of NGS reads
- setting up clustering analysis – single species
- read pre-processing and clustering analysis from the command line
- comparative clustering of multiple samples

19:30 → : Dinner at “Masné krámy”

Wednesday (May 25)

9:00 - 10:30

- Ilia Leitch (Royal Botanic Gardens, Kew, UK) – **Genome diversity across land plants**
- Steven Dodsworth (Royal Botanic Gardens, Kew, UK) – **Repeat dynamics in *Nicotiana***
- Jamie McCann (University of Vienna, Austria) – **Repeat dynamics in *Melampodium***

10:30 - 11:00 *Coffee break*

- Tony Heitkam (TU Dresden, Germany) – **Repeat analysis in the *Camellia japonica* genome**
- Anja Kögler (TU Dresden, Germany) – **Exploring the origin of the Pillnitz *Camellia* (*Camellia japonica*)**
- Gerhard Menzel (TU Dresden, Germany) – **Repetitive DNA families of the saffron crocus (*Crocus sativus* L.)**

12:20 - 13:30 *Lunch*

13:30 - (18:00) Practical training II

- identification and phylogenetic analysis of retrotransposon protein domains
- cluster annotation and repeat composition of the genome
- re-clustering and cluster merging
- SeqGrapheR – visualization and annotation of the cluster graphs
- comparative clustering of multiple species – data interpretation
- repeat quantification (principles, sensitivity and reproducibility)

Thursday (May 26)

9:00 – 10:30

- A. Sousa (Univ. Munich) – **Genomic and in situ analyses of the repetitive DNA on the Y-chromosome of *Coccinia grandis* in comparison to the X and autosomes**
 - N. Waminal (Seoul Nat. University) – **A high-copy satellite DNA of *Panax ginseng* was derived from CACTA transposon and is a useful barcode for identifying individual chromosomes**
 - I. Kirov (Institut. Agricultural and Fisheries Res.) – **Towards a FISH-based karyotype of *Rosa***
 - R. van Velzen (Wageningen Univ.) – **Expansion of Gypsy-type LTR retrotransposons in the nodulating non-legume *Parasponia***
- 10:30 - 11:00 *Coffee break*
- V. Peska (Masaryk University, Brno) – **BAL31-NGS identification of *Allium* telomeres**
 - I. Vargás Jentsch (University of Cologne) – **Repetitive trouble: the repetitive landscape of the milkweed bug *Oncopeltus fasciatus***
 - H. Becher (Queen Mary Univ.) – **Transcribed repeats and Numts in a grasshopper with a giant genome**

12:30 - 13:30 *Lunch*

13:30 – (18:00) - Practical training III

- design of hybridization probes based on RE output
 - **NEW TOOL**: automated identification and characterization of satellite repeats
 - **NEW TOOL**: evaluation of ChIP-seq data
- advanced topics, troubleshooting

Acknowledgments

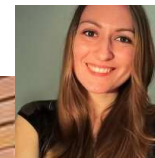


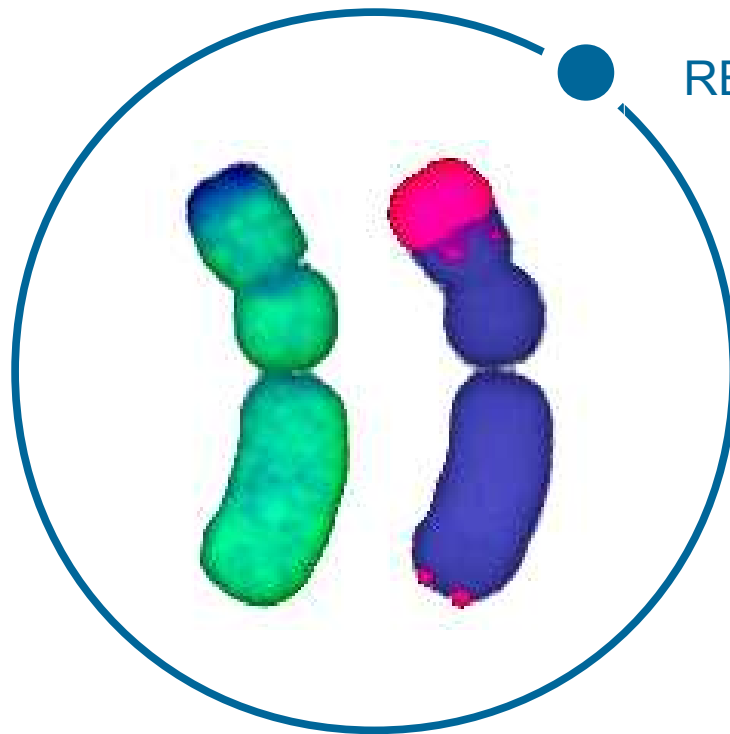
Biology Centre CAS, České Budějovice
Institute of Plant Molecular Biology

Laboratory of Molecular
Cytogenetics

Petr Novák
Pavel Neumann
Iva Fuková
Laura A. Robledillo
Nina Hošťáková

Renata Marecová
Andrea Koblížková
Vlasta Tetourová
Jana Látalová
Katharina Böttinger





REPEATS IN PLANT GENOMES

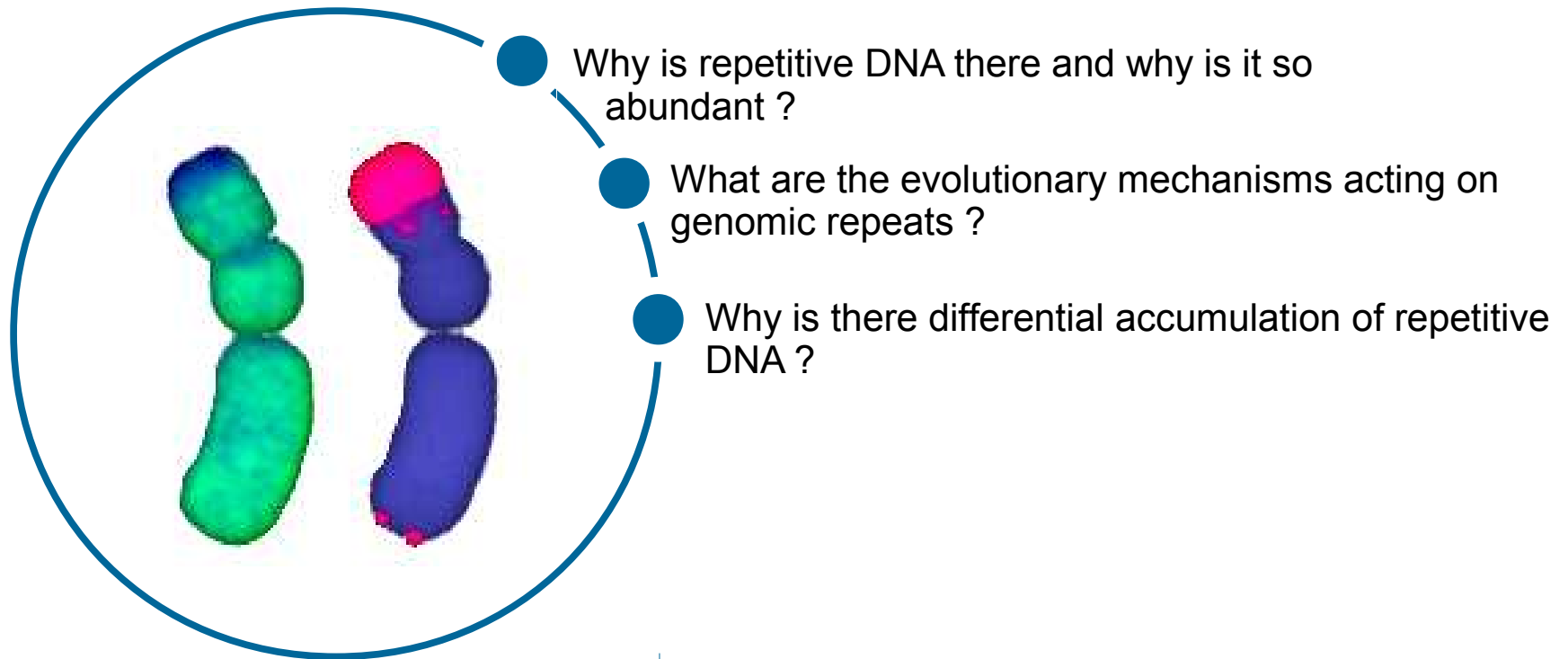


Satellite DNA

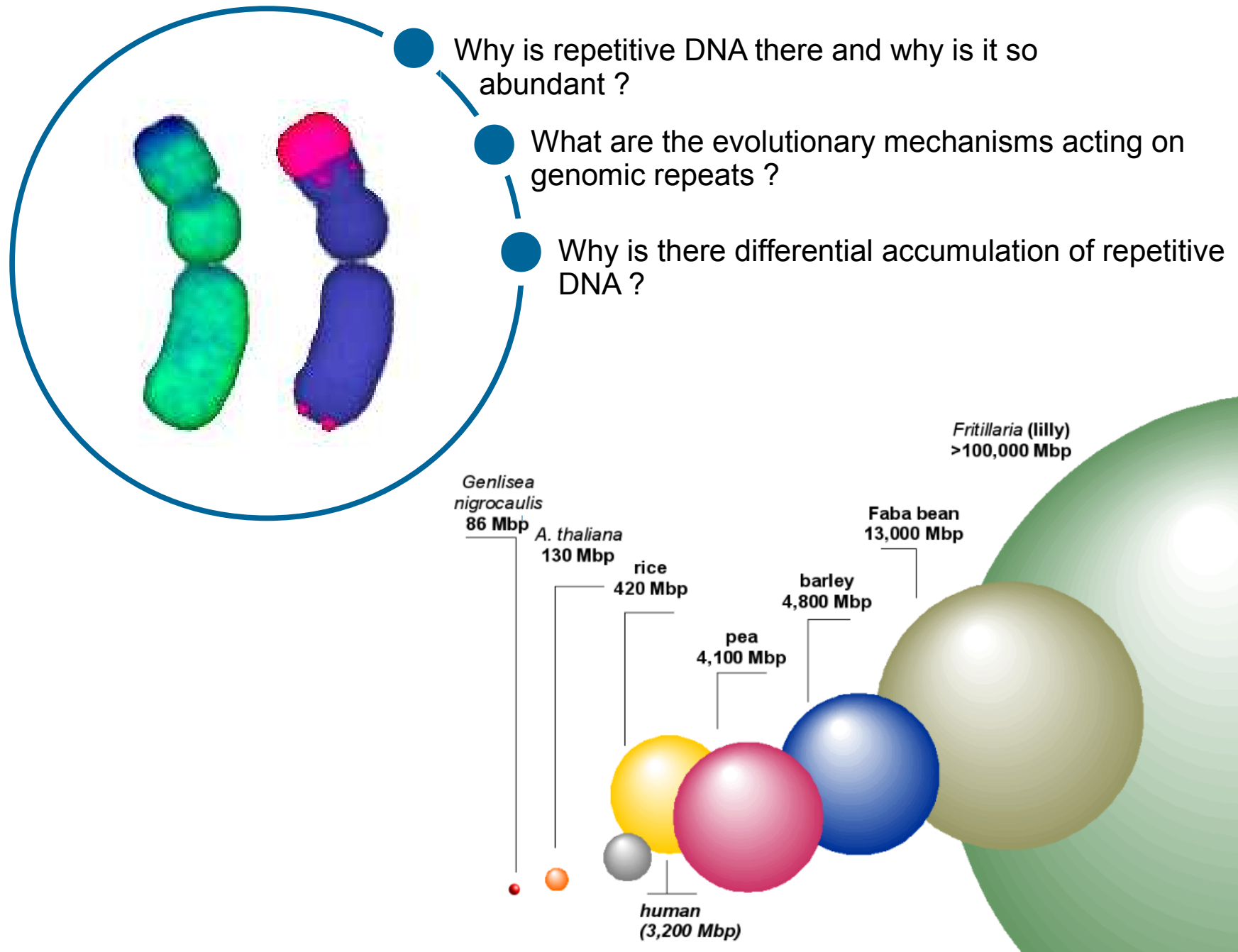


Retrotransposon

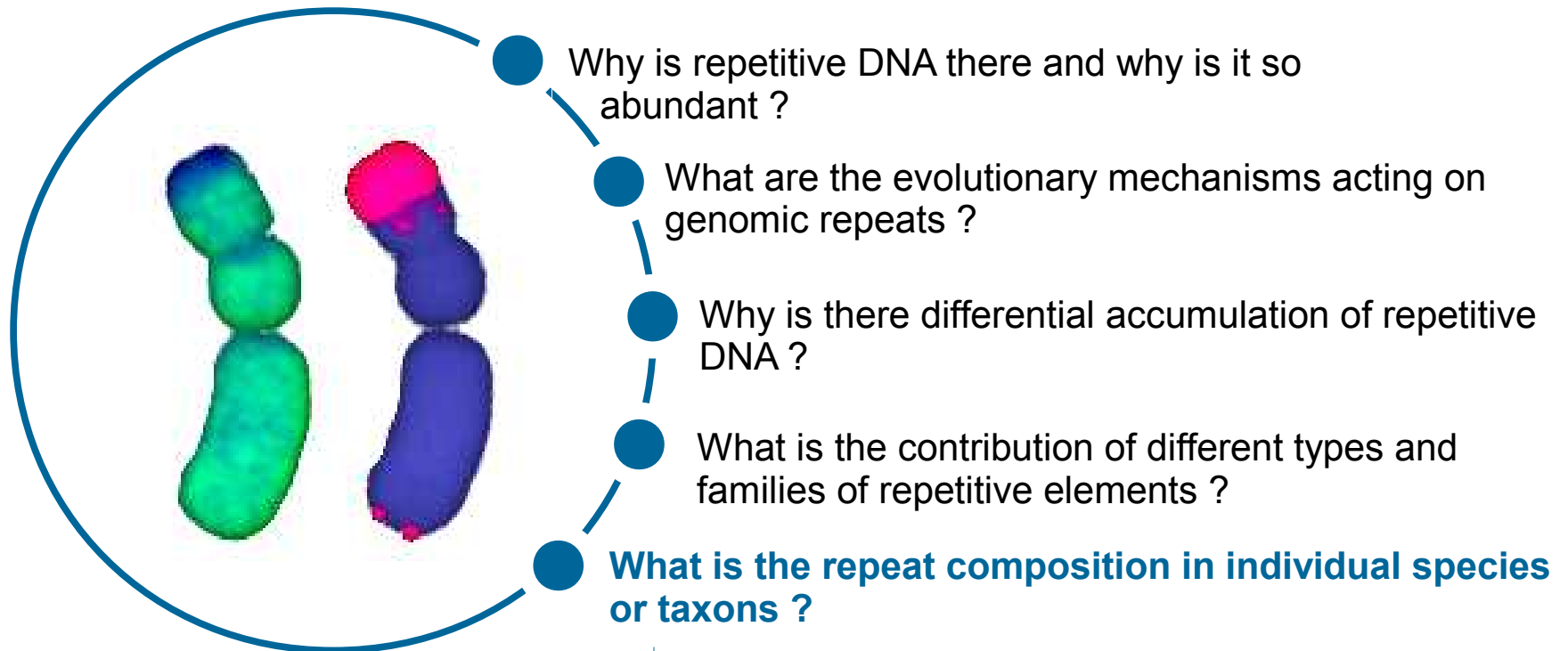
Repeats in plant genomes



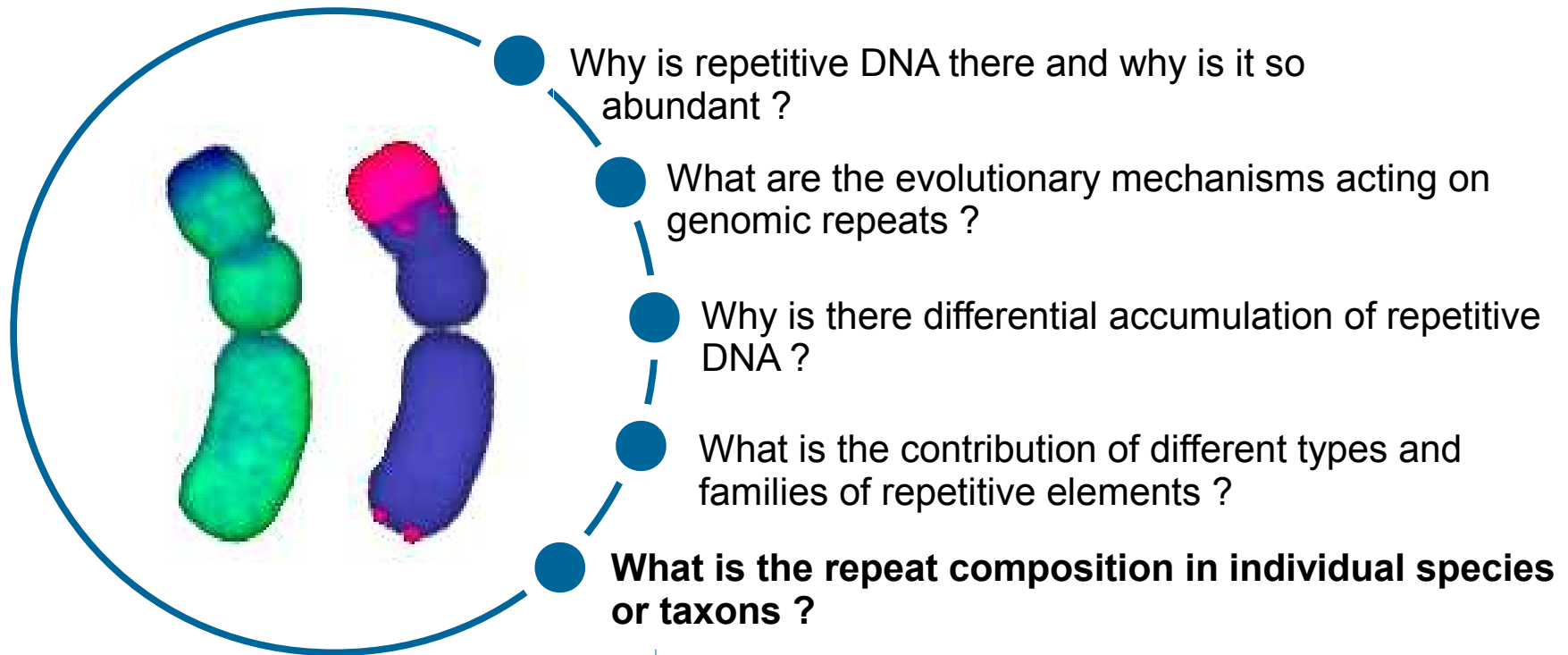
Repeats in plant genomes



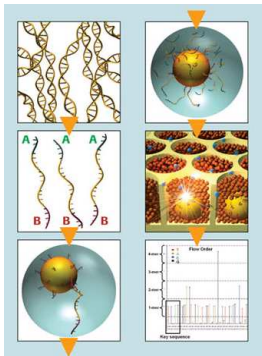
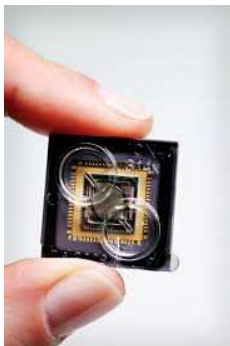
Repeats in plant genomes



The challenge of repeat identification in complex genomes



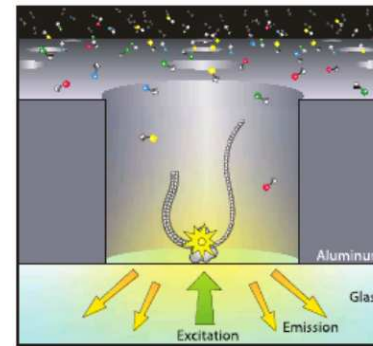
Next generation sequencing: getting sequence data is no longer a limiting factor



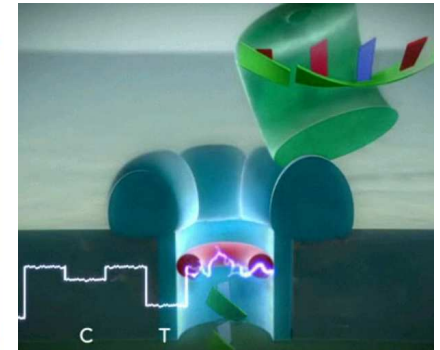
454/Roche



Illumina



Pacific Biosciences



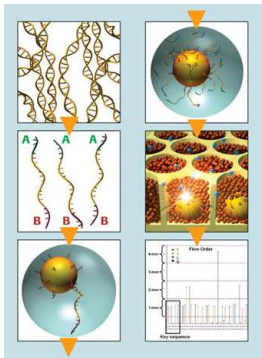
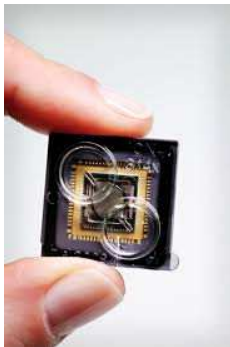
Oxford Nanopore

The challenge of repeat identification in complex genomes

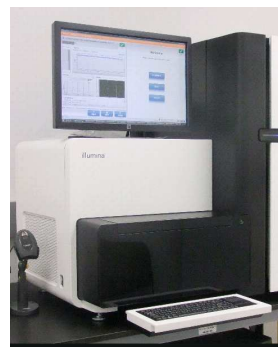
Repeat identification in sequence data

- Pure algorithmic (evaluates “repetitiveness” of substrings in DNA sequence)
- Signature or feature-based (searches for specific structures of biological importance)
- Similarity to known repeats

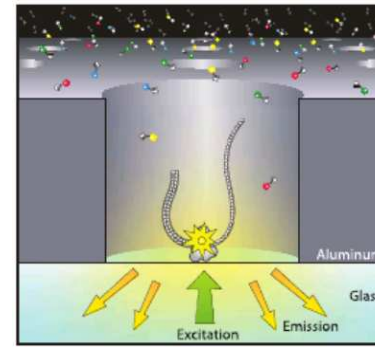
Next generation sequencing: getting sequence data is no longer a limiting factor



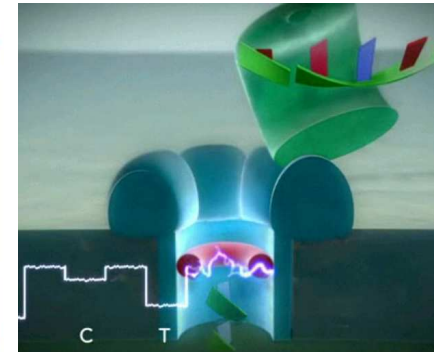
454/Roche



Illumina



Pacific Biosciences



Oxford Nanopore

Principles and tools for global repeat identification

		Assembled genome or long contigs (BACs etc.)	
		Y E S	N O
Reference database	Y E S		
	N O		

Repeat identification in sequence data

- Pure algorithmic (evaluates “repetitiveness” of substrings in DNA sequence)
- Signature or feature-based (searches for specific structures of biological importance)
- Similarity to known repeats

Principles and tools for global repeat identification

		Assembled genome or long contigs (BACs etc.)	
		Y E S	N O
Reference database	Y E S	• <i>first “model” species</i> • <i>extensive resources (+ wet lab)</i>	• <i>model-related taxa</i> • <i>often shotgun, “assembled”</i>
	N O		

Repeat identification in sequence data

- Pure algorithmic (evaluates “repetitiveness” of substrings in DNA sequence)
- Signature or feature-based (searches for specific structures of biological importance)
- Similarity to known repeats

Principles and tools for global repeat identification

		Assembled genome or long contigs (BACs etc.)	
		YES	NO
Reference database	YES	RepeatMasker / RepBase • <i>first “model” species</i> • <i>extensive resources (+ wet lab)</i>	• <i>model-related taxa</i> • <i>often shotgun, “assembled”</i>
	NO		

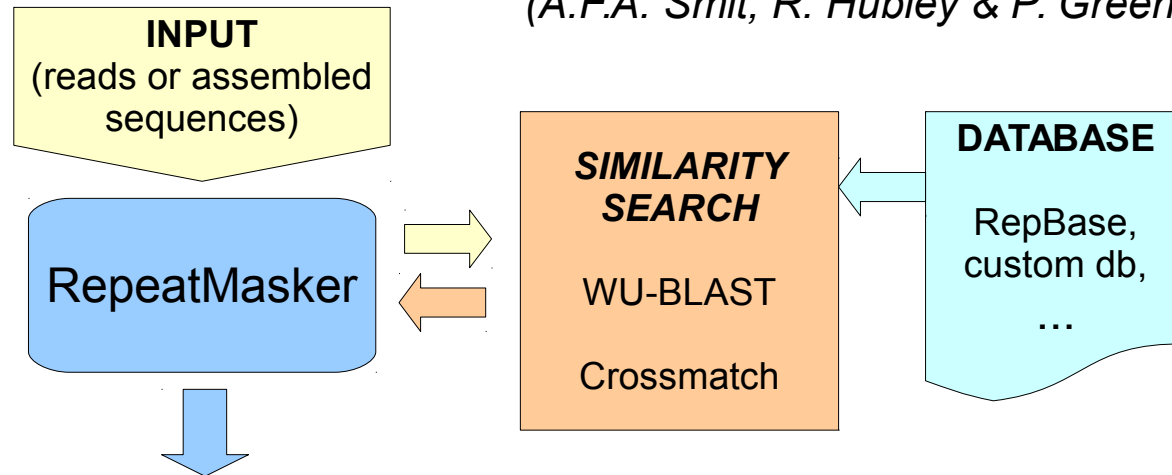
Repeat identification in sequence data

- Pure algorithmic (evaluates “repetitiveness” of substrings in DNA sequence)
- Signature or feature-based (searches for specific structures of biological importance)
- Similarity to known repeats (uses curated databases of known repeats)

Principles and tools for global repeat identification

RepeatMasker

(A.F.A. Smit, R. Hubley & P. Green)



OUTPUT

table of hits

SW score	perc div.	perc del.	perc ins.	query sequence	position begin	position end	query (left)	matching repeat	repeat class/family
459	20.0	1.1	0.0	LAL_1000603f	2	96	(4) +	Ogre_PS1_from_AY299398	LTR/Gypsy/Ogre
361	27.8	0.0	0.0	LAL_1000670r	1	90	(10) C	Ogre_PA_z_c700	LTR/Gypsy/Ogre_PA
449	23.5	0.0	0.0	LAL_1001438f	1	98	(2) +	PSC454_CL1Contig253	LTR/Gypsy/Ogre
491	23.5	0.0	0.0	LAL_1001438r	3	100	(0) C	PSC454_CL1Contig253	LTR/Gypsy/Ogre
487	24.0	0.0	0.0	LAL_1001478f	1	100	(0) C	PSC454_CL1Contig253	LTR/Gypsy/Ogre
398	28.0	0.0	0.0	LAL_1001478r	1	100	(0) C	PSC454_CL1Contig2773	LTR/Gypsy/Ogre
417	20.6	2.0	2.0	LAL_1002130r	2	100	(0) C	PSC454_CL1Contig253	LTR/Gypsy/Ogre
373	29.0	0.0	0.0	LAL_1002357f	1	100	(0) +	Ogre_PS1_from_AY299398	LTR/Gypsy/Ogre

[illegible]

< masked sequence

```
summary >
table
```

	number of elements*	length occupied	percentage of sequence
Retroelements	1	12014 bp	96.40 %
SINES:	0	0 bp	0.00 %
Penelope	0	0 bp	0.00 %
LINES:	0	0 bp	0.00 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	0	0 bp	0.00 %
R1/L0A/Jockey	0	0 bp	0.00 %
R2/R4/NeSL	0	0 bp	0.00 %
RTE/Bov-B	0	0 bp	0.00 %
L1/CIN4	0	0 bp	0.00 %
LTR elements:	1	12014 bp	96.40 %
BEL/Pao	0	0 bp	0.00 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	1	12014 bp	96.40 %
Retroviral	0	0 bp	0.00 %
DNA transposons	0	0 bp	0.00 %
hobo-Activator	0	0 bp	0.00 %
Tcl-IS630-Pogo	0	0 bp	0.00 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	0	0 bp	0.00 %
Other (Mirage, P-element, Transib)	0	0 bp	0.00 %

Principles and tools for global repeat identification

		Assembled genome or long contigs (BACs etc.)	
		YES	NO
Reference database	YES	RepeatMasker / RepBase • <i>first “model” species</i> • <i>extensive resources (+ wet lab)</i>	• <i>model-related taxa</i> • <i>often shotgun, “assembled”</i>
	NO	• dot matrix (dot-plot)	

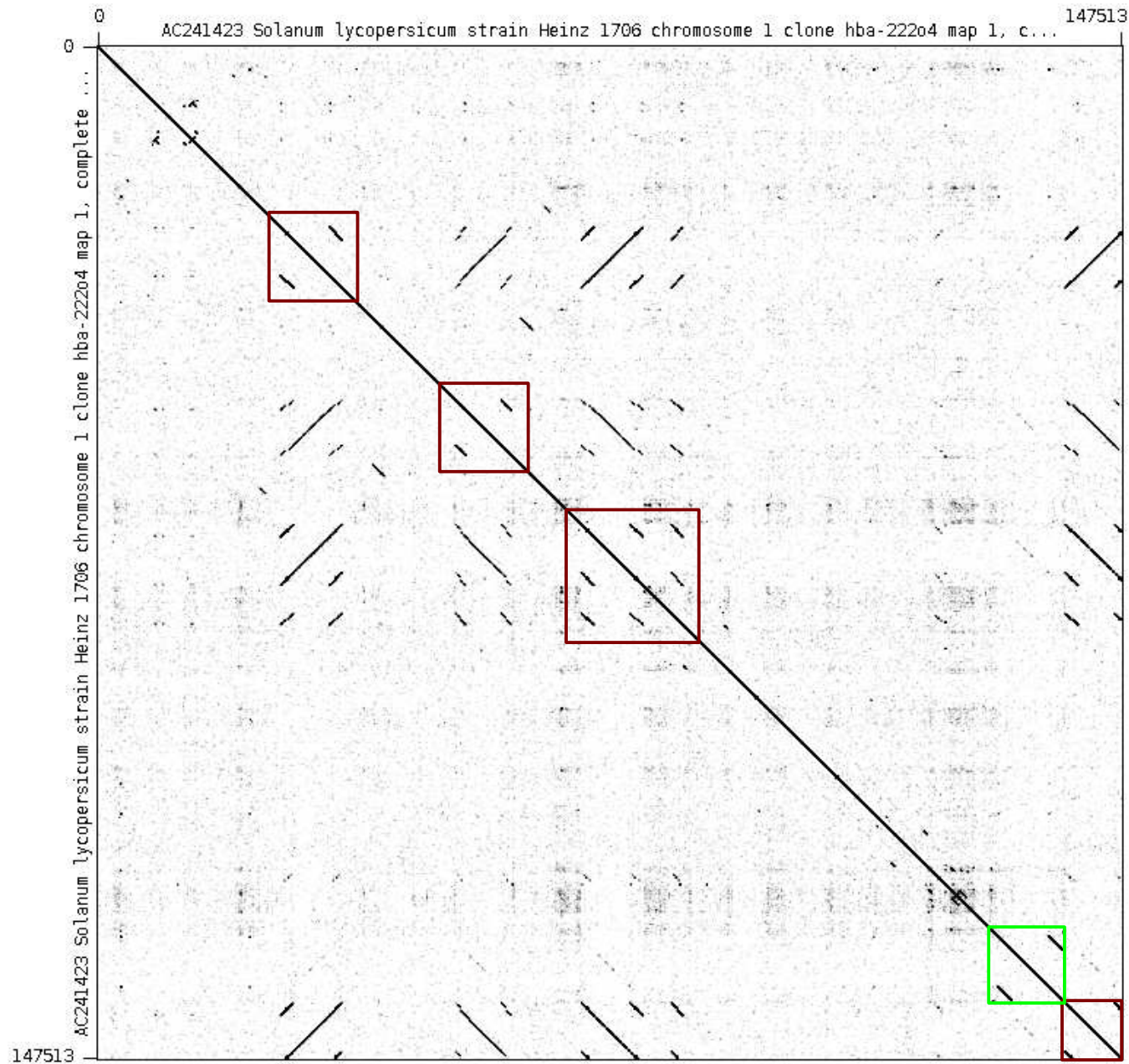
Repeat identification in sequence data

- **Pure algorithmic (evaluates “repetitiveness” of substrings in DNA sequence)**
- Signature or feature-based (searches for specific structures of biological importance)
- Similarity to known repeats (uses curated databases of known repeats)

Principles and tools for global repeat identification

dot-plot

(GEnome PAir - Rapid Dotter, Krumsiek et al., 2007)



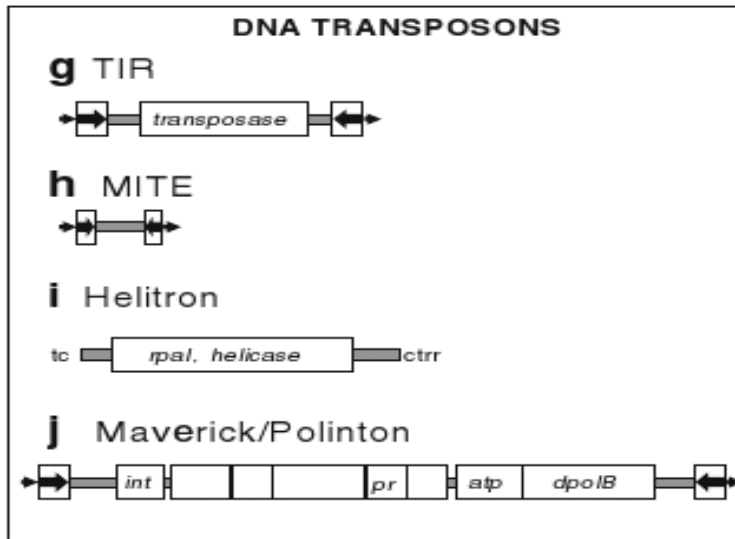
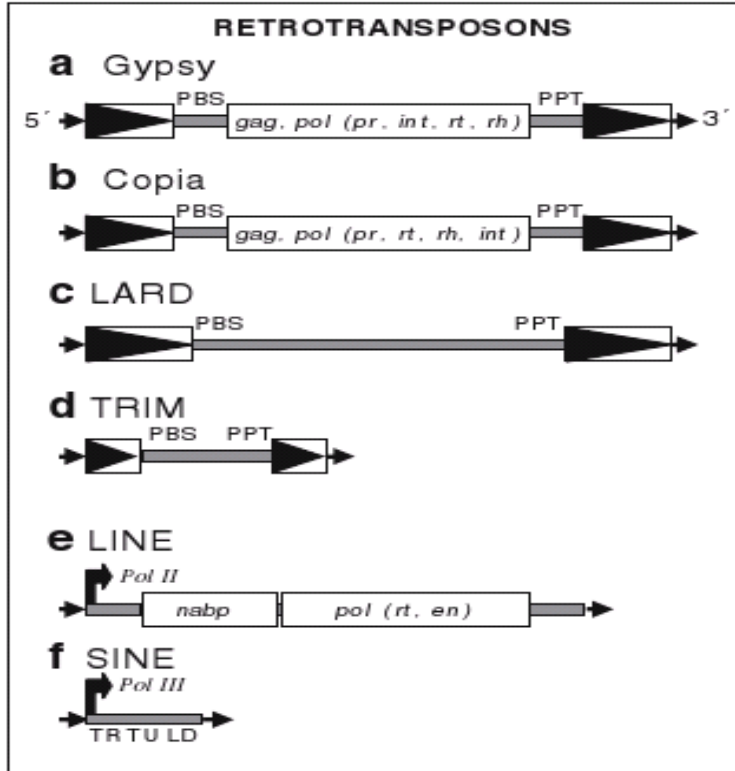
Principles and tools for global repeat identification

		Assembled genome or long contigs (BACs etc.)	
		YES	NO
Reference database	YES	RepeatMasker / RepBase • <i>first “model” species</i> • <i>extensive resources (+ wet lab)</i>	• <i>model-related taxa</i> • <i>often shotgun, “assembled”</i>
	NO	• dot matrix (dot-plot) • PILER (local alignment) • LTR_STRUC, MITE-Hunter , etc.	

Repeat identification in sequence data

- Pure algorithmic (evaluates “repetitiveness” of substrings in DNA sequence)
- **Signature or feature-based (searches for specific structures of biological importance)**
- Similarity to known repeats (uses curated databases of known repeats)

Principles and tools for global repeat identification



Signature / structure-based approaches

LTR_STRUC (McCarthy and McDonald, 2003)

LTR_FINDER (Xu and Wang, 2007)

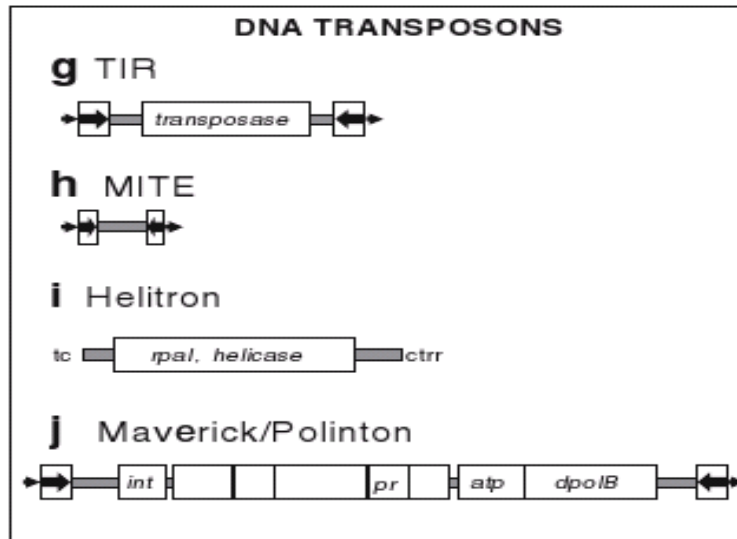
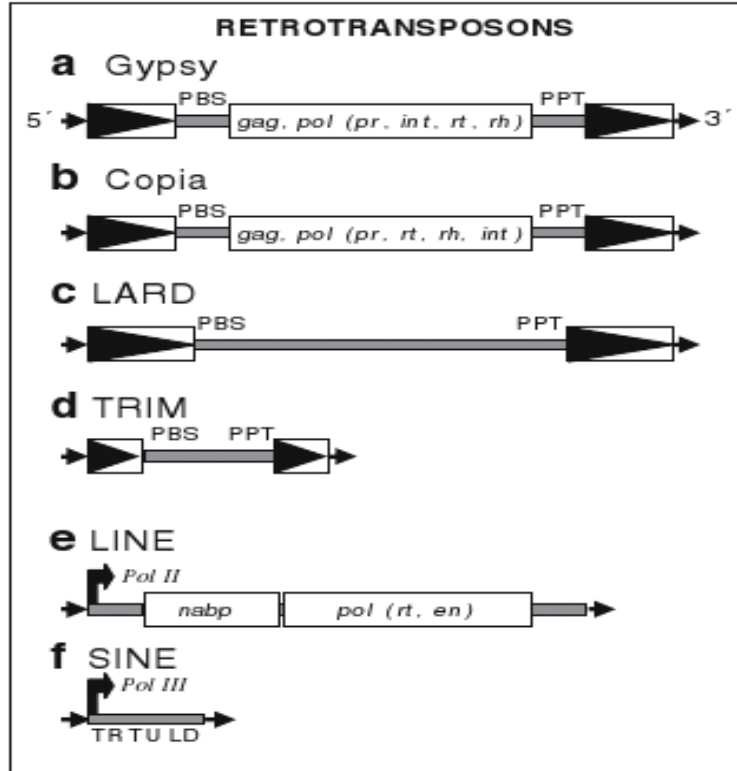
SINE-Finder (Wenke et al. 2011)

MITE-Hunter (Han and Wessler, 2010)

MITE Analysis Toolkit (Yang and Hall, 2003)

HelitronFinder (Du et al. 2008) [HelA helitrons in maize]

Principles and tools for global repeat identification



Saha et al. (20008)

Signature / structure-based approaches

LTR_STRUC (McCarthy and McDonald, 2003)

LTR_FINDER (Xu and Wang, 2007)

SINE-Finder (Wenke et al. 2011)

MITE-Hunter (Han and Wessler, 2010)

MITE Analysis Toolkit (Yang and Hall, 2003)

HelitronFinder (Du et al. 2008) [HelA helitrons in maize]

Tandem Repeats Finder (Benson, 1999)

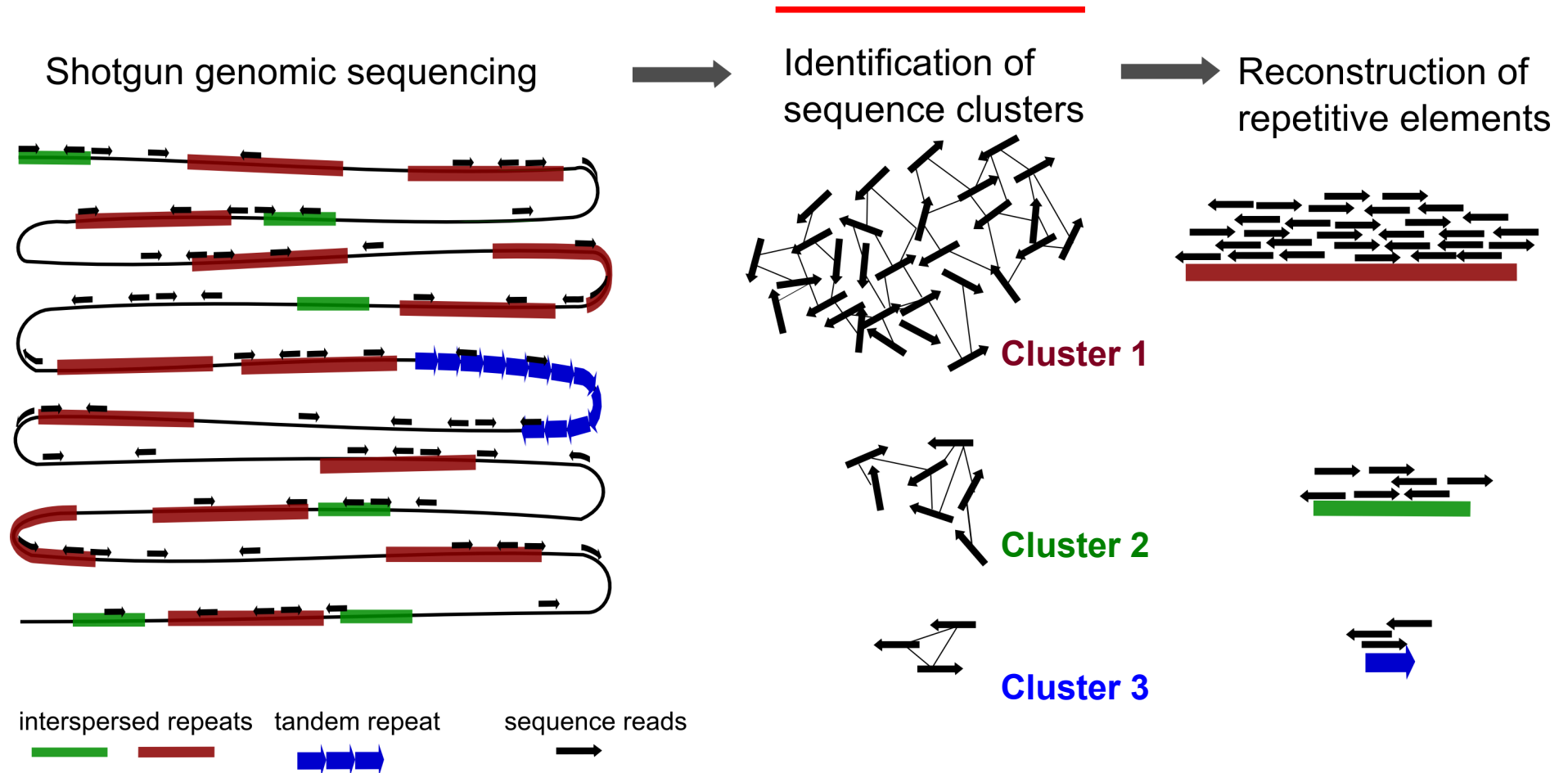
Principles and tools for global repeat identification

		Assembled genome or long contigs (BACs etc.)	
		Y E S	N O
Reference database	Y E S	RepeatMasker / RepBase • <i>first “model” species</i> • <i>extensive resources (+ wet lab)</i>	• <i>model-related taxa</i> • <i>often shotgun, “assembled”</i>
	N O	• dot matrix (dot-plot) • PILER (local alignment) • LTR_STRUC, MITE-Hunter , etc.	• RECON • <i>(Next-gen sequencing)</i> • direct assembly (phrap,...) • clustering-based

Repeat identification in sequence data

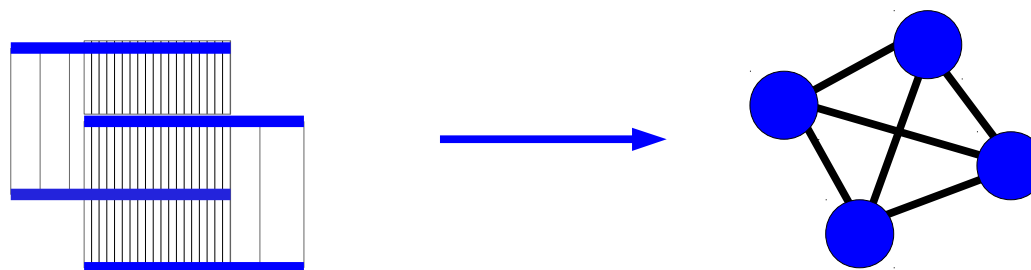
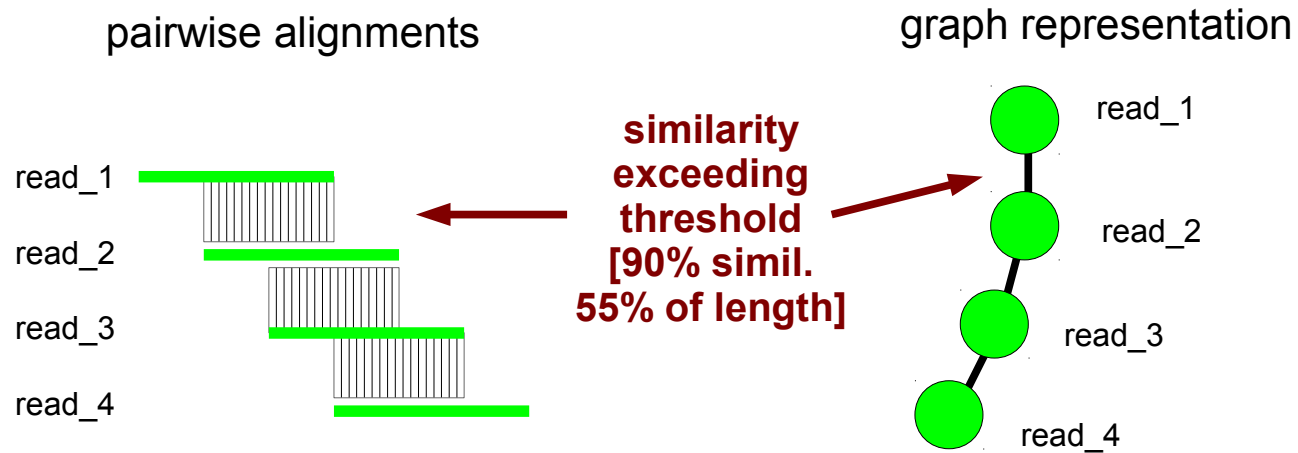
- Pure algorithmic (evaluates “repetitiveness” of substrings in DNA sequence)
- Signature or feature-based (searches for specific structures of biological importance)
- Similarity to known repeats (uses curated databases of known repeats)

Clustering-based repeat identification

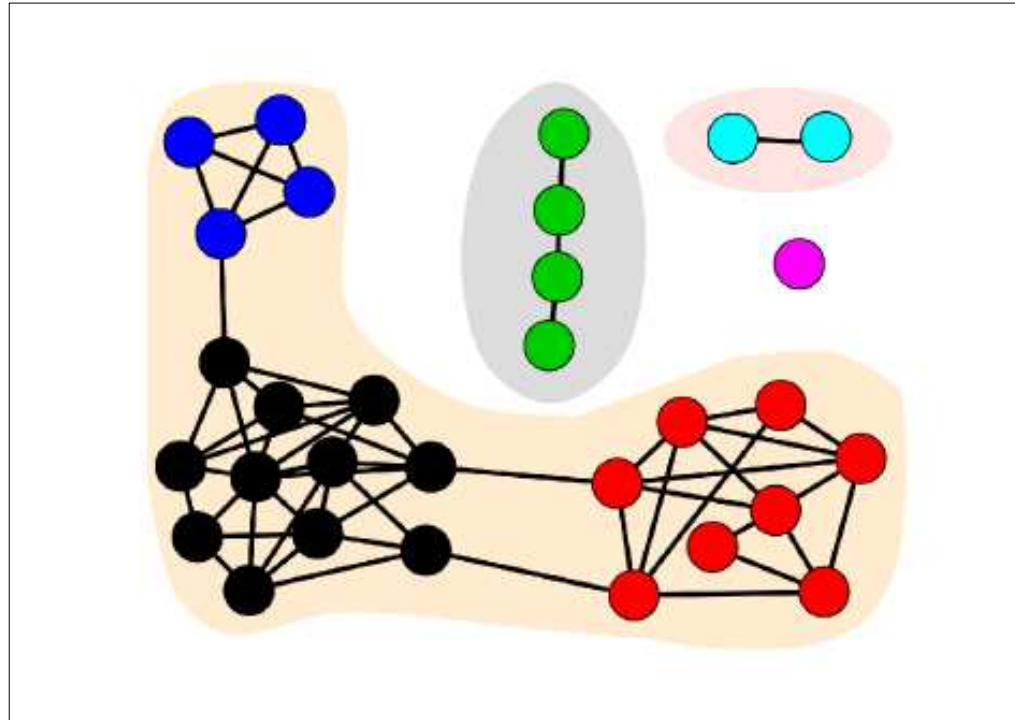


CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

Clustering-based repeat identification



Clustering-based repeat identification

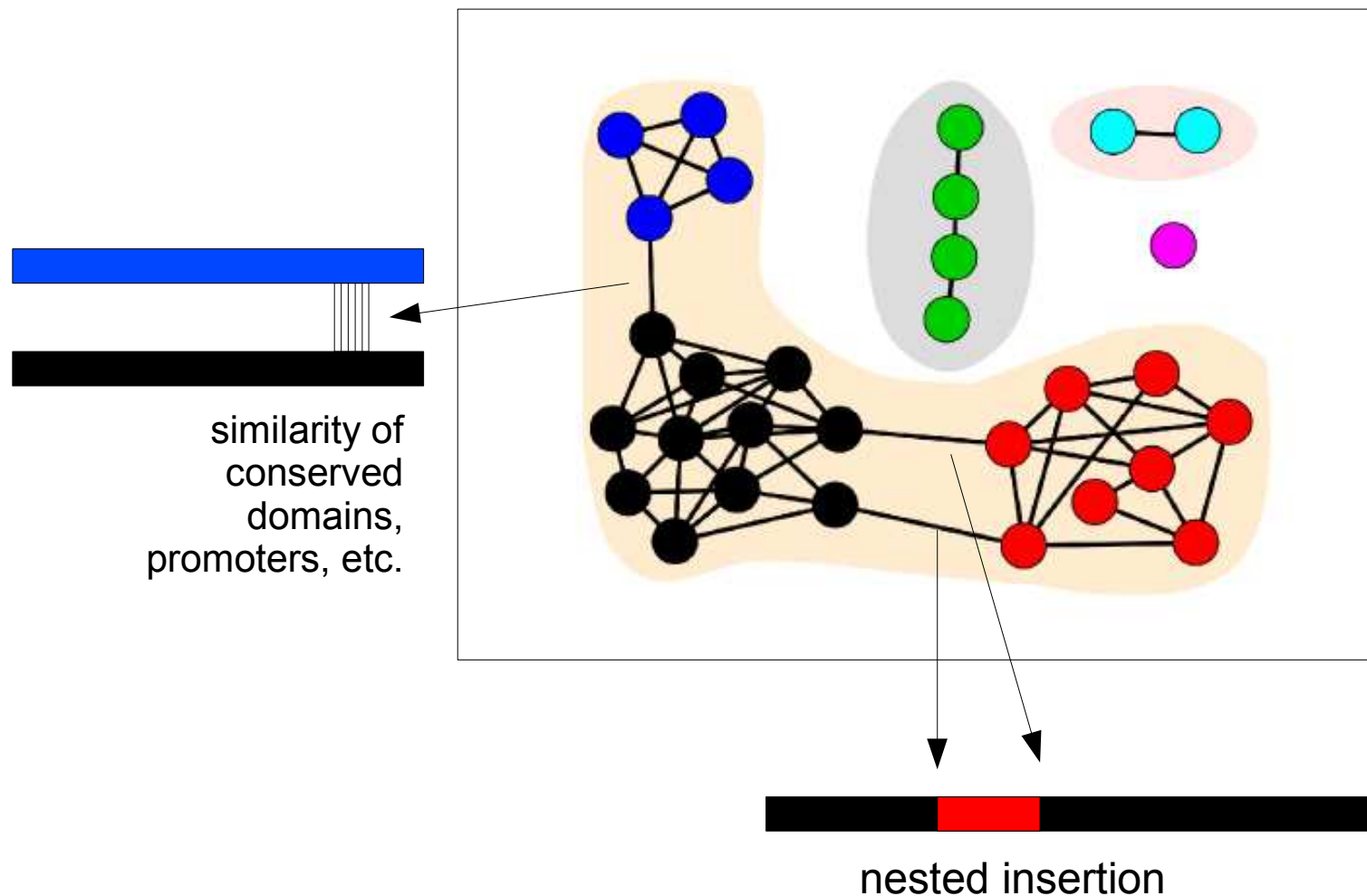


Single linkage clustering => connected components

TGICL
(TIGR Gene Indices clustering tool)
Pertea et al., 2003

Macas et al. (2007) - Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*.

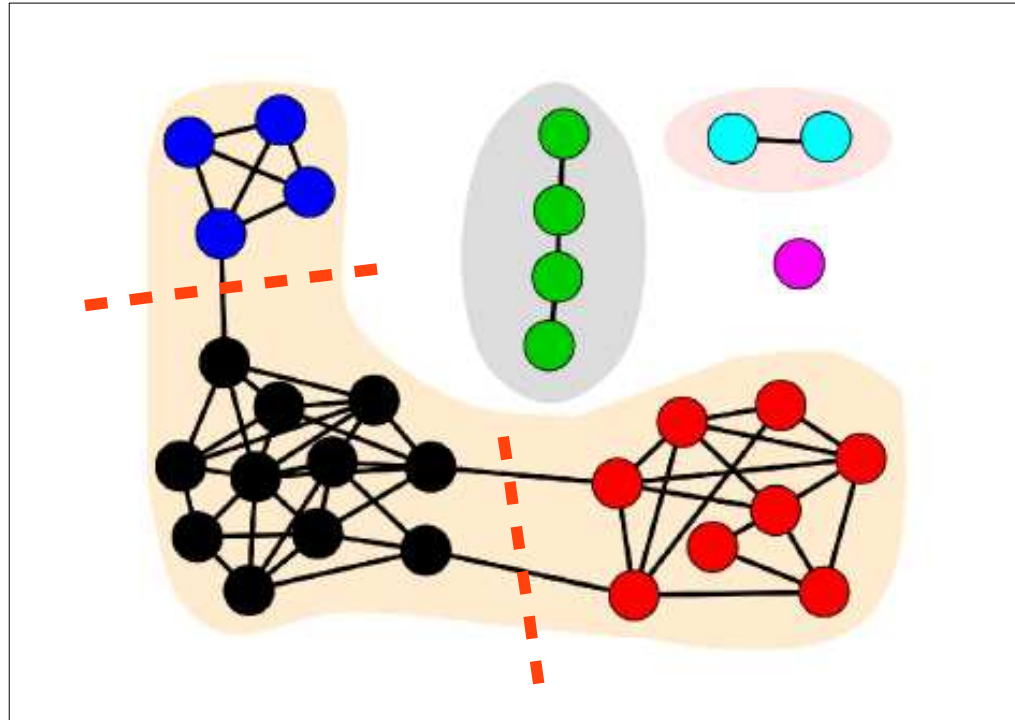
Clustering-based repeat identification



Single linkage clustering => connected components

TGICL
(TIGR Gene Indices clustering tool)
Pertea et al., 2003

Graph-based clustering

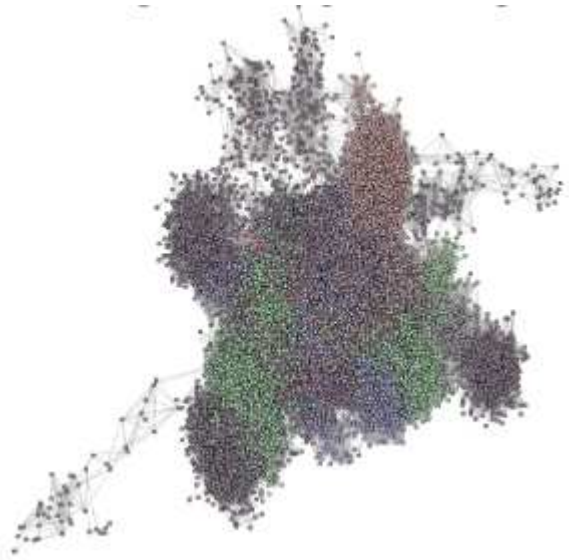


Graph-based clustering

- Sequence overlaps between the reads are transformed to a graph where the **reads** are represented as **nodes** and their **similarities** as **edges** connecting the nodes
- Graph structure is examined to detect *communities of frequently connected nodes* which are split to separate clusters



Graph-based characterization of repeat clusters

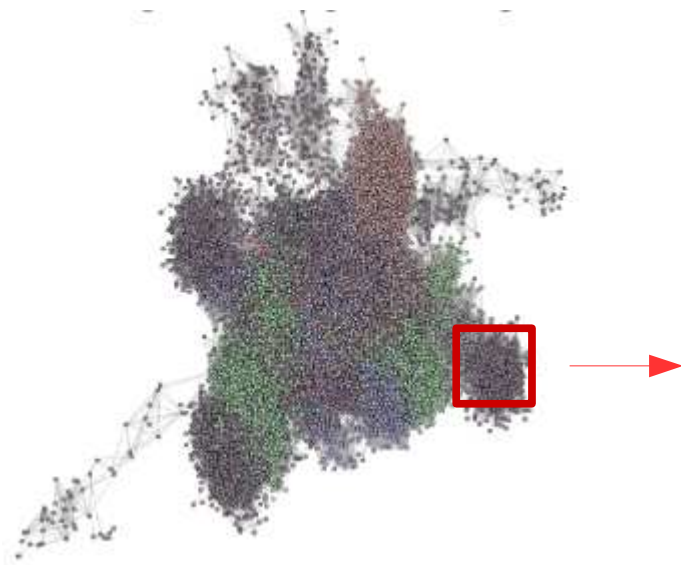


Virtual graphs used
to analyze real data
contain **up to millions**
of nodes (reads)

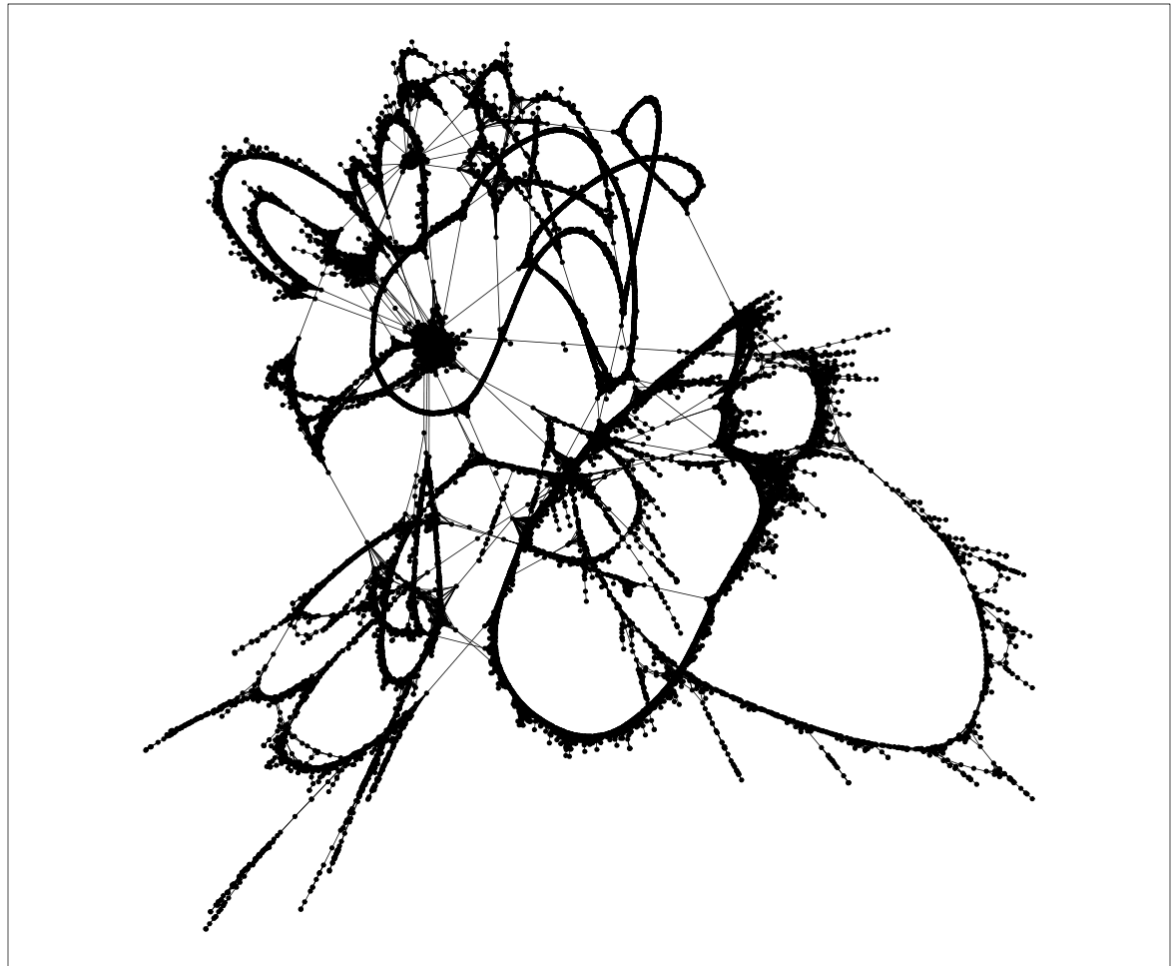
Graph-based clustering

- Sequence overlaps between the reads are transformed to a graph where the **reads** are represented as **nodes** and their **similarities** as **edges** connecting the nodes
- Graph structure is examined to detect *communities of frequently connected nodes* which are split to separate clusters

Graph-based characterization of repeat clusters



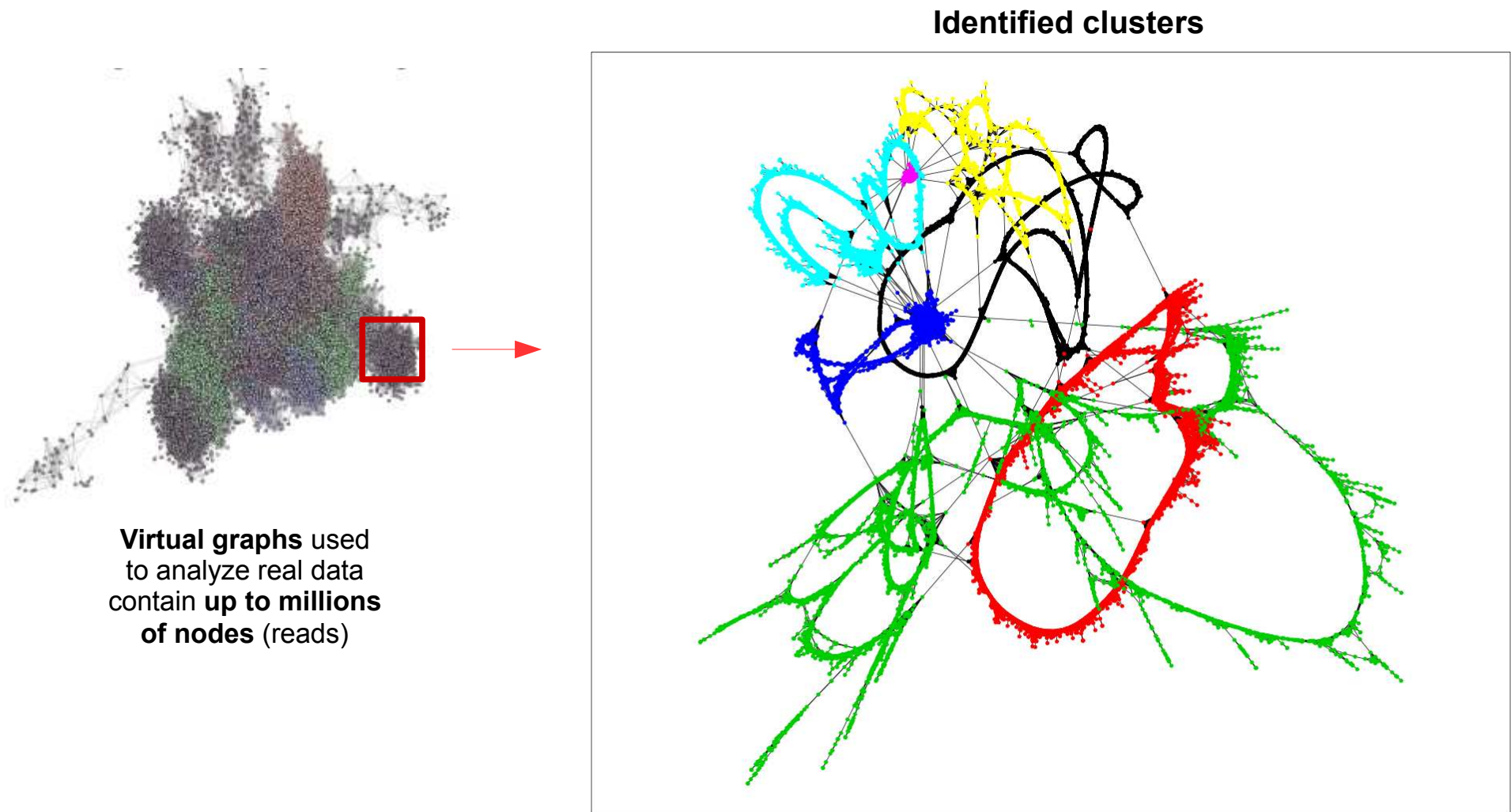
Virtual graphs used to analyze real data contain **up to millions of nodes** (reads)



Graph-based clustering

- Sequence overlaps between the reads are transformed to a graph where the **reads** are represented as **nodes** and their **similarities** as **edges** connecting the nodes
- Graph structure is examined to detect *communities of frequently connected nodes* which are split to separate clusters

Graph-based characterization of repeat clusters



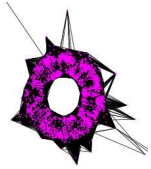
Graph-based clustering

- Sequence overlaps between the reads are transformed to a graph where the **reads** are represented as **nodes** and their **similarities** as **edges** connecting the nodes
- Graph structure is examined to detect *communities of frequently connected nodes* which are split to separate clusters

Graph-based characterization of repeat clusters

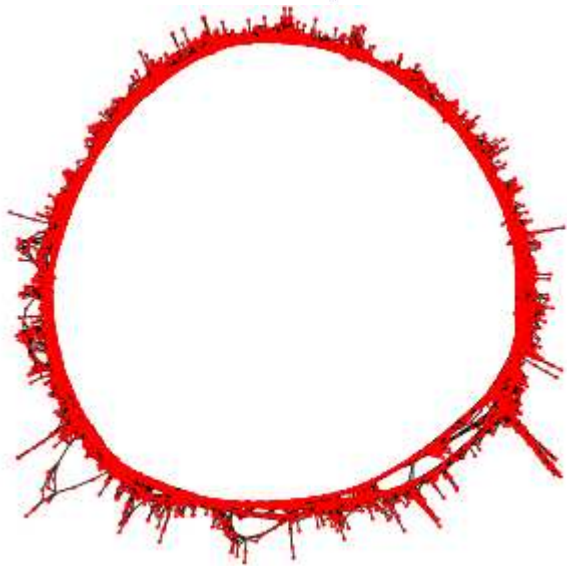
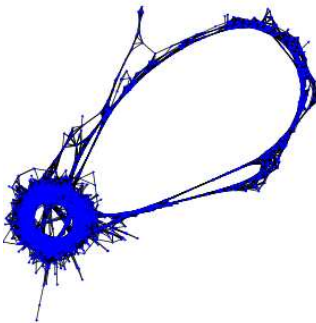
Extraction of individual clusters
(graph shapes indicate repeat types)

Identified clusters

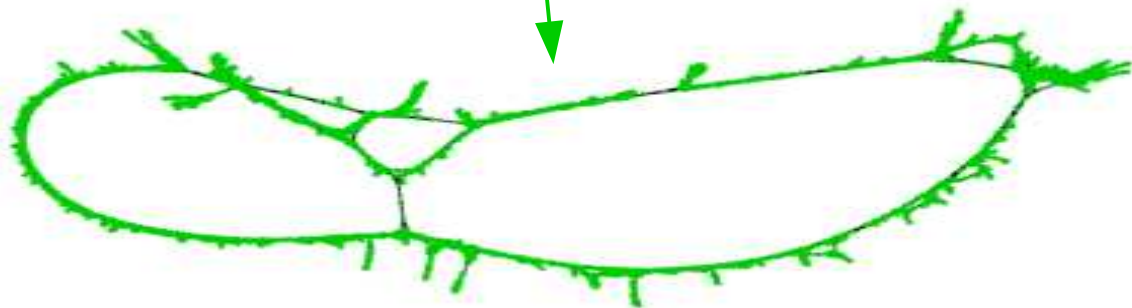
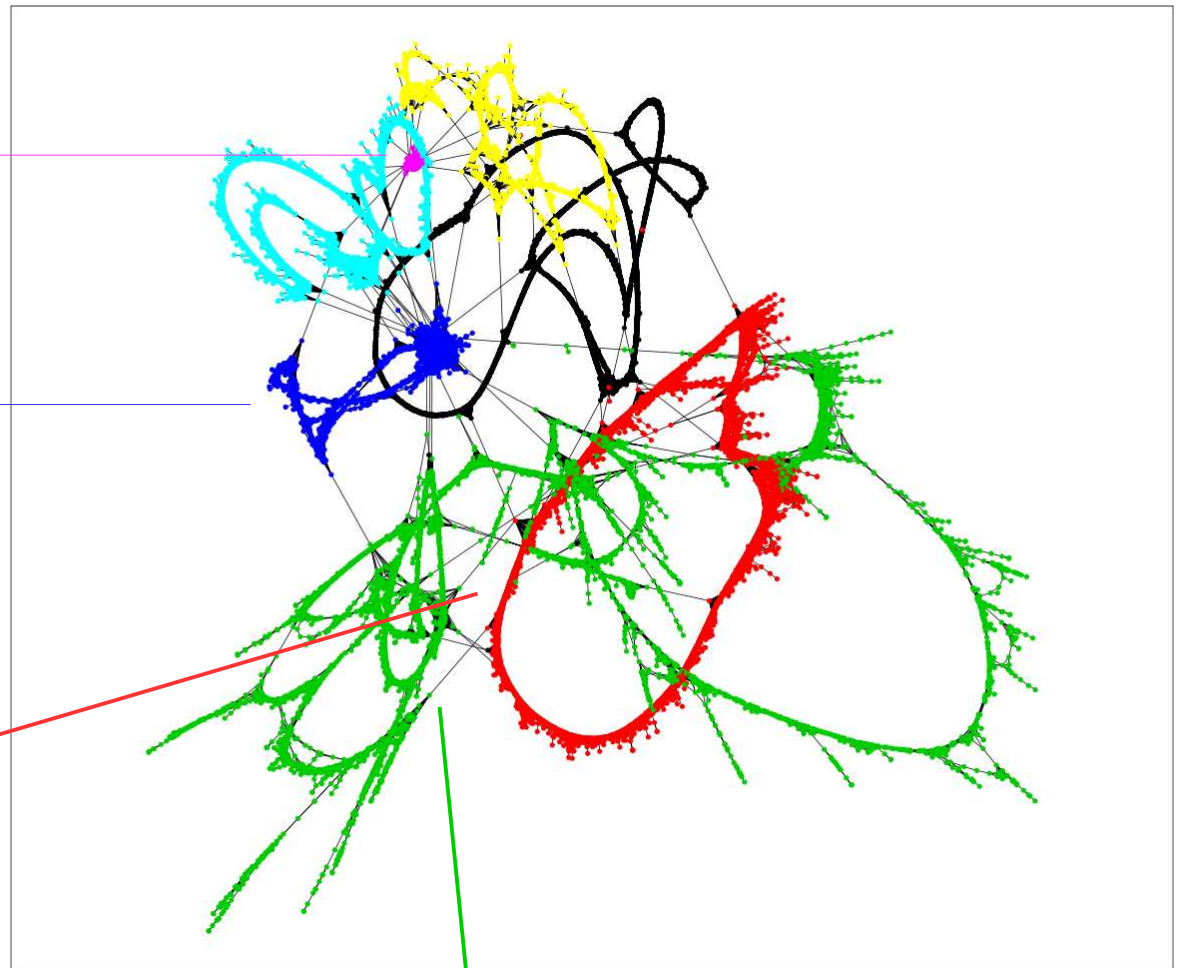


Satellite
(simple)

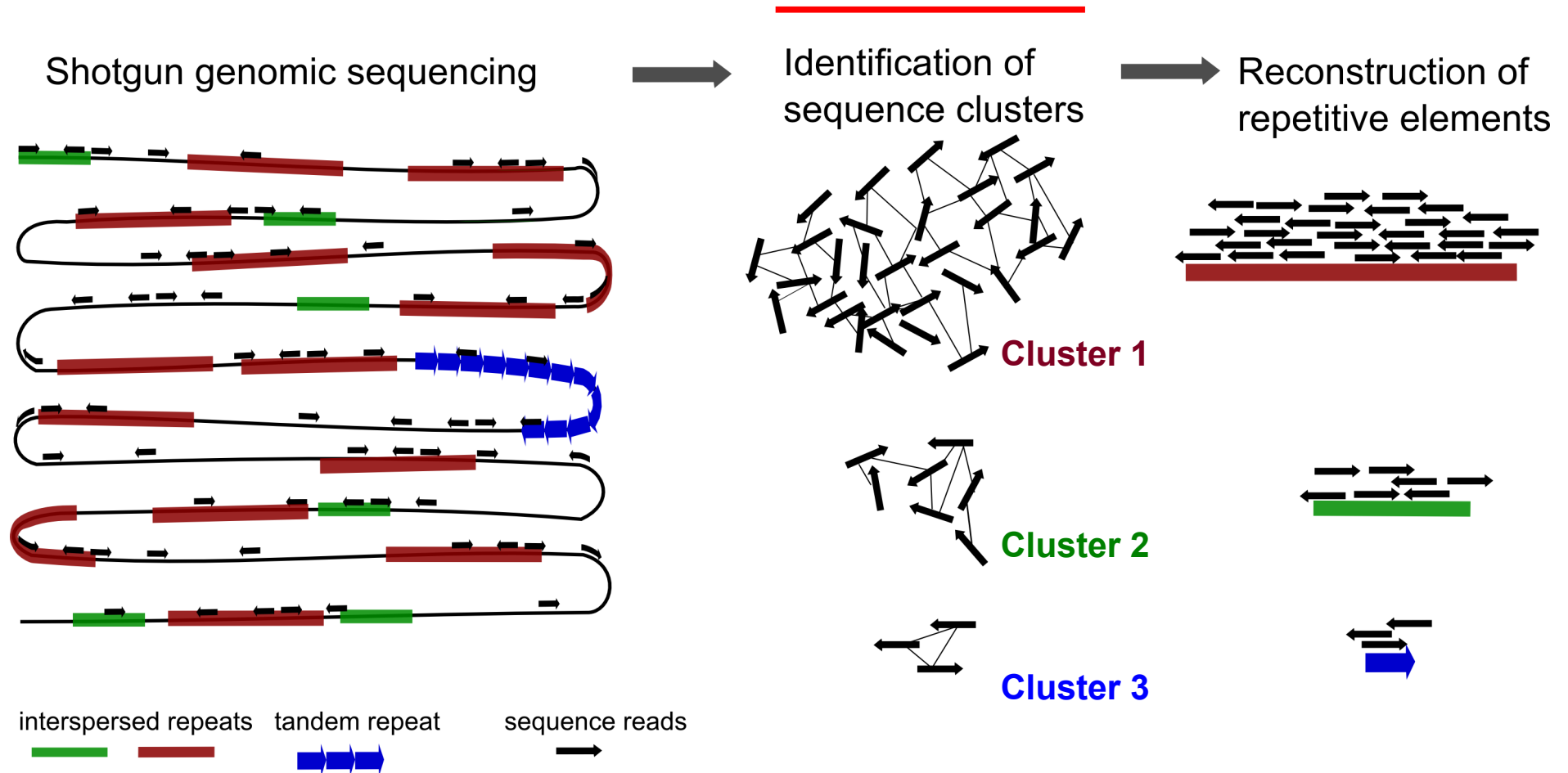
Satellite family
(short + long monomers) >



LTR-retrotransposons



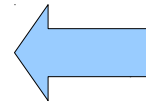
Clustering-based repeat identification



CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

Repeat characterization

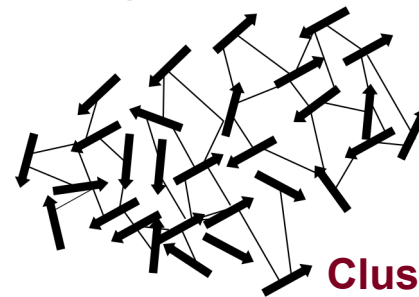
Cluster annotation and quantification



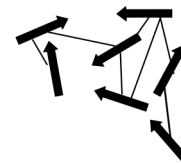
Identification of sequence clusters



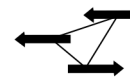
Reconstruction of repetitive elements



Cluster 1



Cluster 2

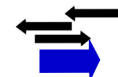


Cluster 3

...

...

(Cluster x)



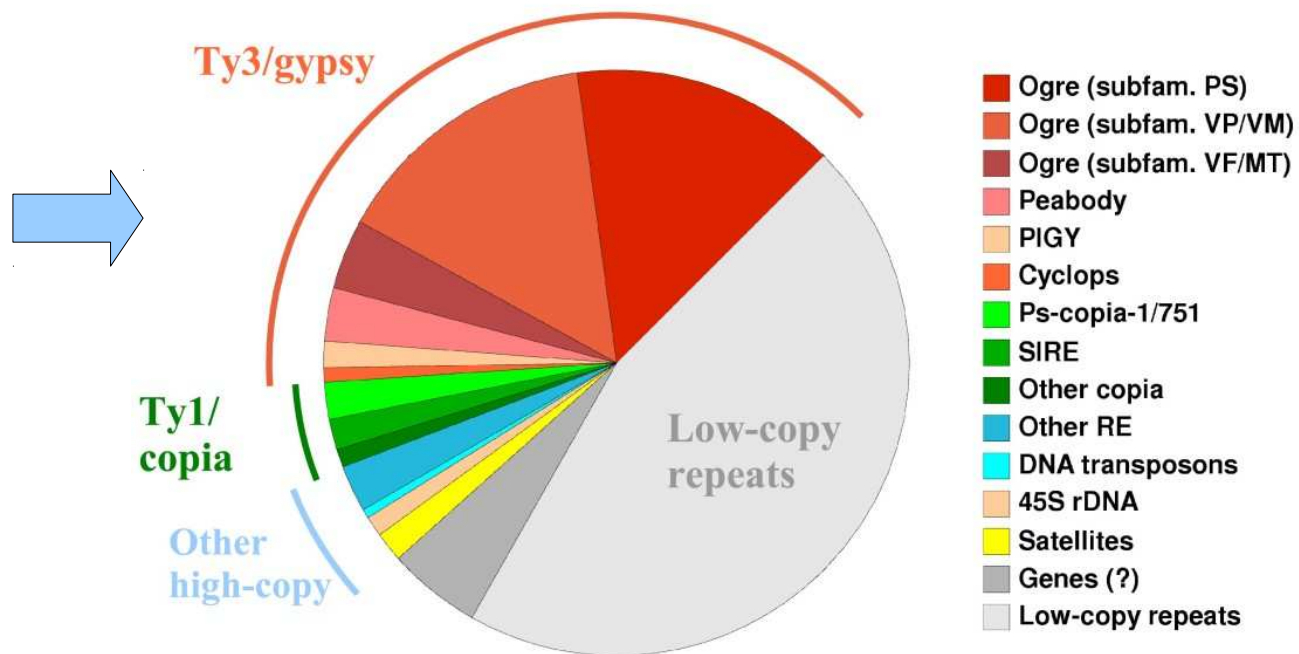
	7192836 (= 100%)	64.1			
CL	reads	genome %	class	type	note
1	304159	4.229	gypsy	Tat	PROT-RT/RH-INT
2	234749	3.264			?, PE->24,18
3	216307	3.007	gypsy	chromo	ALL domains
4	202822	2.820	copla	Maximus	ALL domains
5	149693	2.081	gypsy	Athila	ALL domains
6	145911	2.029	gypsy	Tat	ALL domains
7	143766	1.999	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			?

Repeat characterization

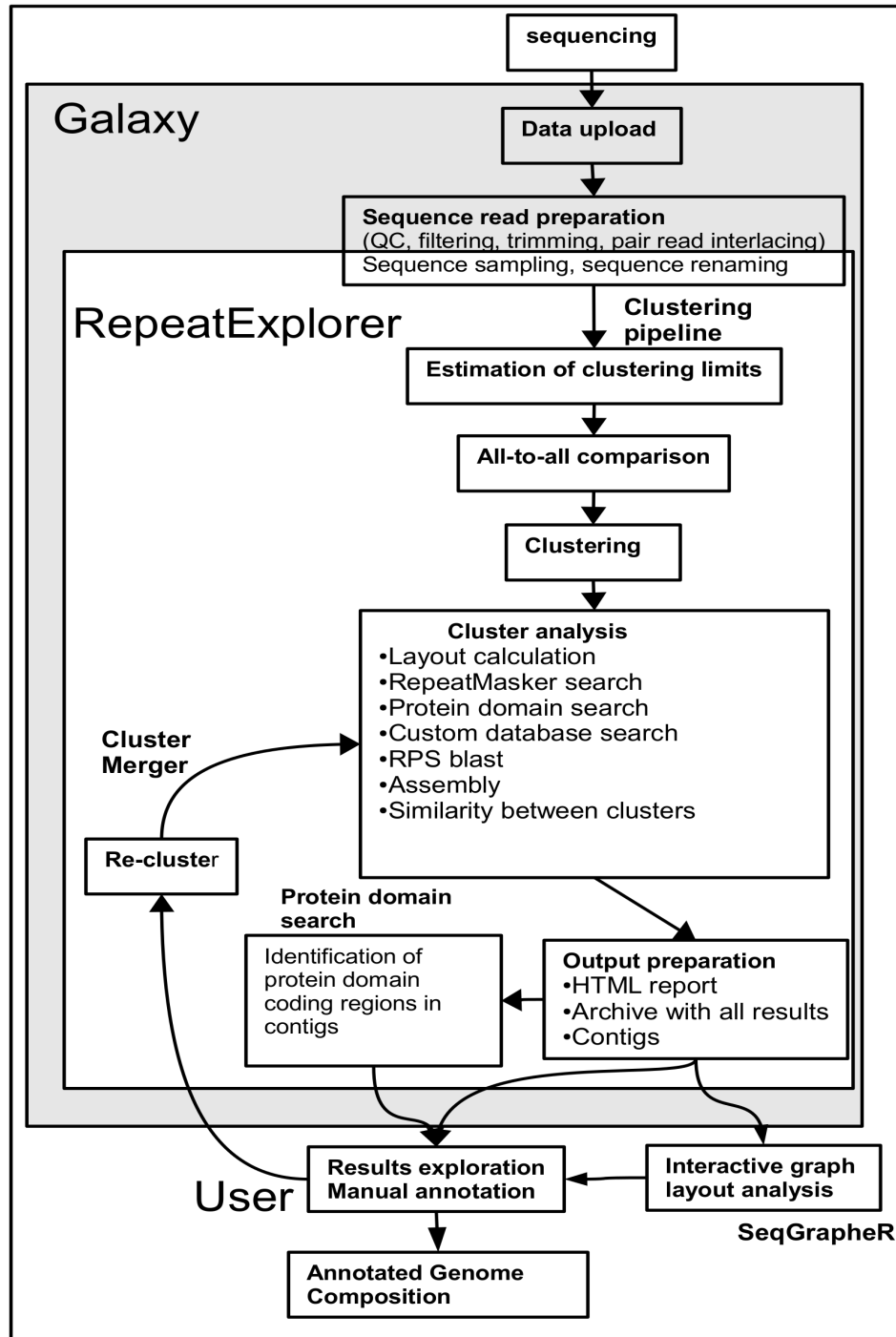
Cluster annotation and quantification

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

Proportions of various repeat types in a genome



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

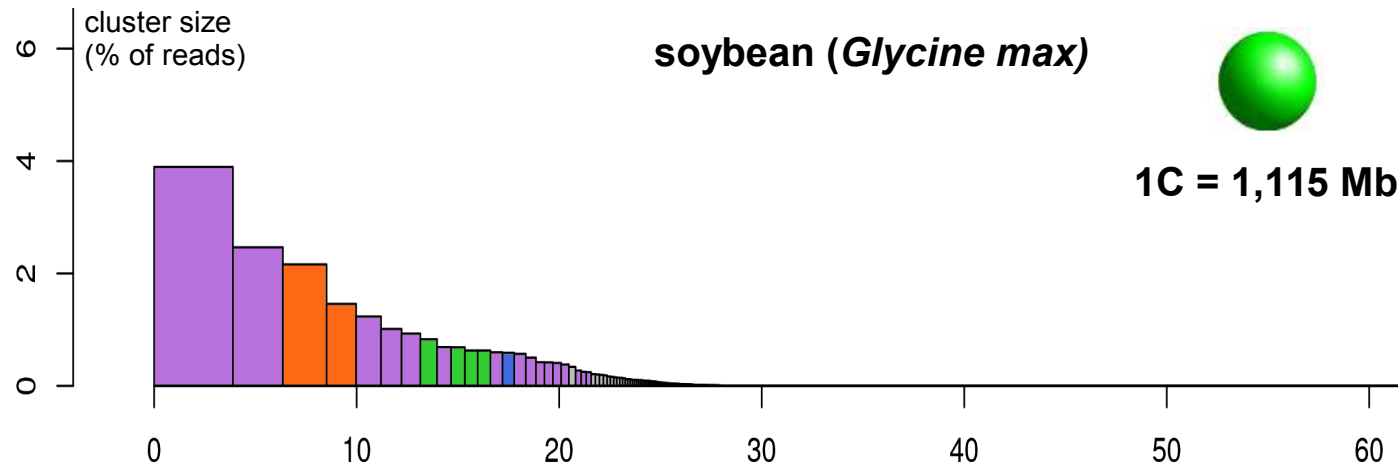
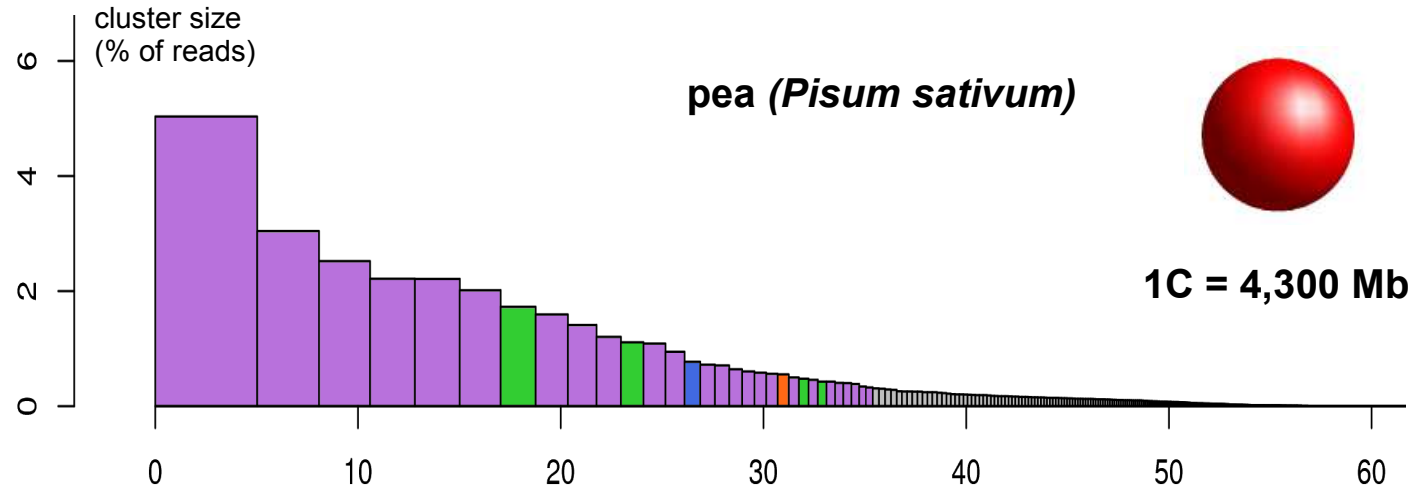
Similarity searches

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

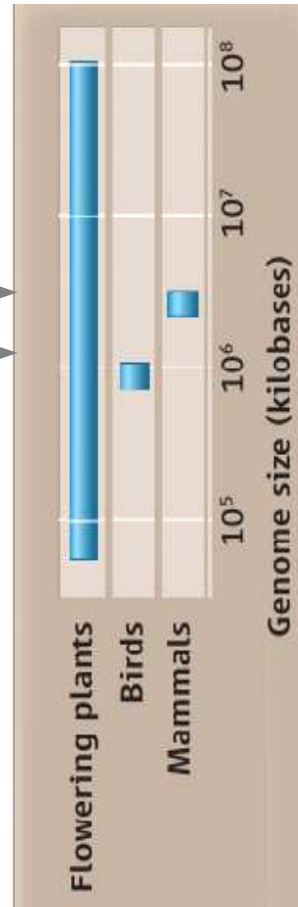
Repeat characterization is achieved by combining all three approaches

Repeat composition of individual species

Repeat type: Ty3/gypsy Ty1/copia Satellite DNA rDNA other/unknown



cumulative % of reads



Comparative analysis

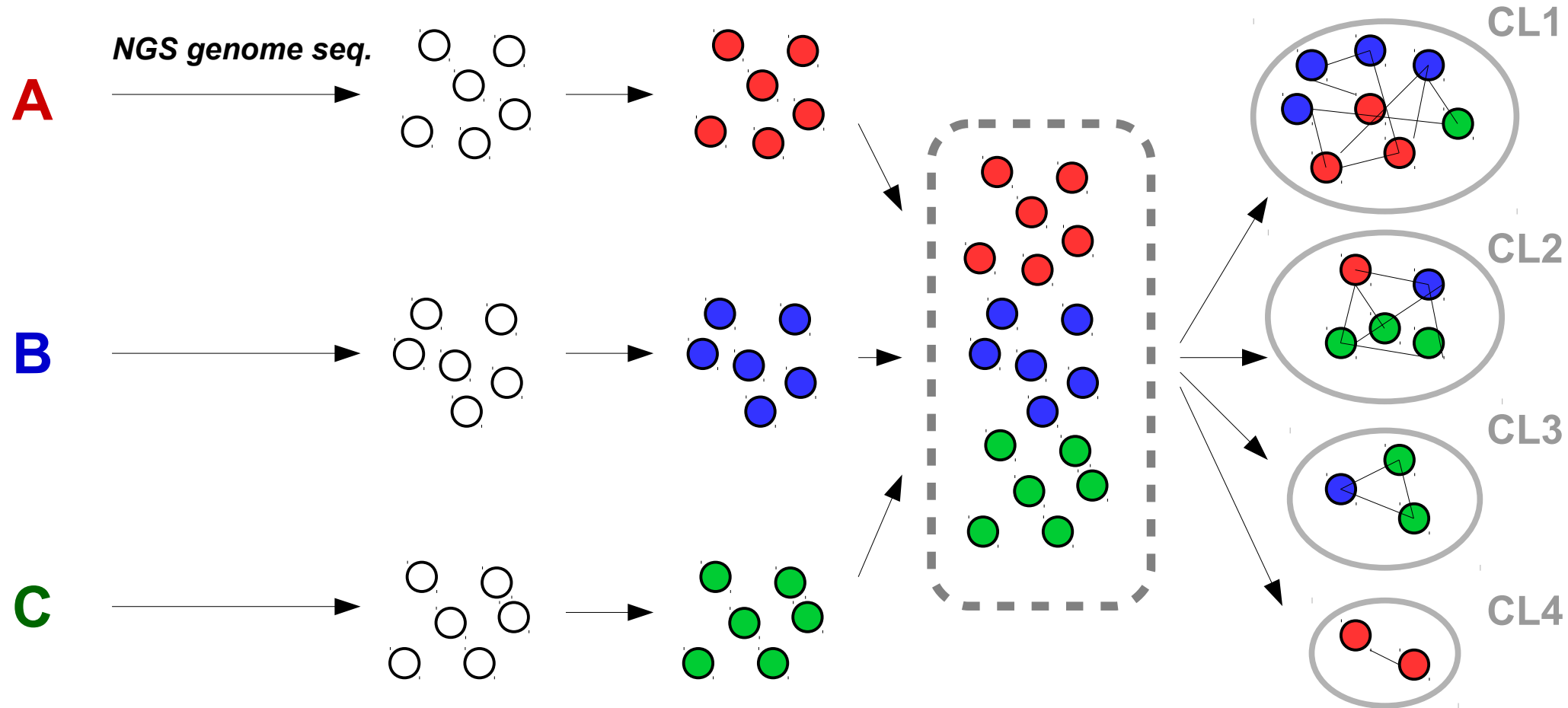
SPECIES
(samples)

Reads

Modify names

Merge

CLUSTERS



Comparative analysis

Sequence reads from **multiple species are mixed** and subjected to clustering analysis

- Efficient identification of homologous repeats from different species
- Easy quantification



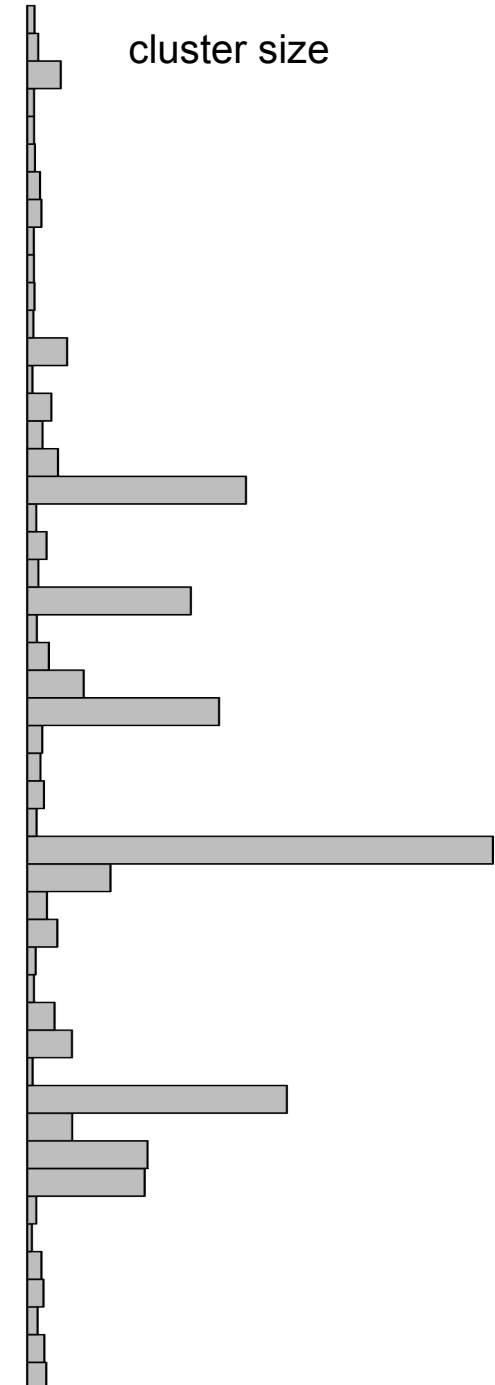
SPECIES:

	<i>Musa ornata</i>
	<i>M. balbisiana</i> (B)
	<i>M. acuminata</i> (A)
	<i>M. textillis</i> (T)
	<i>M. beccarii</i>
	<i>Ensete gillettii</i>

proportion
of reads from
individual
species >>>



cluster size



Comparative analysis

A

SPECIES:

Ensete gillettii

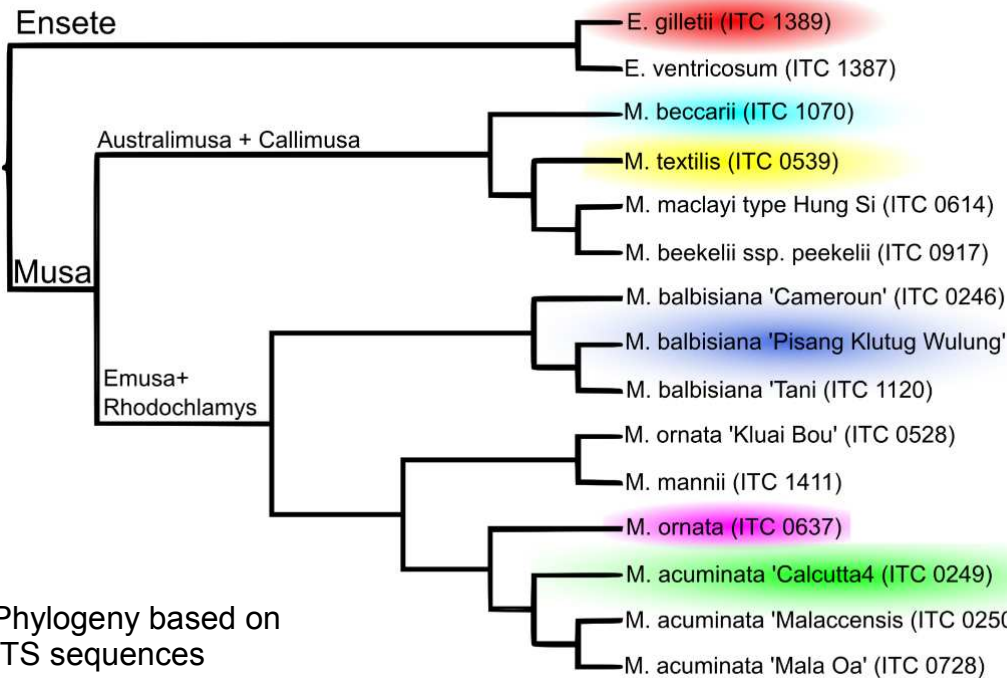
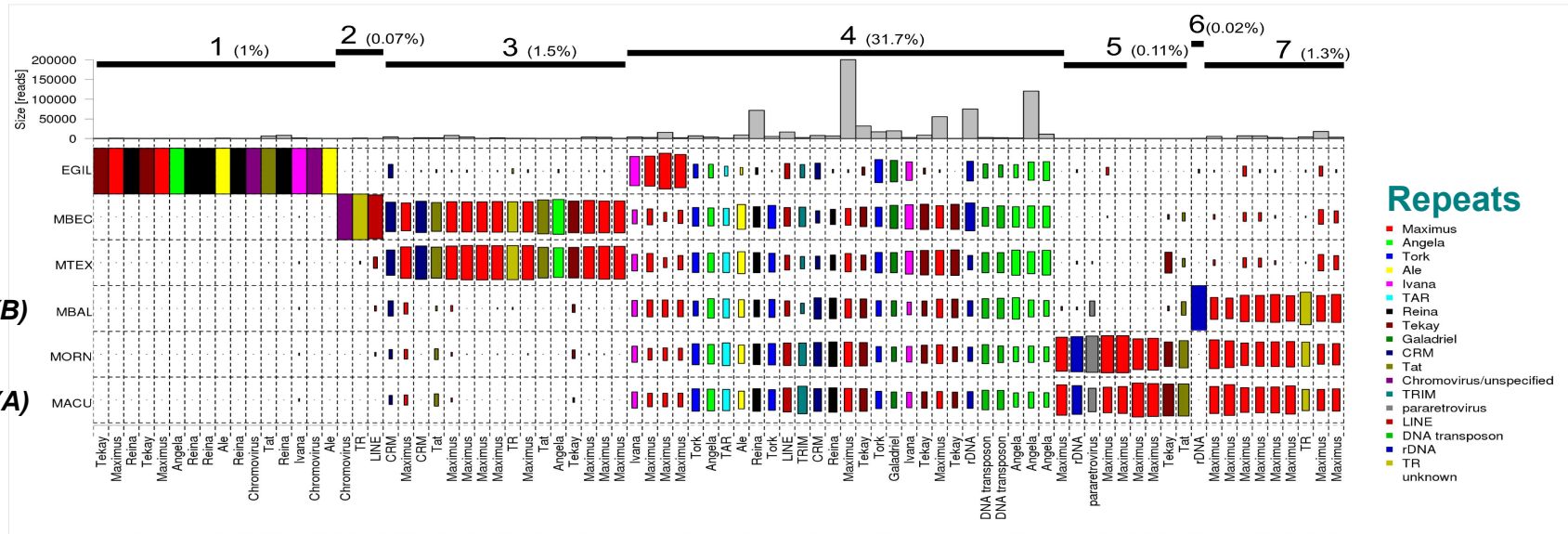
Musa beccarii

M. textillis (T)

M. balbisiana (B)

M. ornata

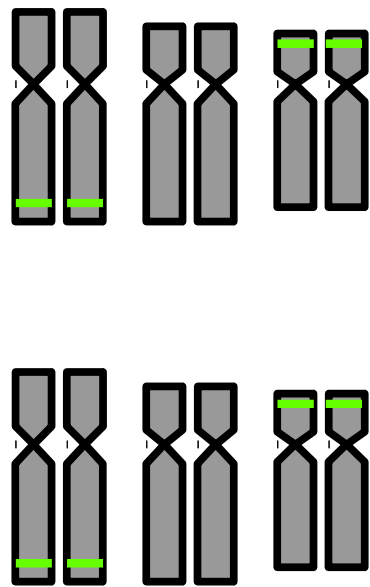
M. acuminata (A)



Phylogeny based on
ITS sequences



Comparative analysis (searching for B-specific repeats)



Next-gen
sequencing

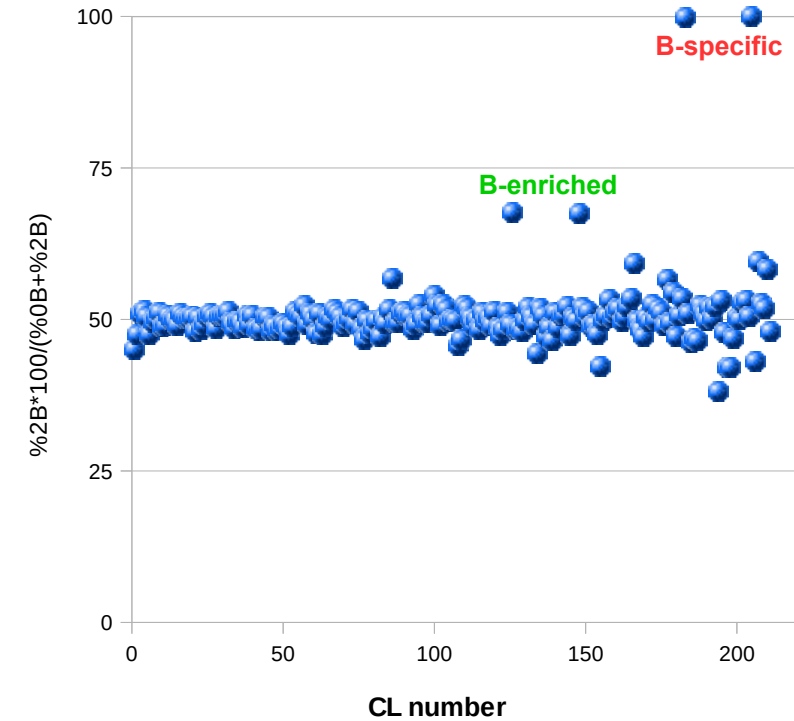
As + Bs

Next-gen
sequencing

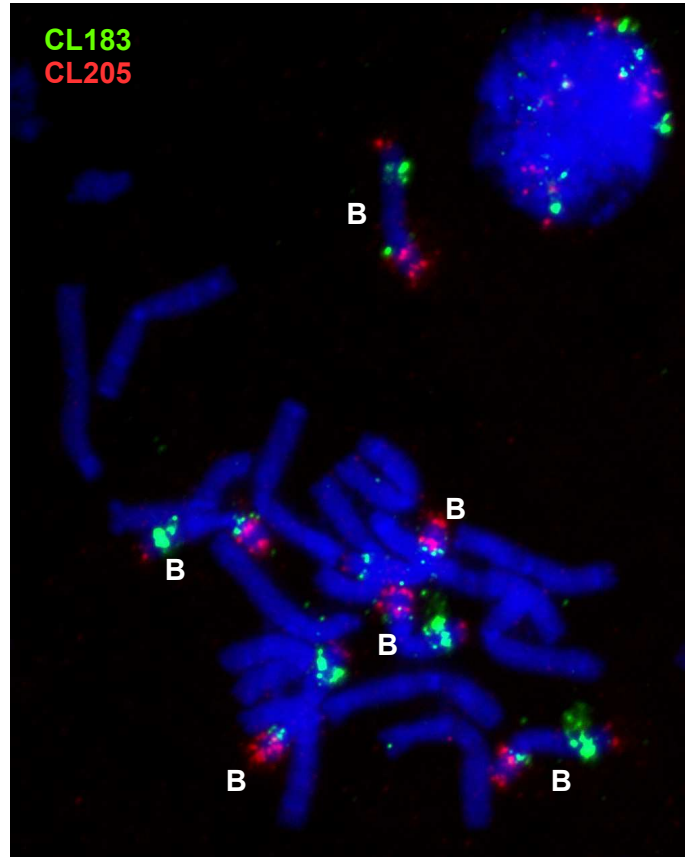
As - 0 -

Mix coded
reads
and
perform
CLUSTERING

Detection of clusters enriched on Bs

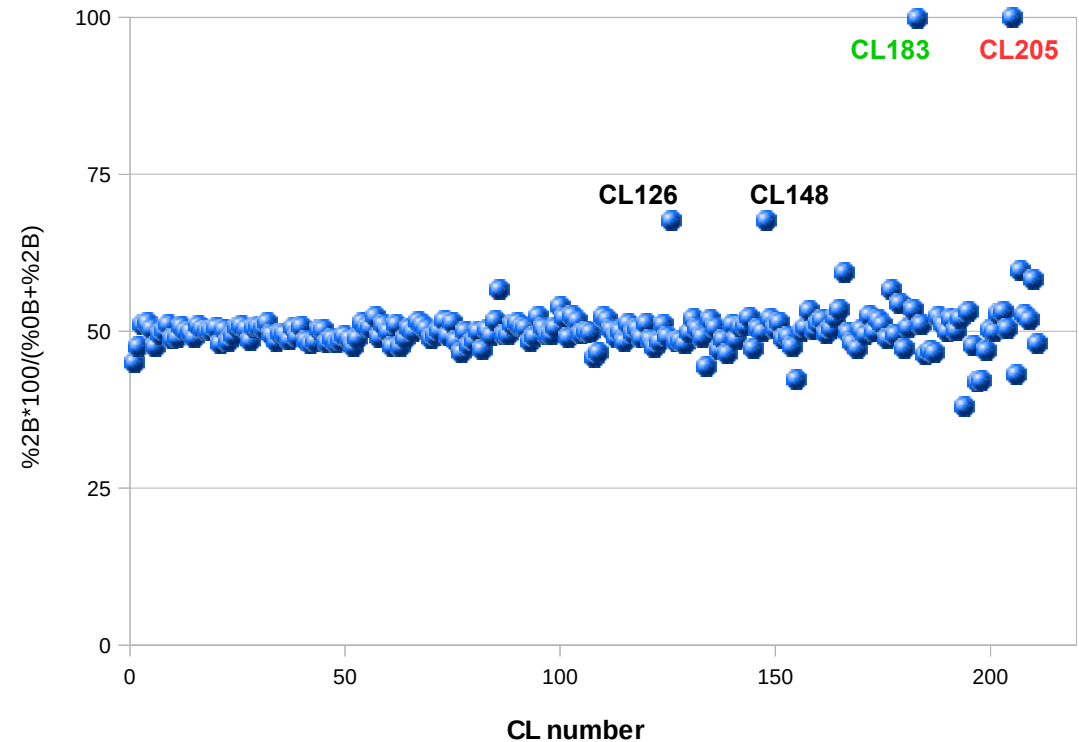


Aegilops speltoides, comparative analysis of 0B / 2B plants



FISH by Alevtina Ruban

Detection of clusters enriched on Bs



Sequencing by Houben lab

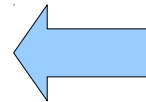
Repeats significantly enriched on Bs

CL	reads	[%]	0B	%	2B	%	$\frac{\%2B \cdot 100}{\%2B + \%0B}$	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

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



CLUSTERS

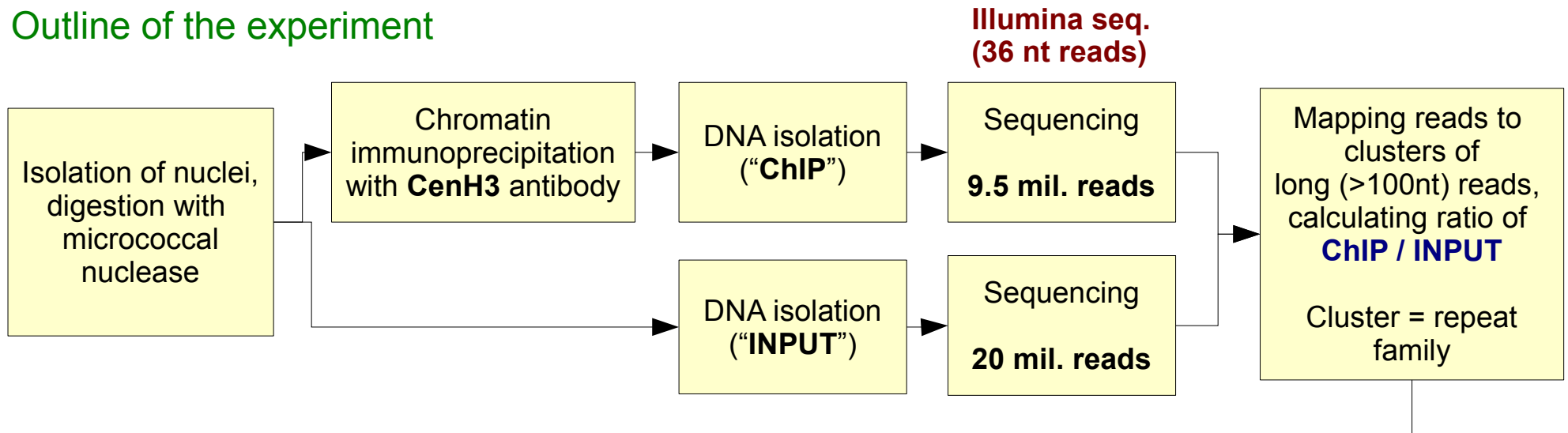
- Represent specific repeat families/variants or their parts
- They are collections of sequence reads capturing full sequence variability of repeat populations

Using clusters as reference for similarity-based classification of:

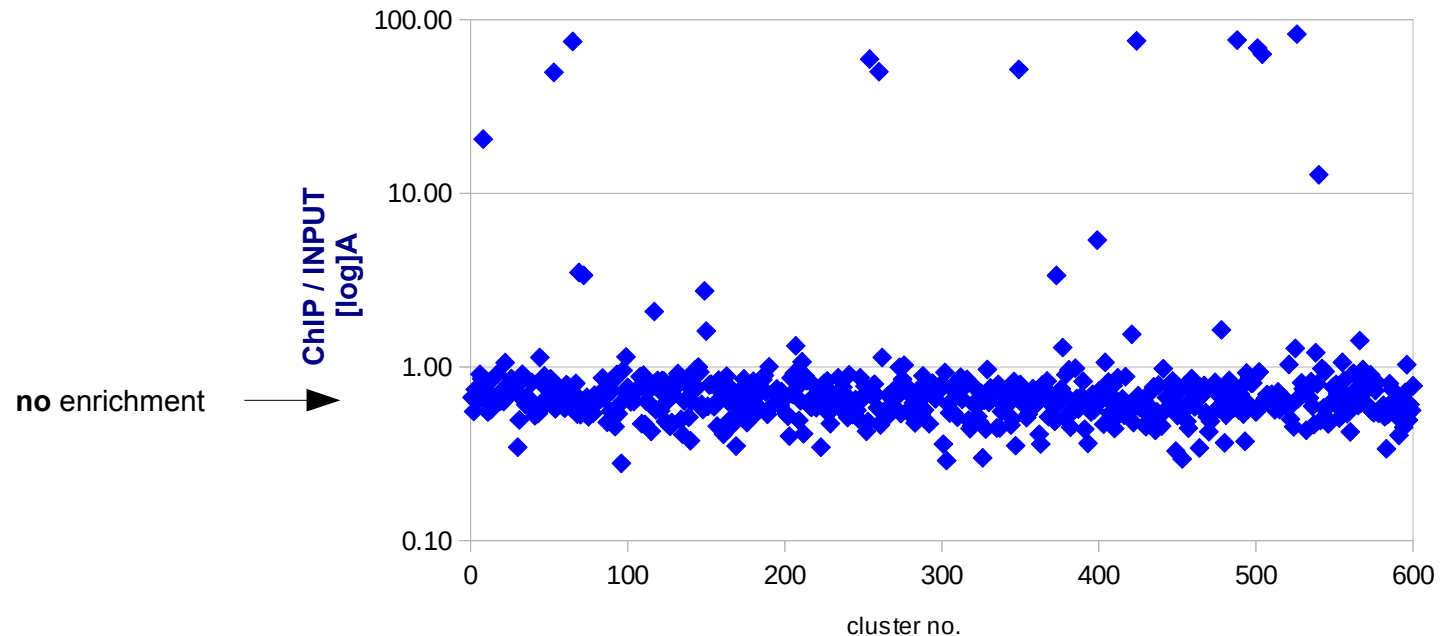
- RNA-seq reads (mRNA, smRNAs,...)
 - Detection of transcribed repeats
 - Comparative analysis (tissues,...)
- ChIP-seq reads
 - Association of repeats with specific types of chromatin

Identification of centromeric repeats by ChIP-seq

Outline of the experiment

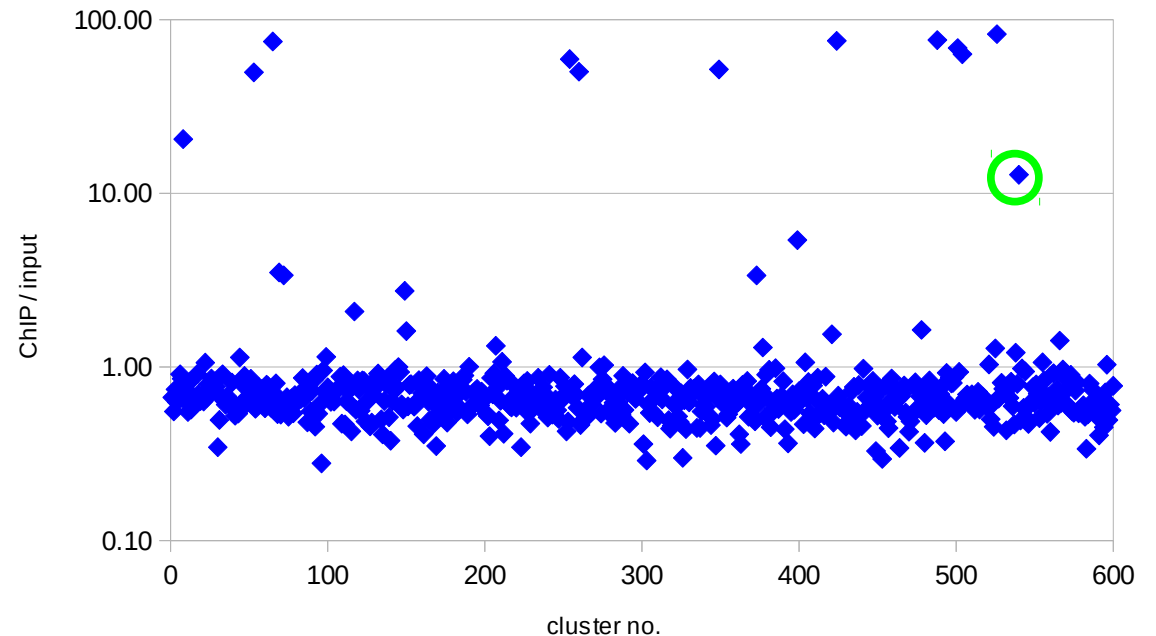
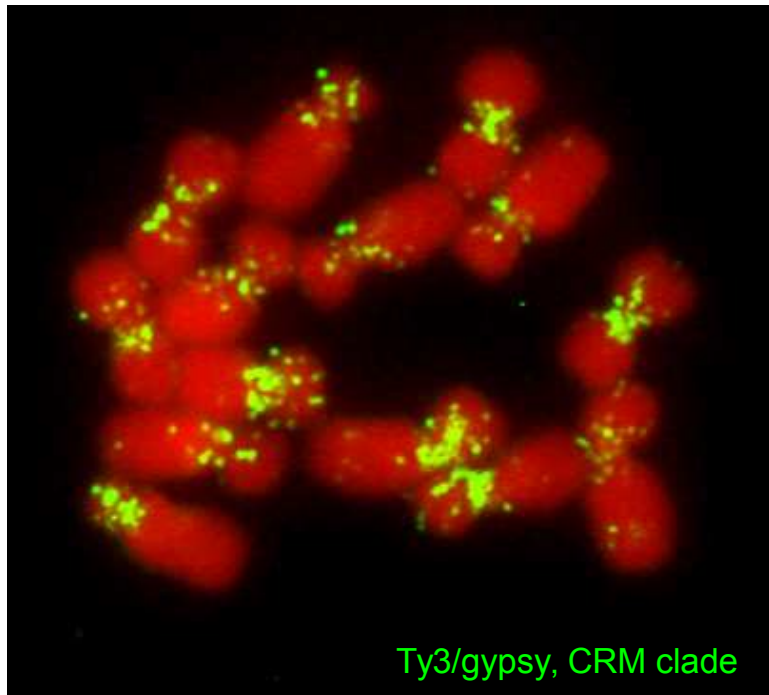
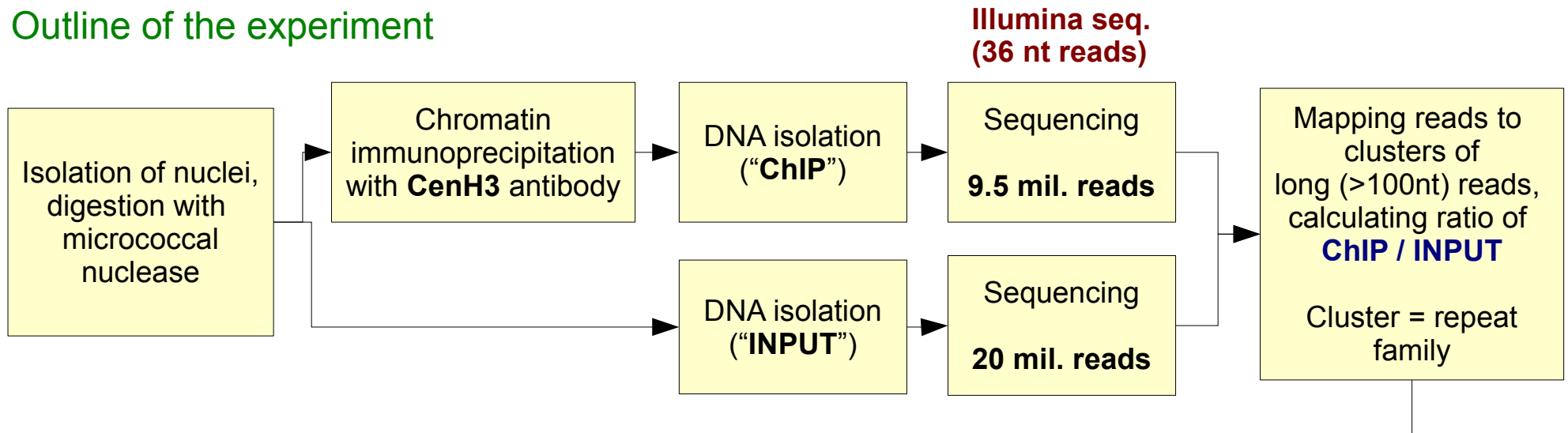


100-fold enrichment



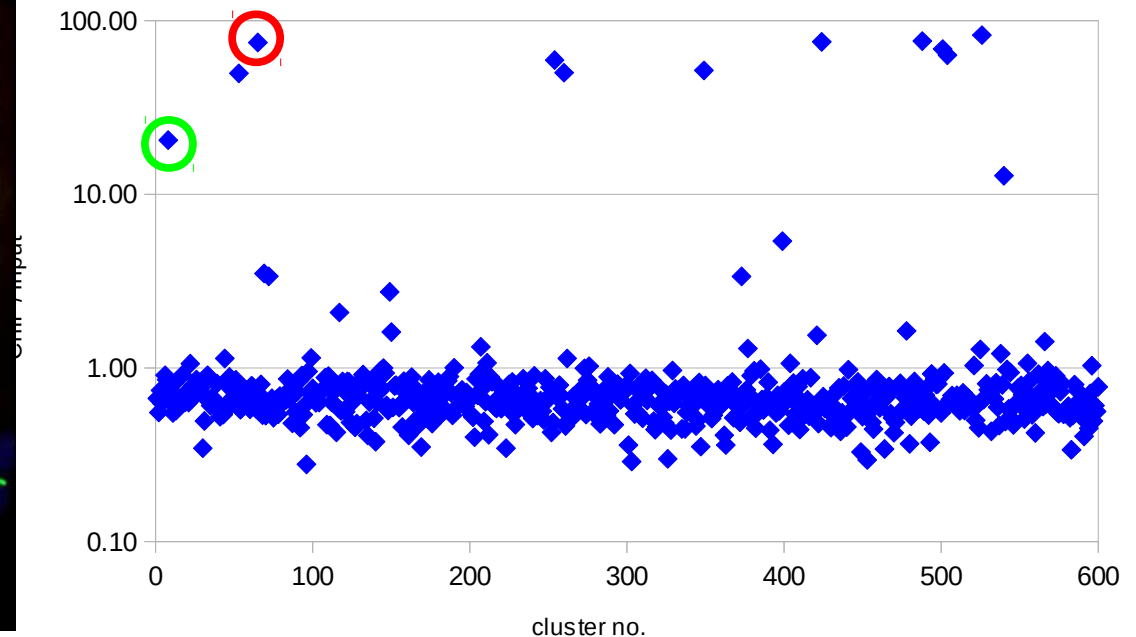
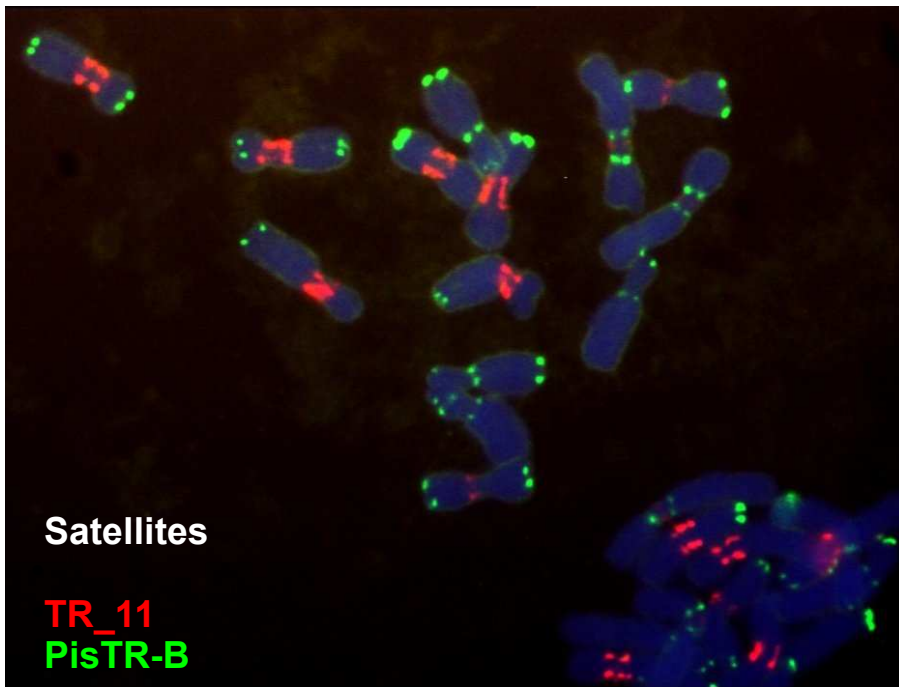
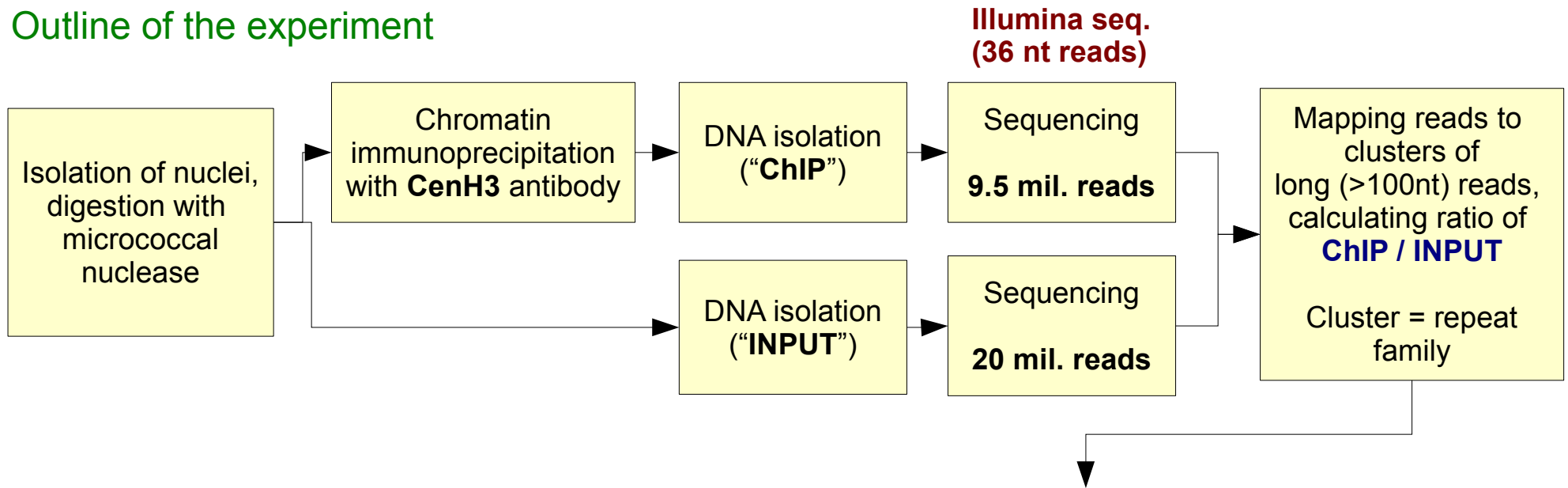
Identification of centromeric repeats by ChIP-seq

Outline of the experiment



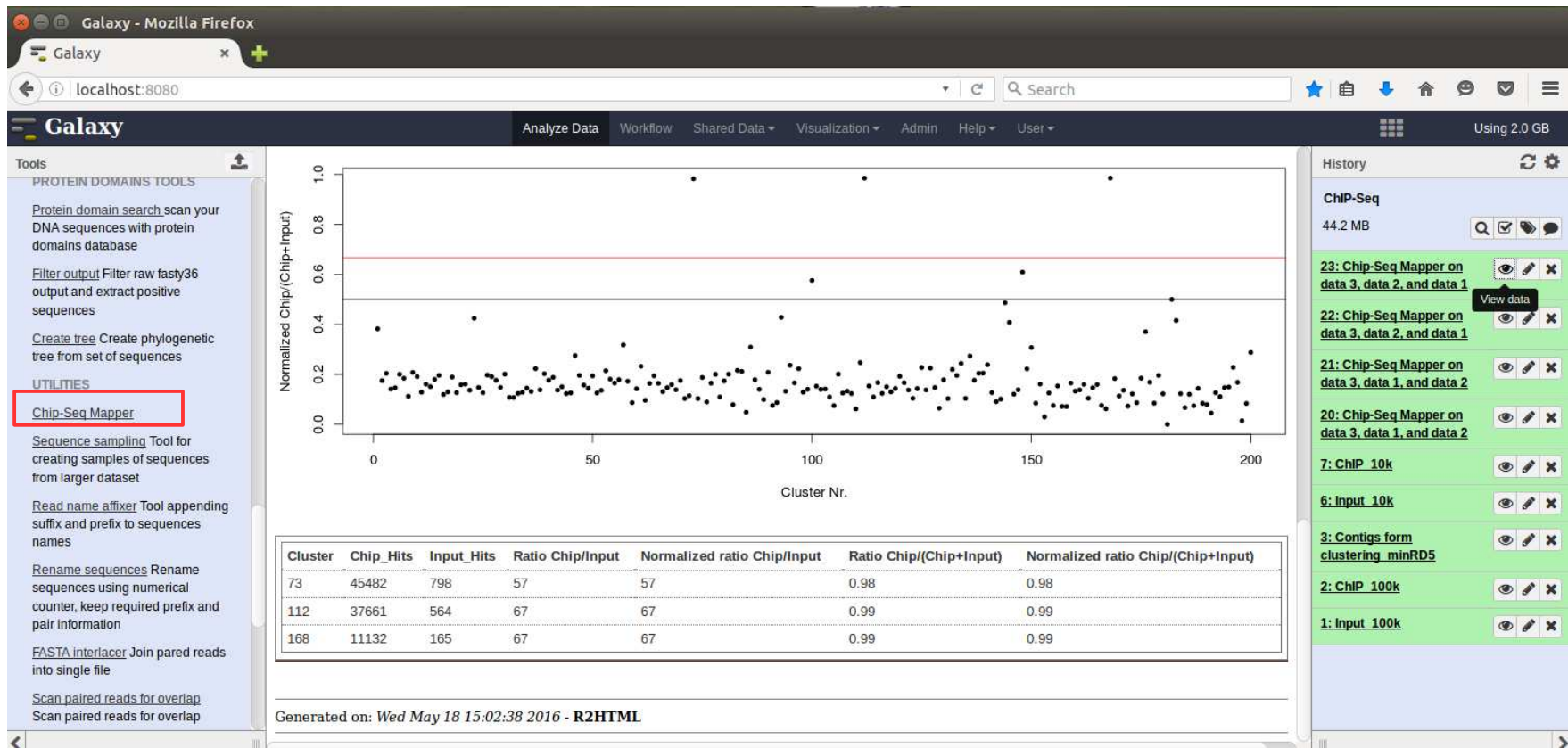
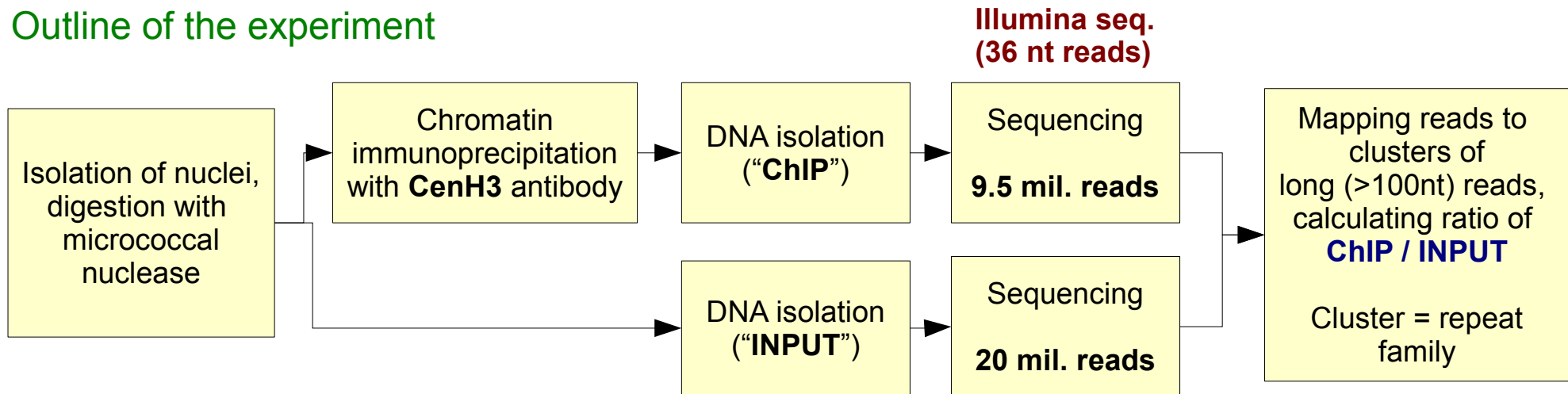
Identification of centromeric repeats by ChIP-seq

Outline of the experiment



ChIP-seq analysis tool in RepeatExplorer

Outline of the experiment



What is new in RepeatExplorer

- ChIP-seq analysis
 - TAREAN - automated identification and monomer reconstruction for satellite repeats
 - *Coming (soon ?) - retrotransposon protein databases covering entire Viridiplantae*
---> *update and official release of automated repeat classification in RE*
-
- Migration of the public RE server to CESNET/CERIT
 - More hardware resources
 - Larger data storage capacity
 - *(now in testing mode)*

