8th RepeatExplorer Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis

Welcome !

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

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 Hoštáková
Ludmila Oliveira
Renata Marecová
Andrea Koblížková
Vlasta Tetourová
Jana Látalová
Tihana Vondrak

Programme

Tuesday (May 21)

9:15 - **Principles and applications of graph-based repeat clustering** (J. Macas)

10:00 - **RepeatExplorer pipeline** (P. Novák)

10:30 - 11:00 *Coffee break*

11:00 - **REXdb database and transposon classification using conserved protein domains** (P. Neumann)

11:30 - **Using RepeatExplorer output for repeat annotation and quantification** (J. Macas)

11:50 - **Additional RE tools** (P. Novák)

12:30 - 13:30 *Lunch*

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

- design of sequencing and repeat analysis experiments
- introduction to Galaxy environment
- quality control and pre-processing of NGS reads, dealing with various read formats
- setting up clustering analysis
- comparative clustering of multiple samples

19:00 → : Dinner at "CITYgastro"

Wednesday (May 22)

8:30 -12:30: **Short talks**

- Steven Dodsworth
- Lucia Campos-Dominguez
- Santelmo Vasconcelos
- Sidonie Bellot

Coffee break

- Tony Heitkam
- Vildana Suljevic & Hanna Schneeweiss
- Ludwig Mann
- Aleš Kovařík

12:30 - 13:30 *Lunch*

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

- identification of satellite DNA using TAREAN
- understanding RepeatExplorer output
- cluster annotation and repeat composition of the genome
- comparative clustering of multiple species – data interpretation
- repeat quantification (principles, sensitivity and reproducibility)
- design of hybridization probes based on RE

Thursday (May 23)

8:30 -12:30: **Short talks**

- Aretuza Sousa
- Kathrin Seibt
- Vratislav Peška
- Jasna Puizina

Coffee break

- Mariana Baez
- Nicola Schmidt
- Mark Johnston
- Horacio Naveira

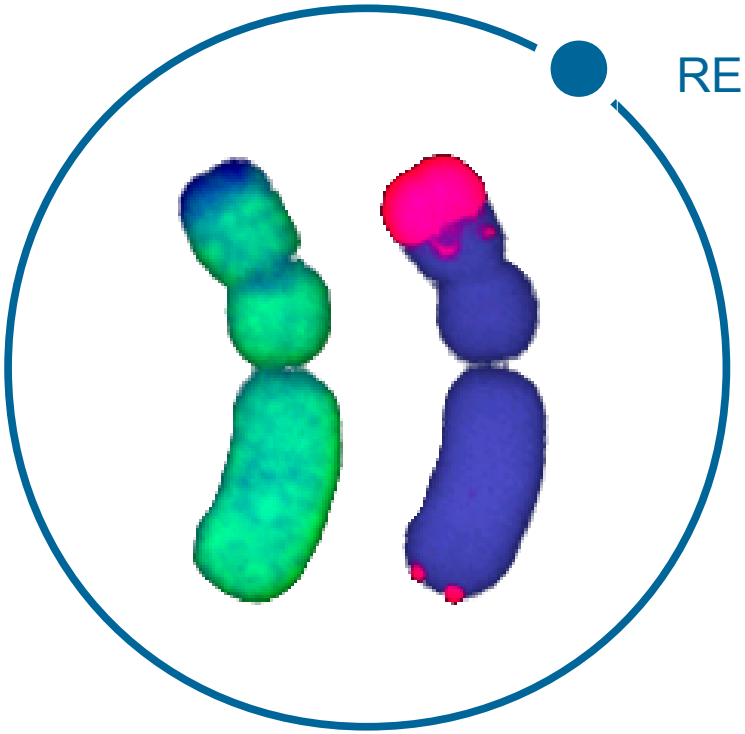
12:30 - 13:30 *Lunch*

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

- combining repeat clustering with ChIP-seq data
- identification and phylogenetic analysis of retrotransposon protein domains
- SeqGrapheR – visualization and annotation of the cluster graphs

- ***advanced topics, troubleshooting***

Principles and applications of graph-based repeat clustering

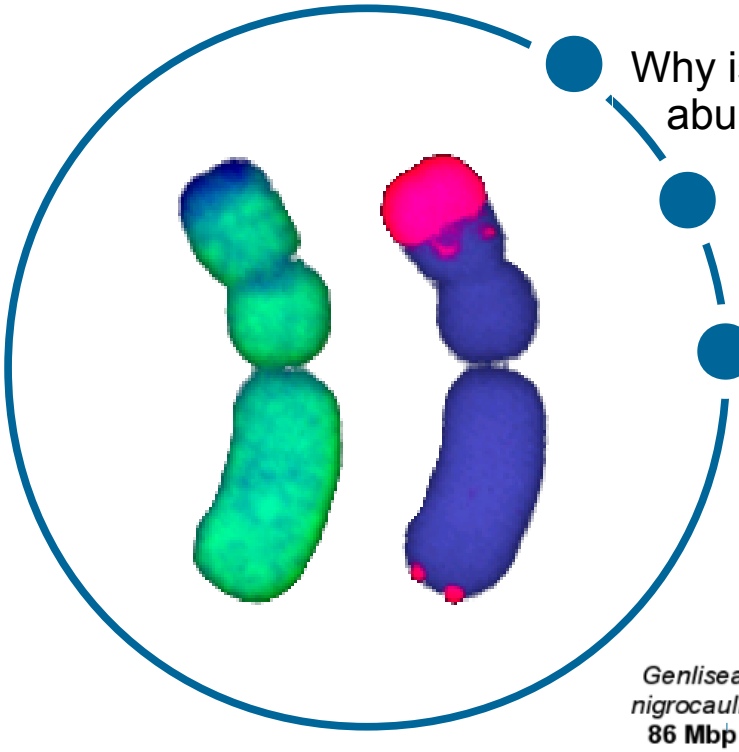


REPEATS IN PLANT GENOMES

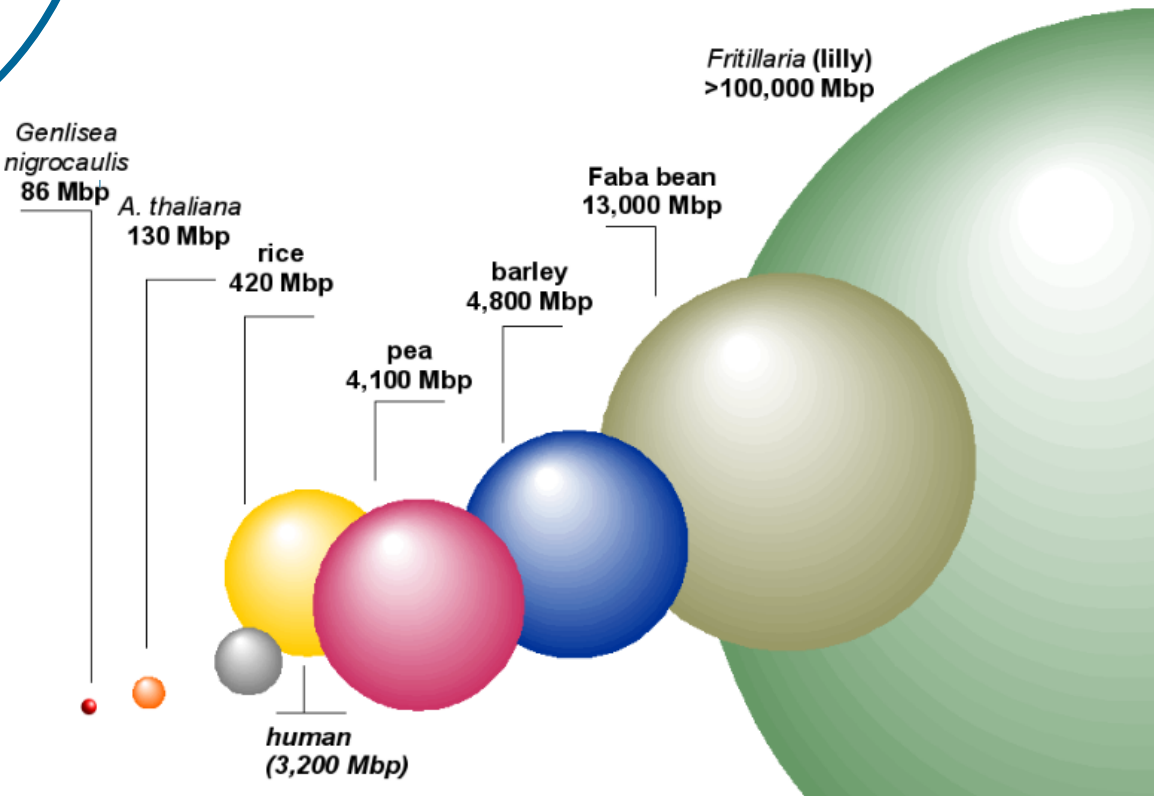
Satellite DNA

Retrotransposon

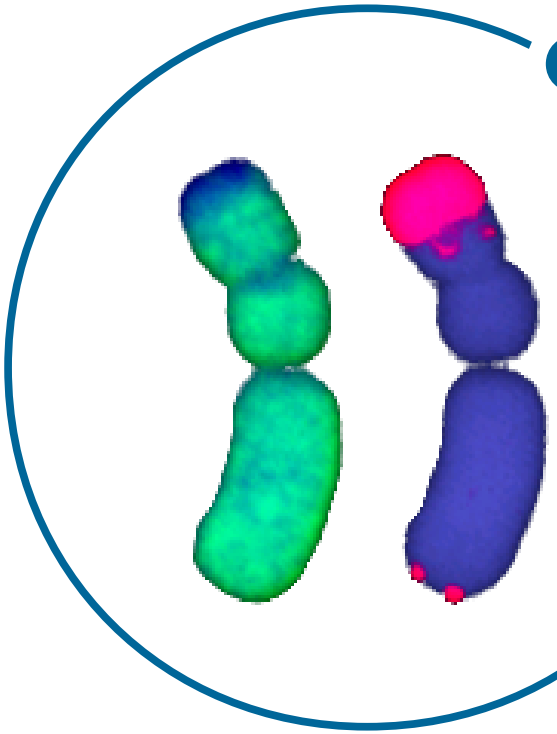
Repeats in plant genomes



- Why is repetitive DNA there and why is it so abundant ?
- What are the evolutionary mechanisms acting on genomic repeats ?
- Why is there differential accumulation of repetitive DNA ?

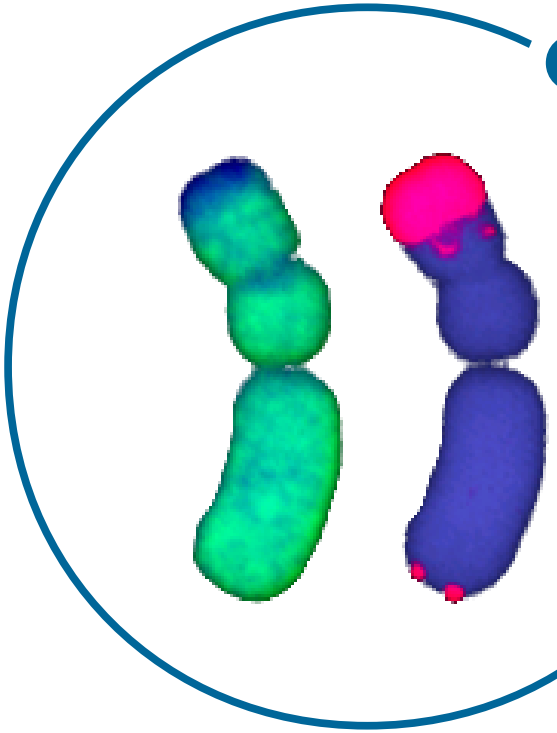


Repeats in plant genomes



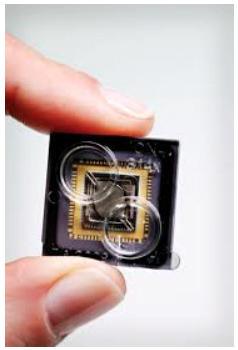
- Why is repetitive DNA there and why is it so abundant ?
- What are the evolutionary mechanisms acting on genomic repeats ?
- Why is there differential accumulation of repetitive DNA ?
- What is the contribution of different types and families of repetitive elements ?
- **What is the repeat composition in individual species or taxons ?**

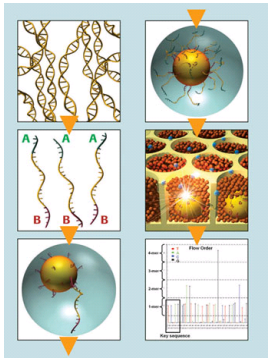
The challenge of repeat identification in complex genomes



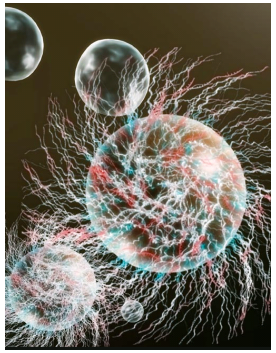
- Why is repetitive DNA there and why is it so abundant ?
- What are the evolutionary mechanisms acting on genomic repeats ?
- Why is there differential accumulation of repetitive DNA ?
- What is the contribution of different types and families of repetitive elements ?
- **What is the repeat composition in individual species or taxons ?**

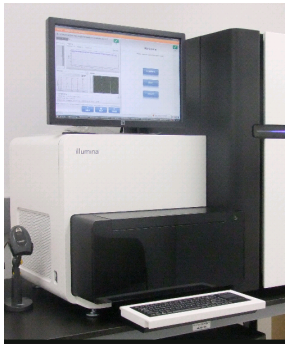
Next generation sequencing: **getting sequence data is no longer a limiting factor**



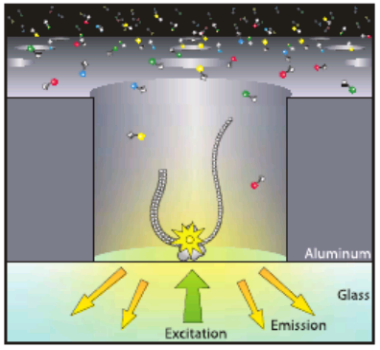


454/Roche

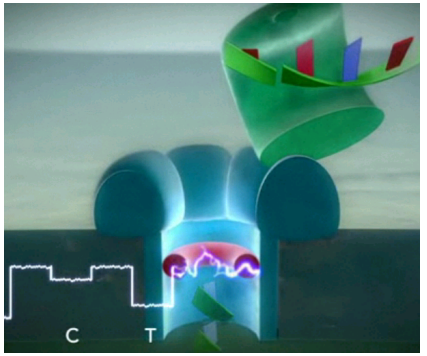




Illumina



Pacific Biosciences



Oxford Nanopore

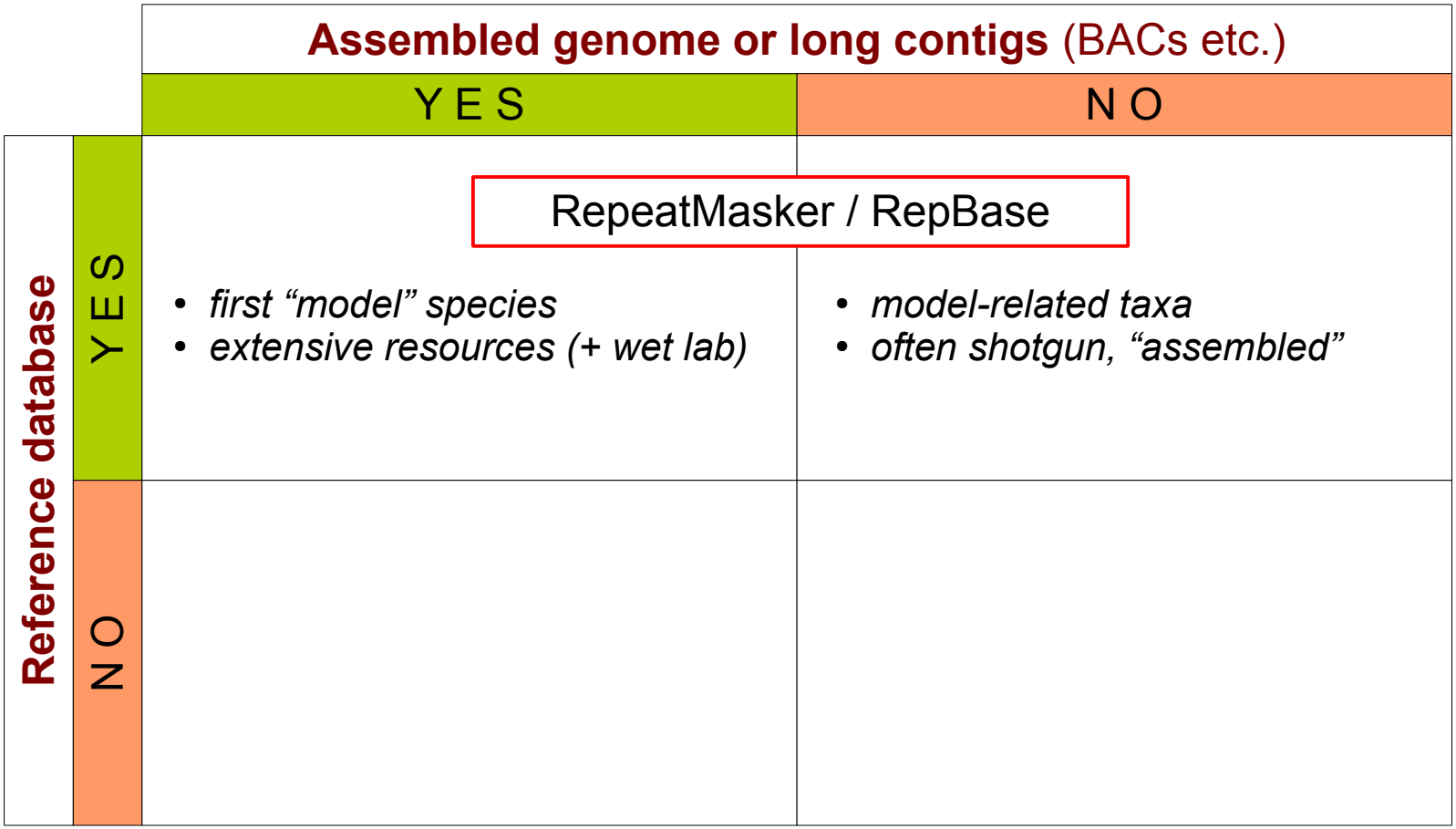
Principles and tools for global repeat identification

		Assembled genome or long contigs (BACs etc.)	
		YES	NO
Reference database	YES		
	NO		

Repeat identification <u>in sequence data</u>

- 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



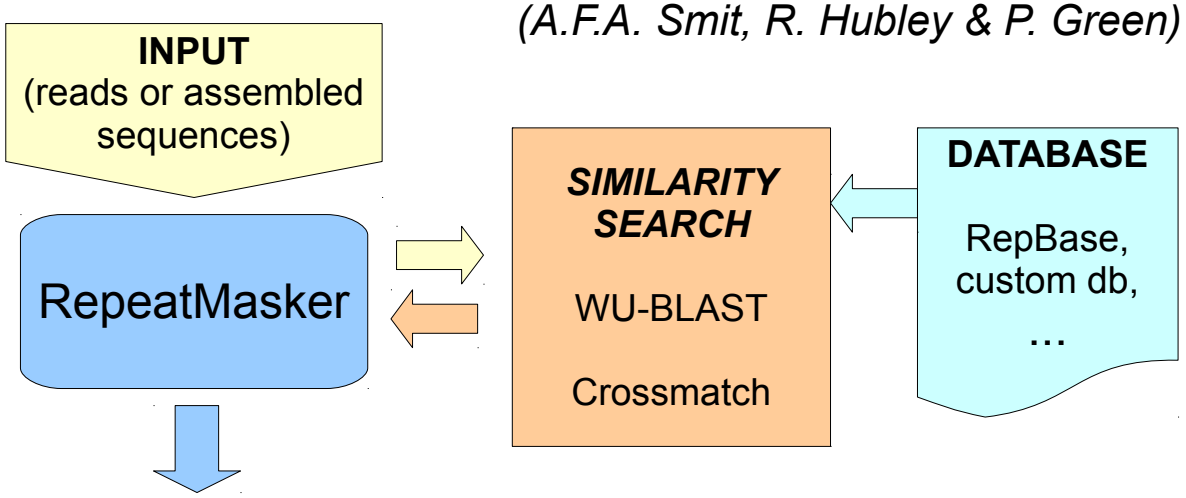
Repeat identification in sequence data

- 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 in query begin	position in query end	(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

< masked sequence

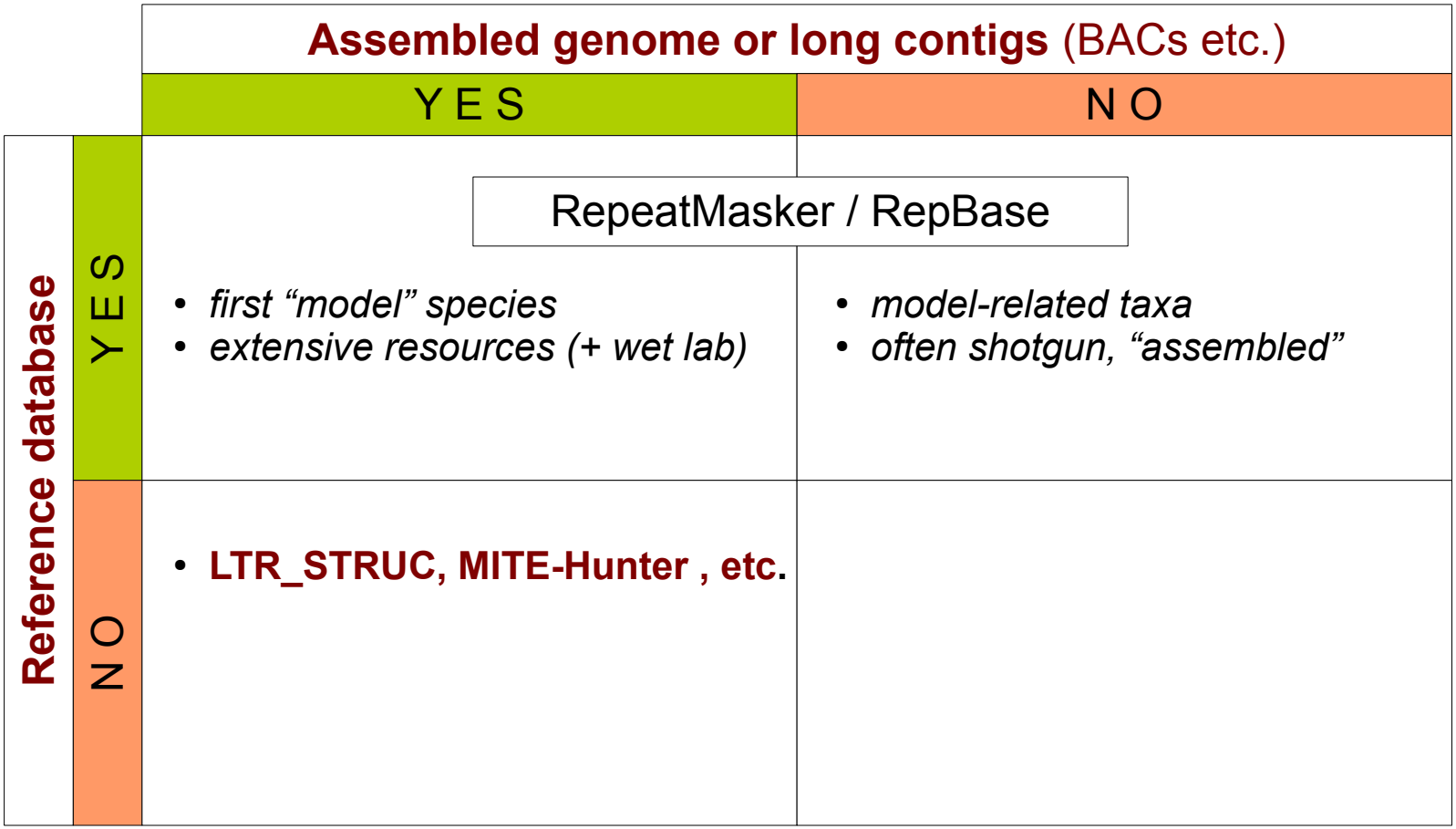
```
>seq_1
NNNNNNNATGAATGCCTGTCCCCGTCAGGGTTGTTGAGTTTTTCCCTAGC
AAATCTCCCGGGCAGAGGTTTGTGTGGATGGTTTTCCCCAGCAGGNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNCCGATCGTCCGTTGACGTCGGGT
TGATCAGATTTTCCCCCAGAGTCACCTCTCCGTTGGTGTCGATAAACGAT
GAGTTTTTCCTATGCGTCCGTTGACGTATAGCTGCATGTTCCCCAAAGAT
CATTTGTTAATGC
```

summary > table

```
=========================================================
                 number of      length   percentage
                 elements*    occupied  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/LOA/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 %
   Tc1-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,        0            0 bp     0.00 %
    P-element, Transib)
```

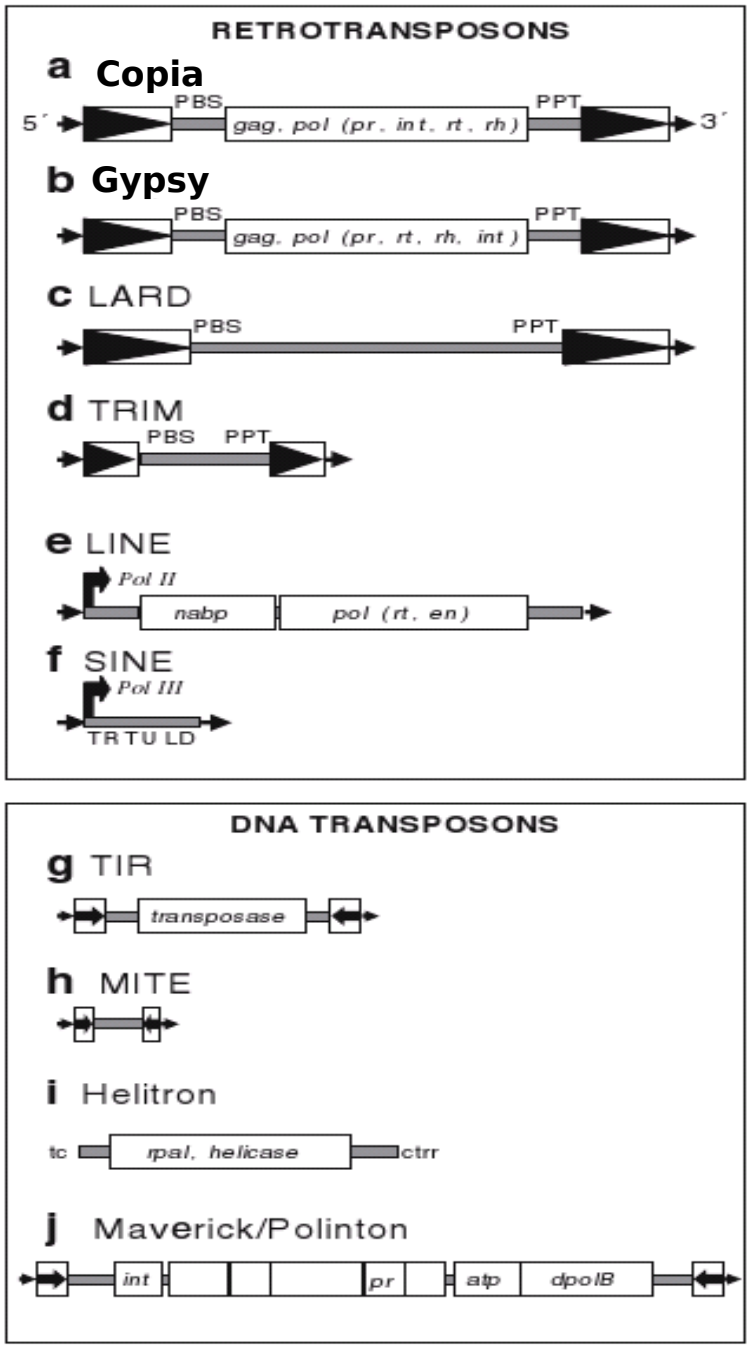
Principles and tools for global repeat identification



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



Saha et al. (20008)

Signature / structure-based approaches

LTR_STRUC *(McCarthy and McDonald, 2003)*

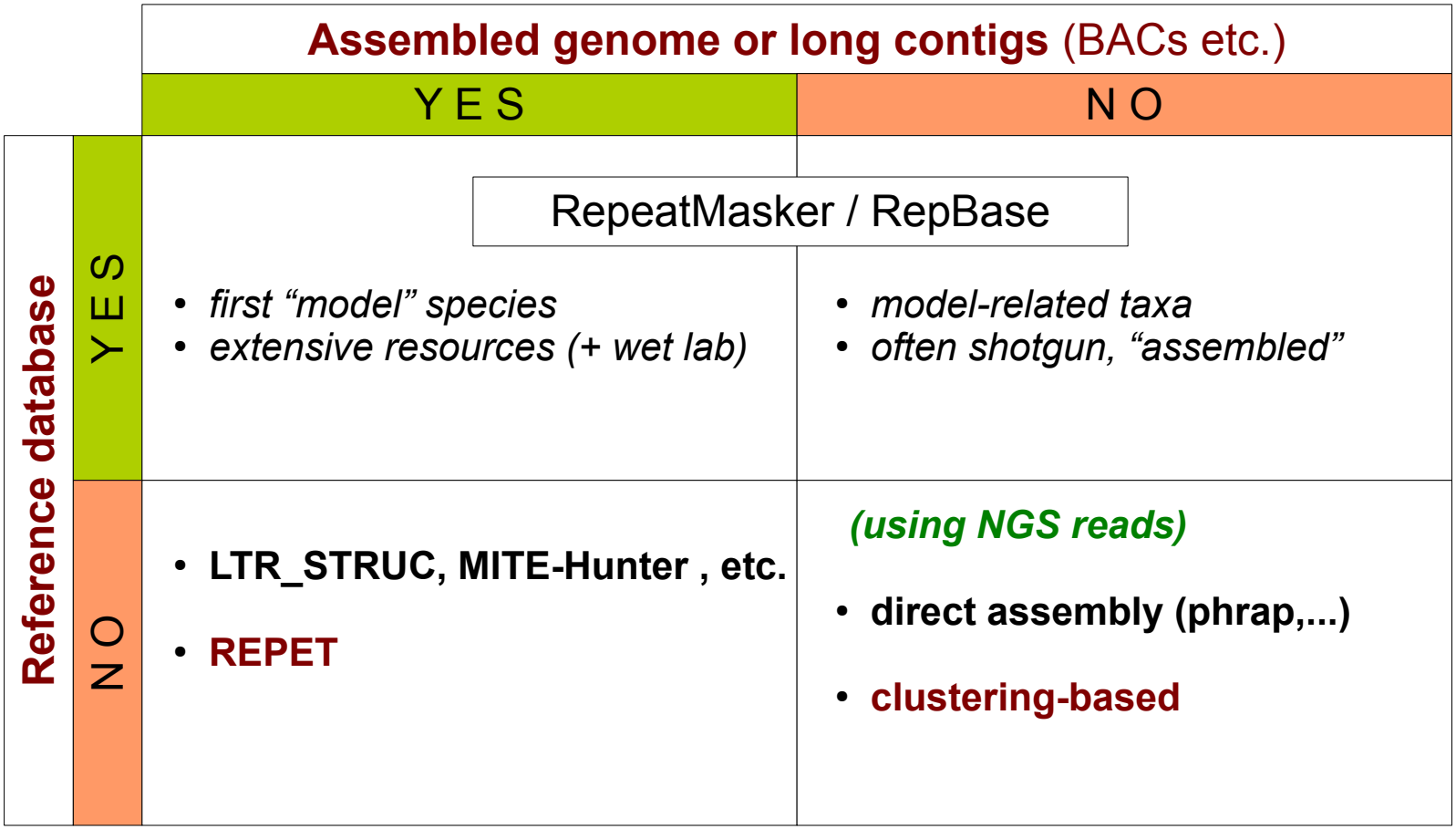
LTR_FINDER *(Xu and Wang, 2007)*

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]

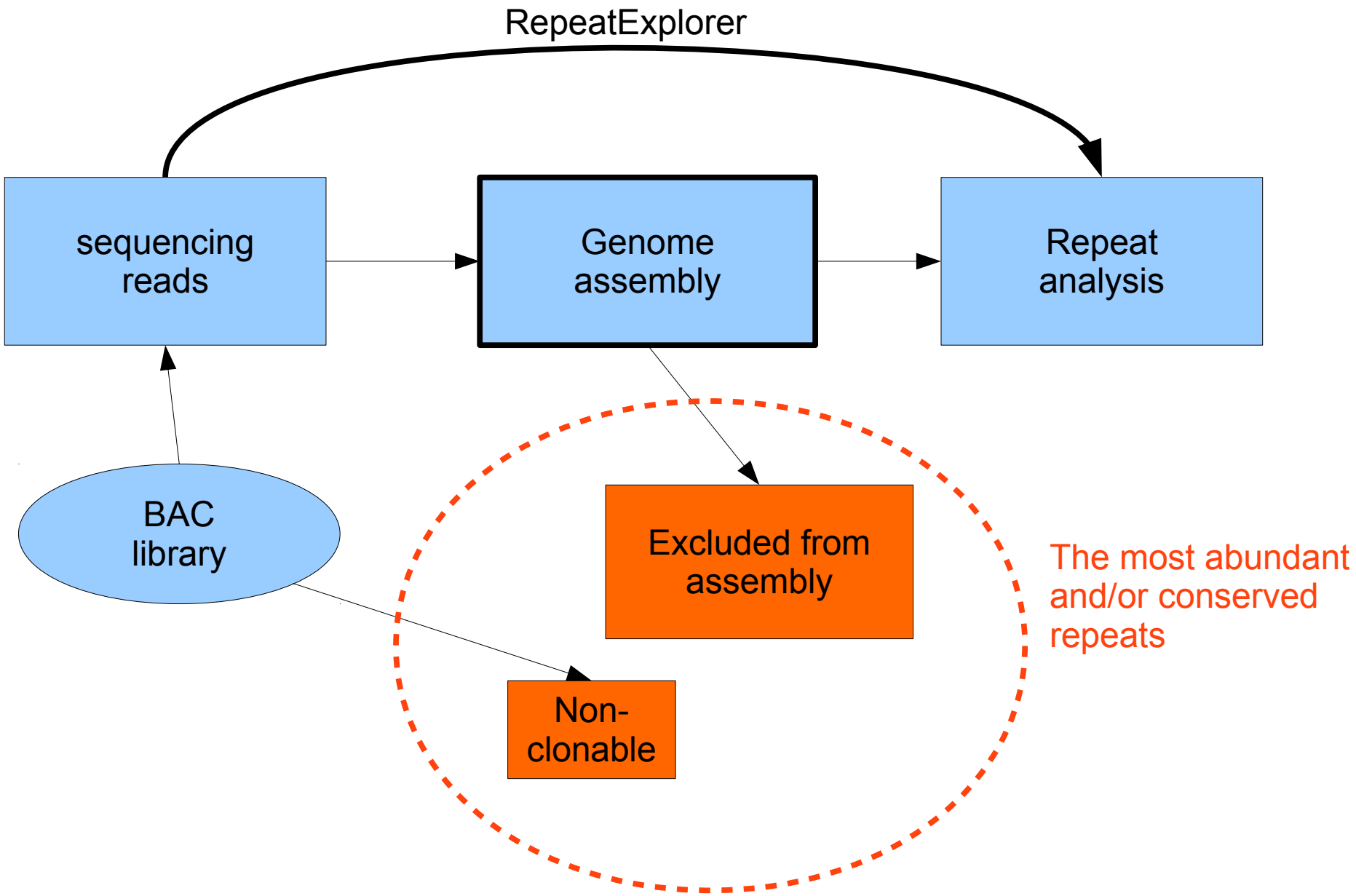
Principles and tools for global repeat identification



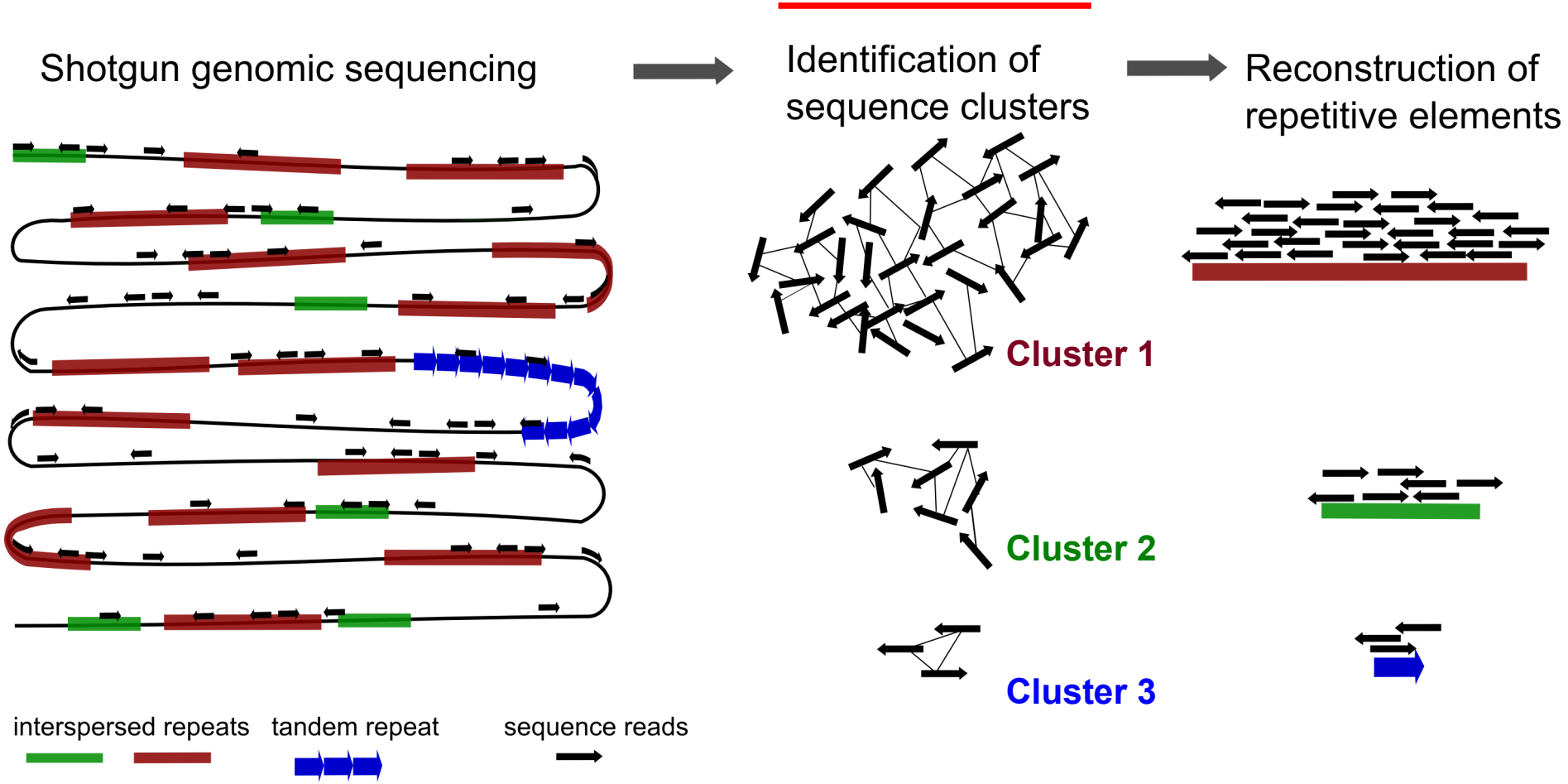
Repeat identification in sequence data

- 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

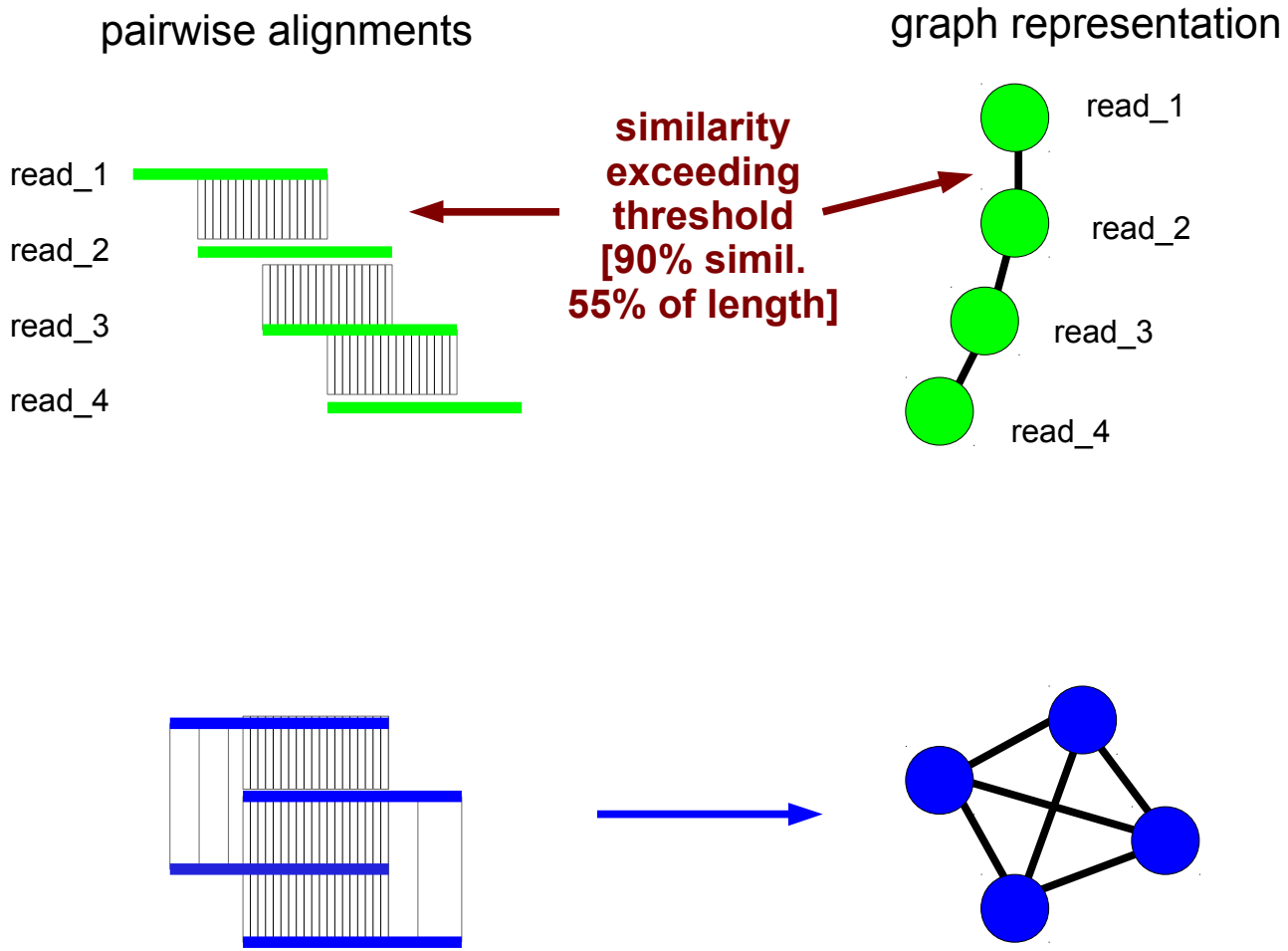


Clustering-based repeat identification

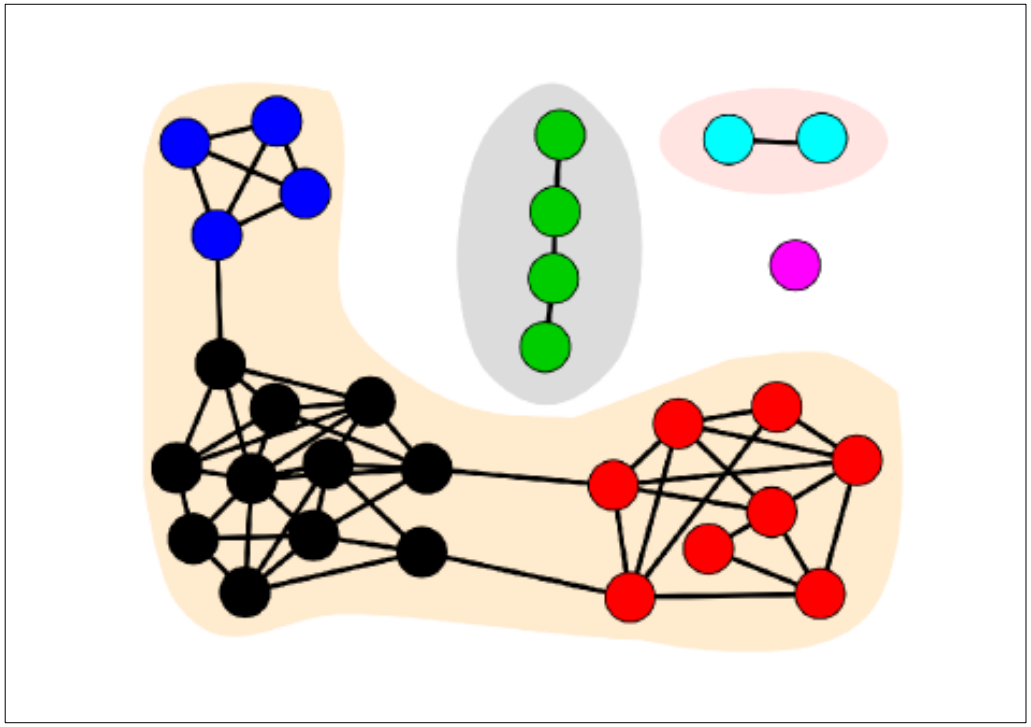


CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

Clustering-based repeat identification



Clustering-based repeat identification

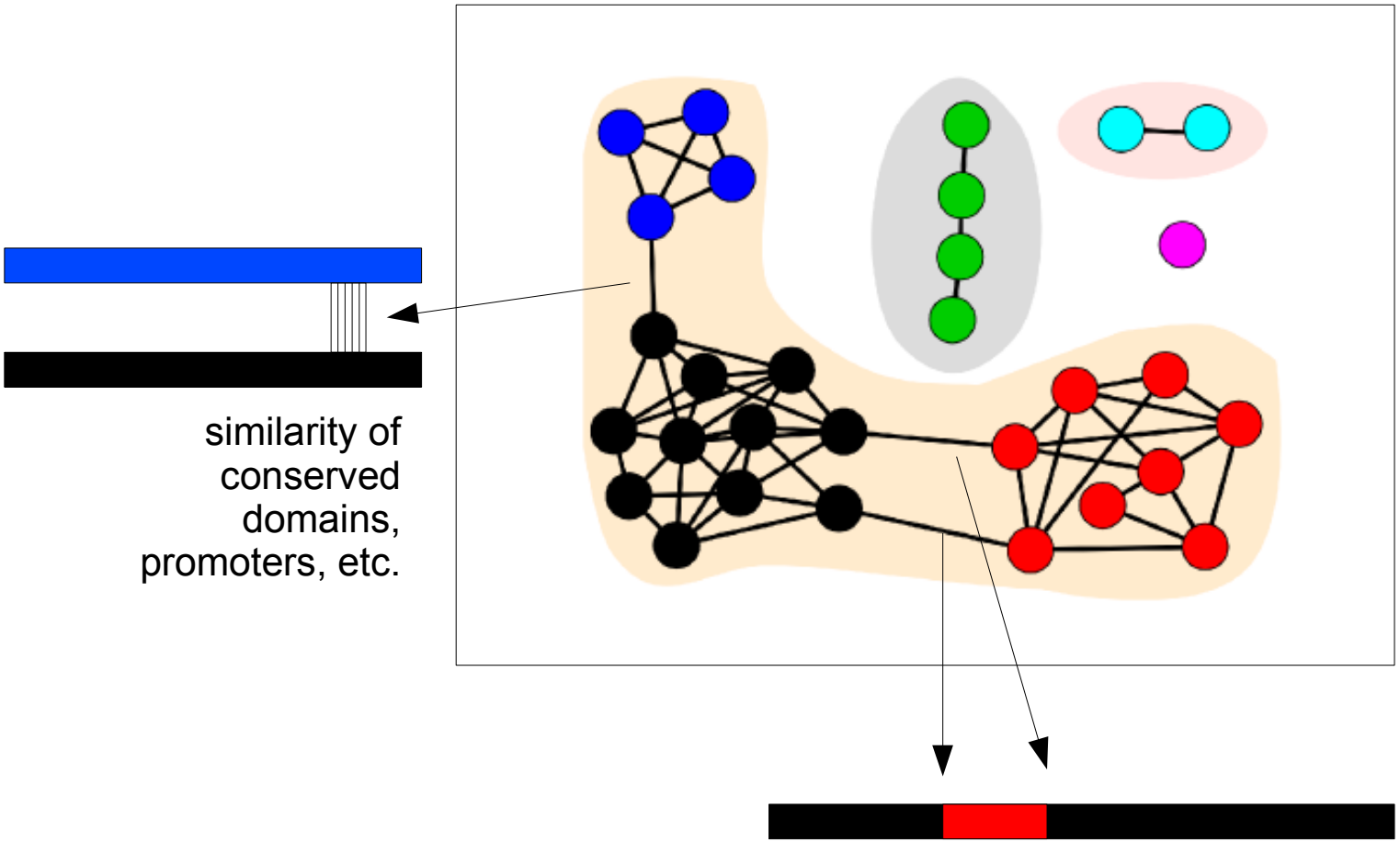


Single linkage clustering => *connected components*

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*.

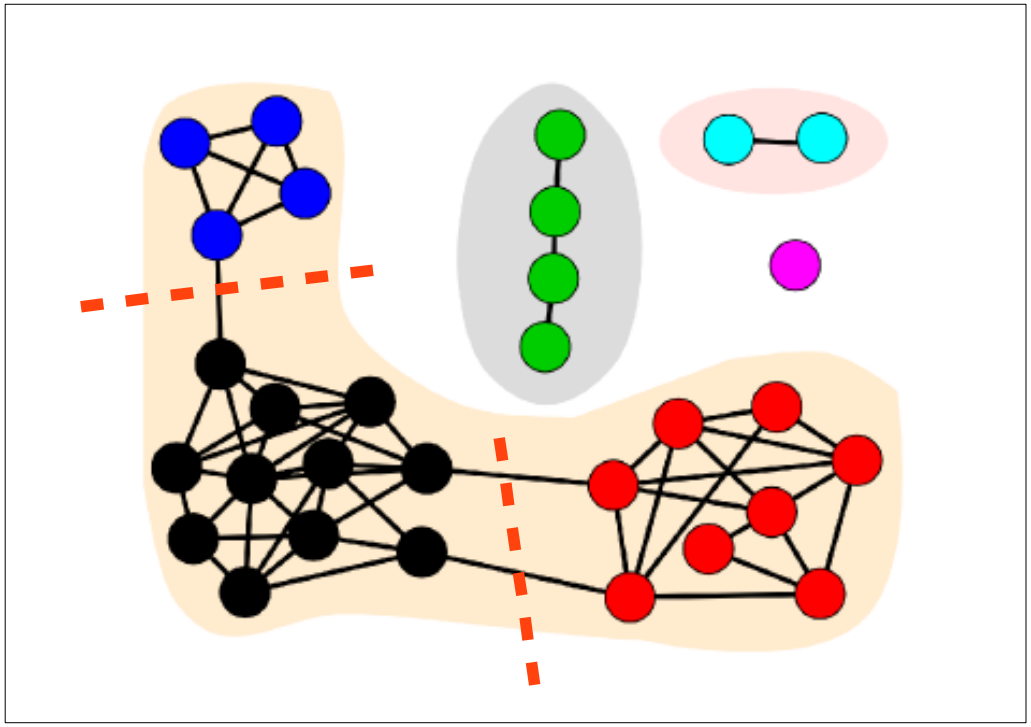
Clustering-based repeat identification



Single linkage clustering => *connected components*

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

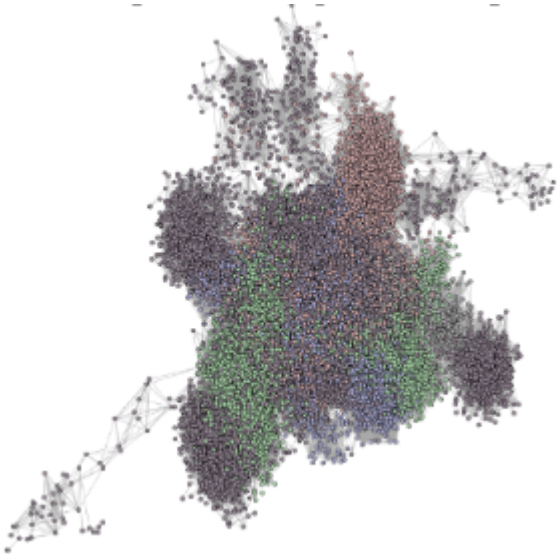
Graph-based clustering



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

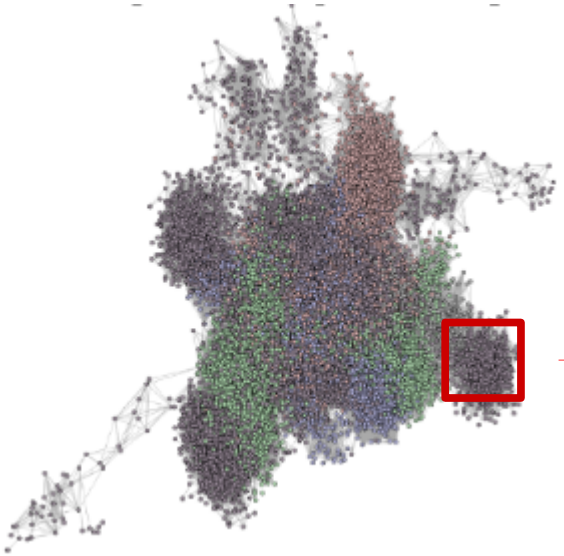


Virtual graphs used to analyze real data contain **up to millions of nodes** (reads)

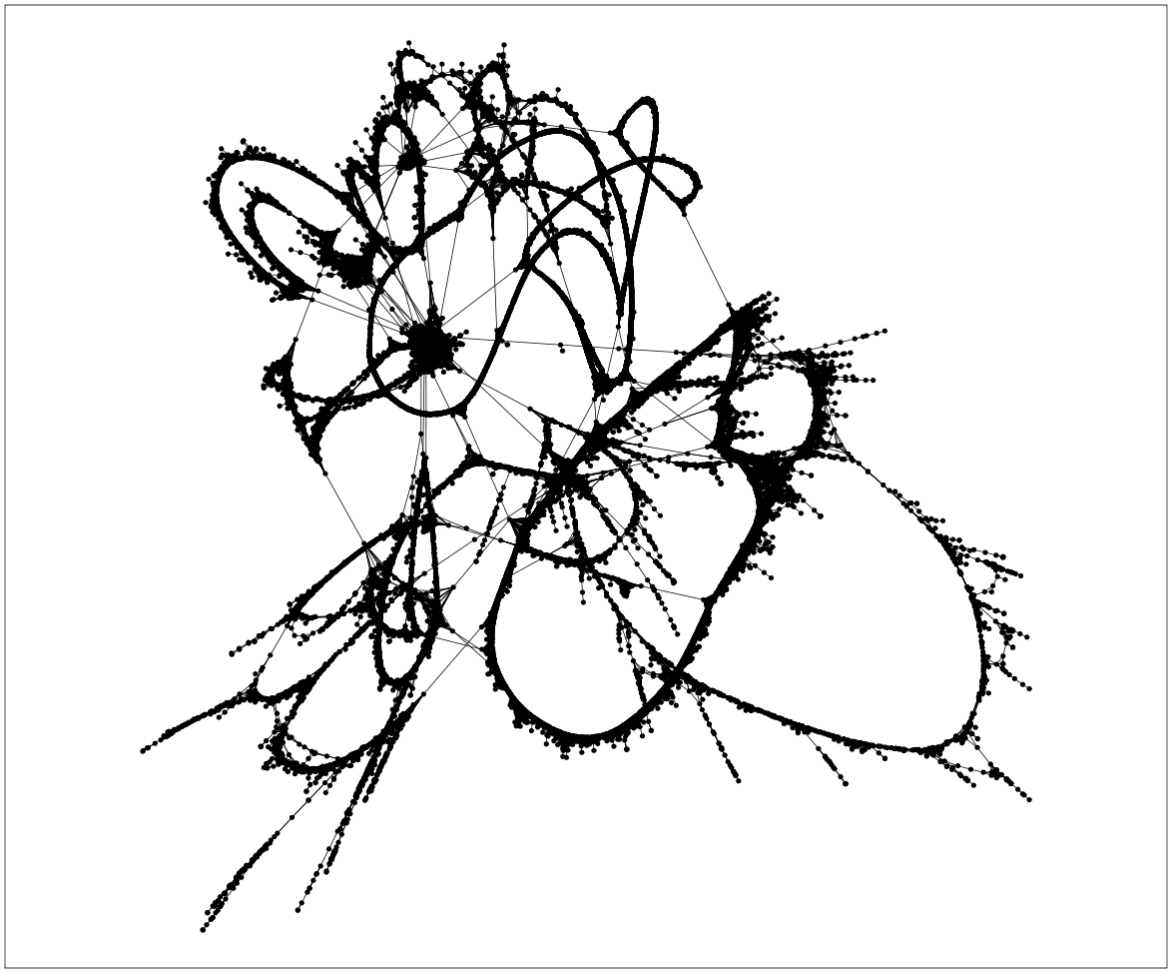
Graph-based clustering

- 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



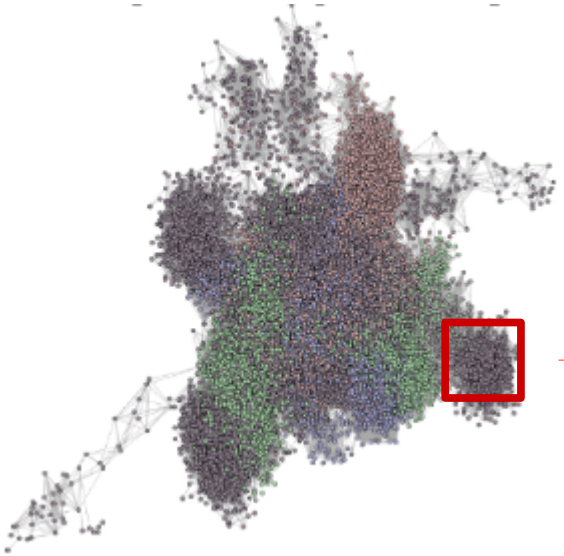
Virtual graphs used to analyze real data contain **up to millions of nodes** (reads)



Graph-based clustering

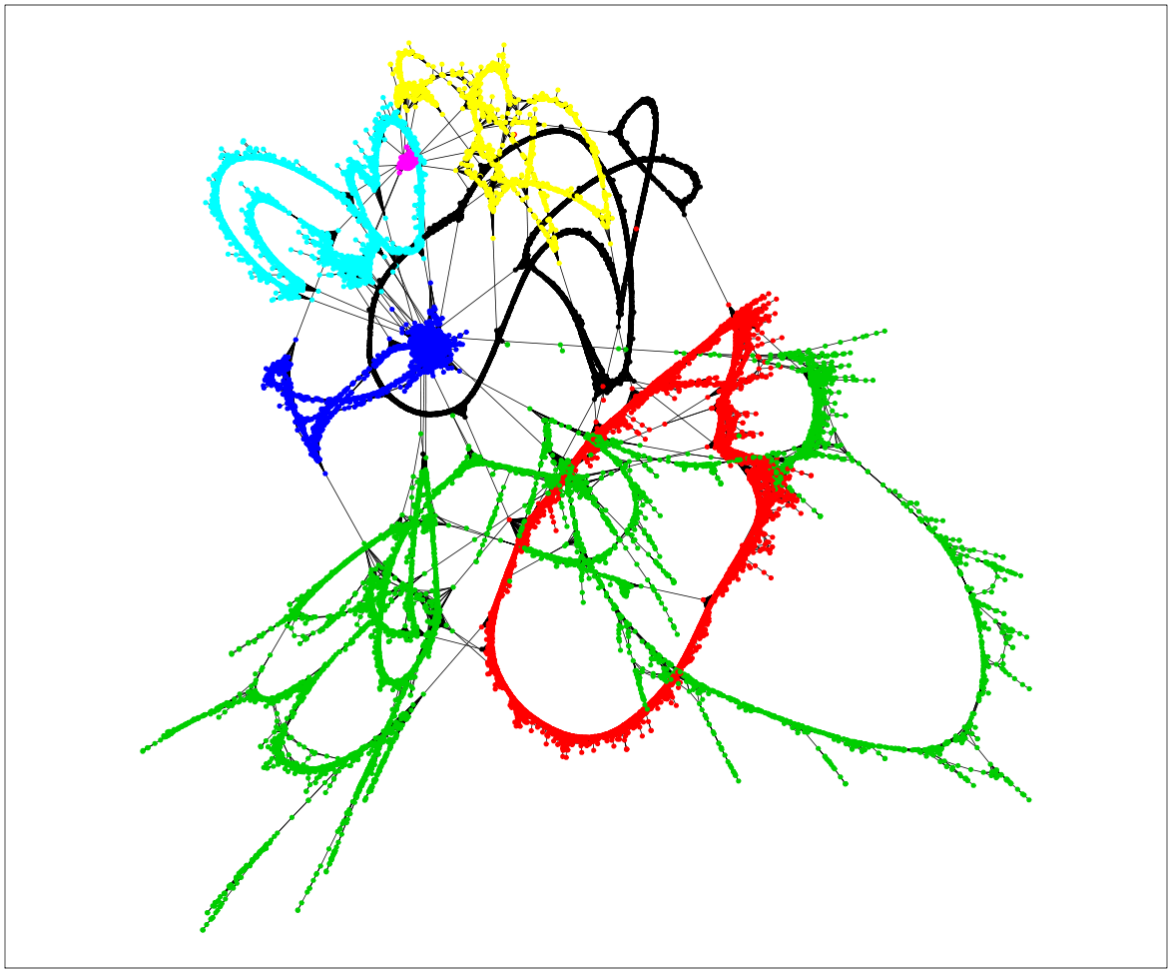
- Sequence overlaps between the reads are transformed to a graph where the **reads** are represented as **nodes** and their **similarities** as **edges** connecting the nodes
- Graph structure is examined to detect ***communities of frequently connected nodes*** which are split to separate clusters

Graph-based characterization of repeat clusters



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

Identified clusters



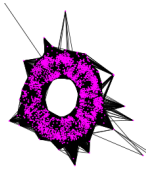
Graph-based clustering

- 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

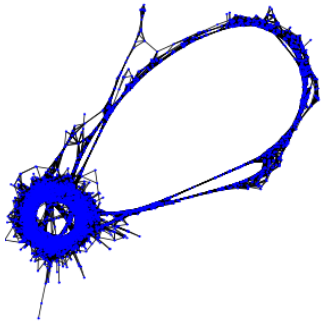
Graph-based characterization of repeat clusters

Extraction of of individual clusters (**graph shapes indicate repeat types**)

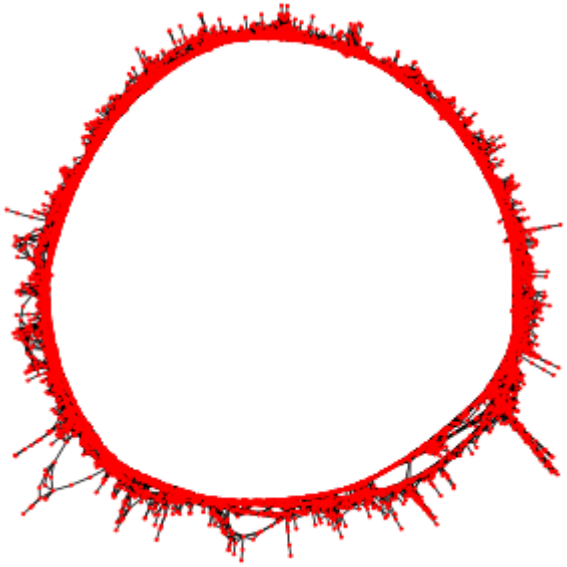
Identified clusters



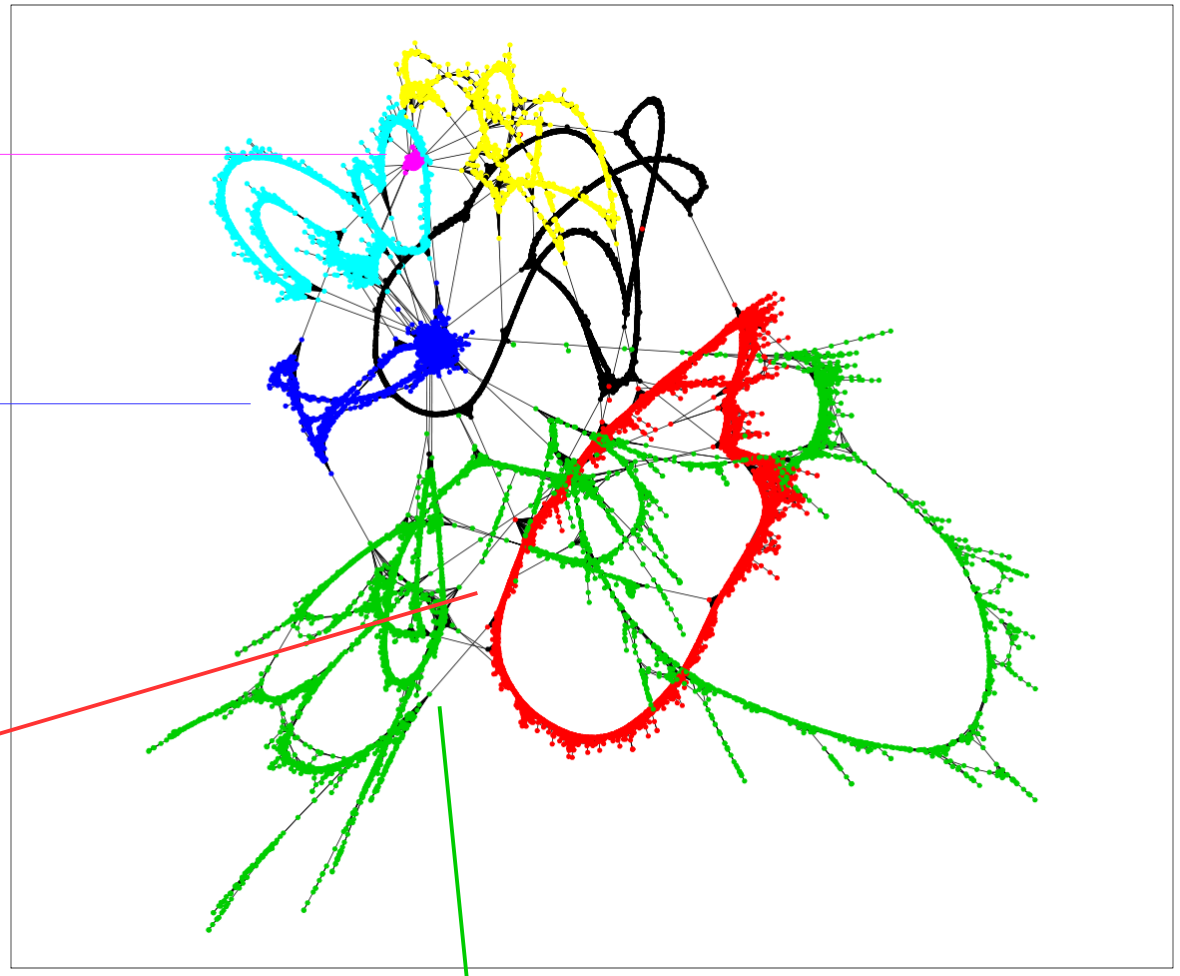
Satellite (simple)

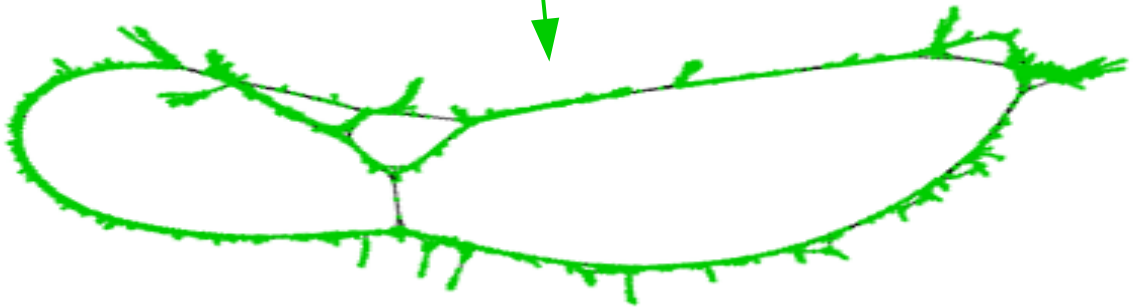


Satellite family (short + long monomers) >

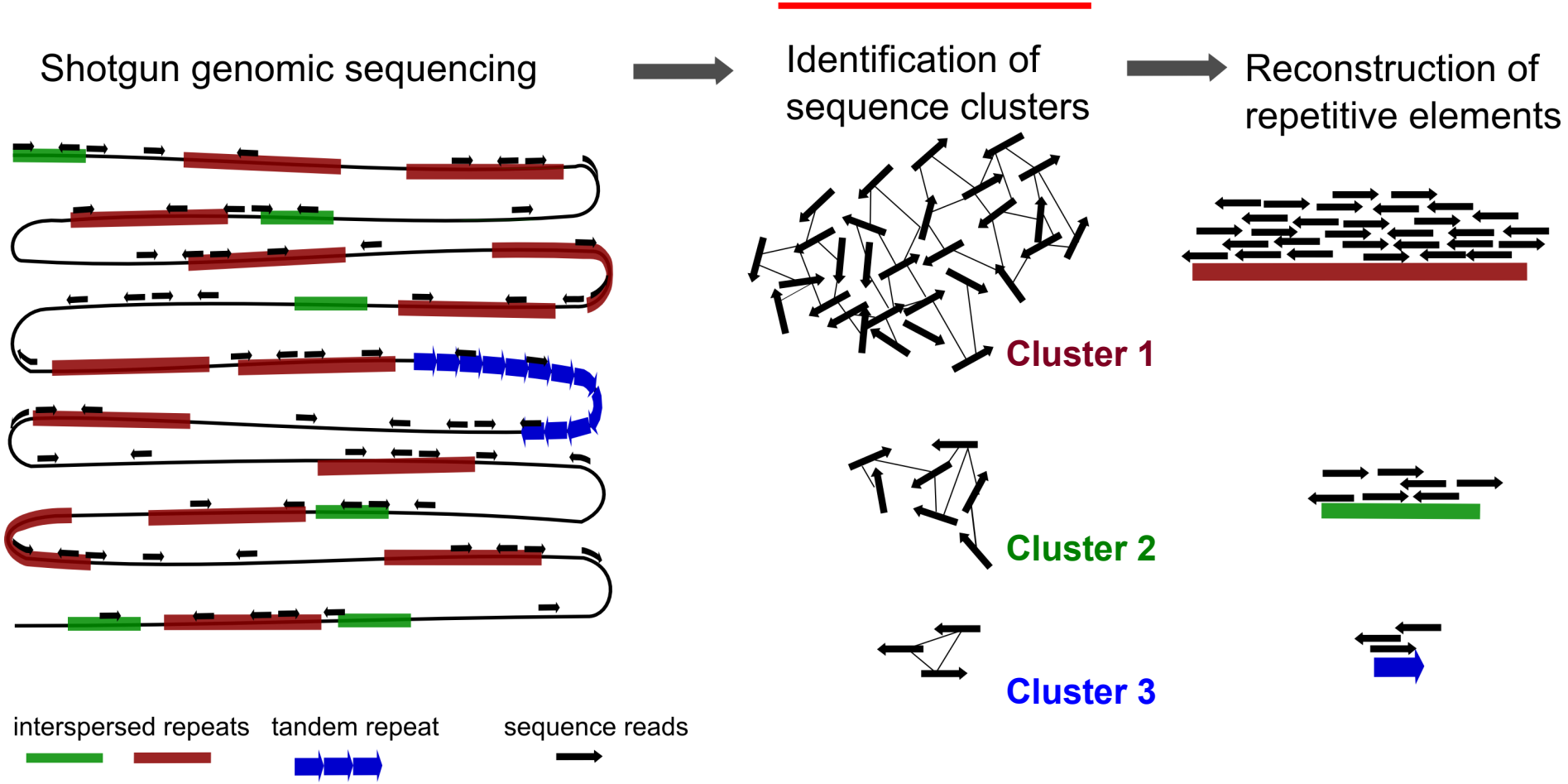


LTR-retrotransposons





Clustering-based repeat identification



CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

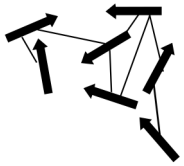
Repeat characterization

Cluster annotation and quantification

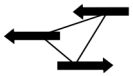
	7192836 (= 100%)	64.1			
CL	reads	genome %	class	type	note
1	304159	4.229	gypsy	Tat	PROT-RT/RH-INT
2	234749	3.264			?, PE->24,18
3	216307	3.007	gypsy	chromo	ALL domains
4	202822	2.820	copia	Maximus	ALL domains
5	149693	2.081	gypsy	Athila	ALL domains
6	145911	2.029	gypsy	Tat	ALL domains
7	143766	1.999	gypsy	chromo	ALL domains
8	142608	1.983	copia	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	copia	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	copia	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			?

Identification of sequence clusters → Reconstruction of repetitive elements

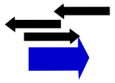
Cluster 1



Cluster 2



Cluster 3



...

...

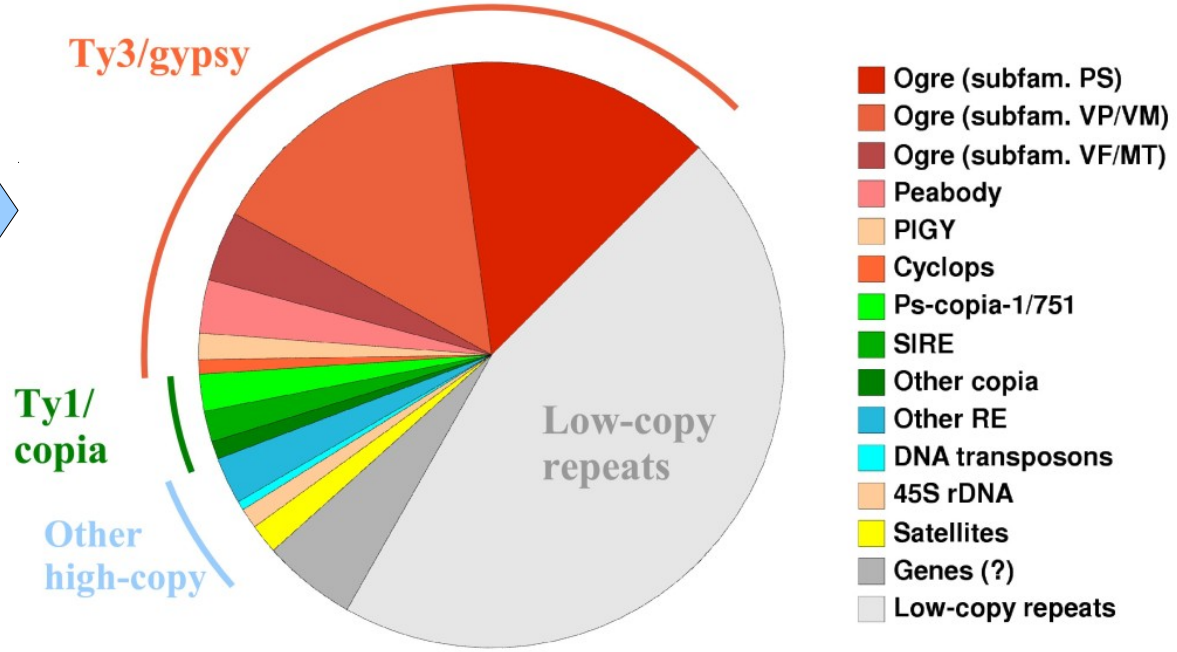
(Cluster x)

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

- *What RepeatExplorer can do for you*:
  - Identify all sequences with certain number of repetitions
  - Repeat quantification
  - Provide models of repeat popoulations (sequence variability)
  - Help in repeat classification/annotation (in plant genomes, *Metazoa in prep.*)

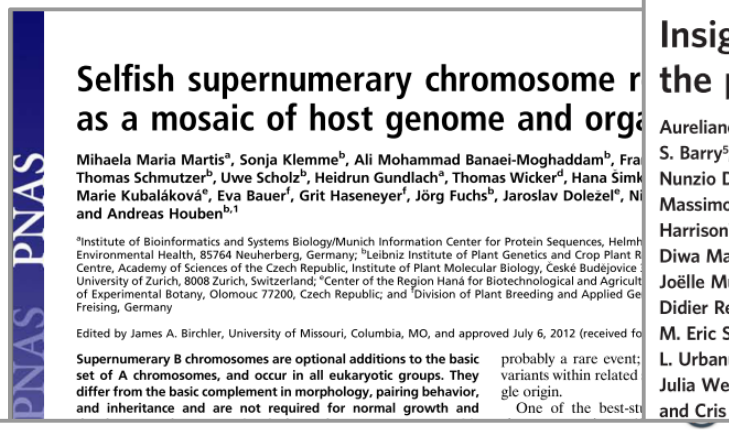
- *What it cannot do*:
  - Genome assembly or reconstruction of individual repeat copies
  - Analyze repeats using assembled genomes as input
  - Use long NGS reads (PacBio, Oxford Nanopore) as input (*work in progress*)
  - Identify some tandem repeats with very short monomers or other low-complexity repeats

Think about proper design of your analysis – RE is just a program designed for a specific type of input (e.g. it needs WGS data as input and it will not work with BAC clones)

Repetitive DNA characterization using RepeatExplorer

Plants

- Over 100 species characterized so far
- Comparative studies
- Whole genome assembly projects



Selfish supernumerary chromosome r
as a mosaic of host genome and orga

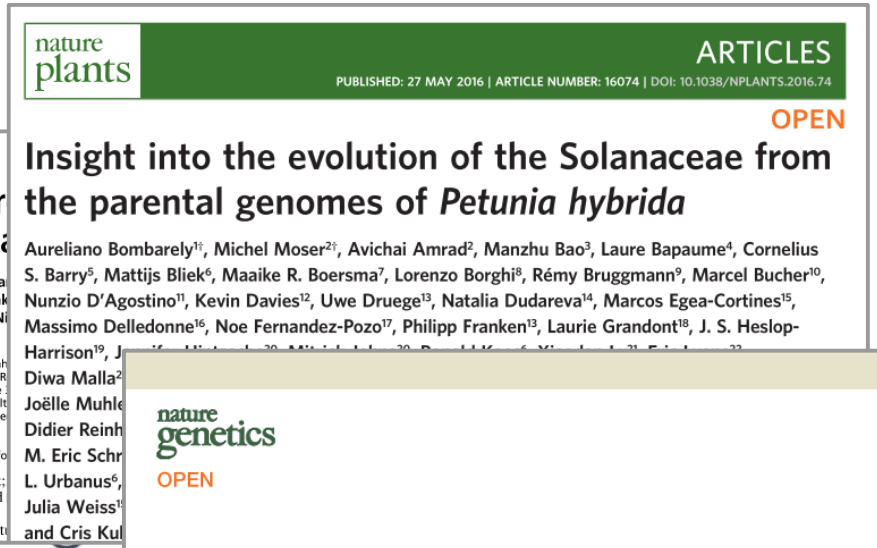
Mihaela Maria Martis[a], Sonja Klemme[b], Ali Mohammad Banaei-Moghaddam[b], Fra
Thomas Schmutzer[b], Uwe Scholz[b], Heidrun Gundlach[a], Thomas Wicker[d], Hana Šimk
Marie Kubaláková[e], Eva Bauer[f], Grit Haseneyer[f], Jörg Fuchs[b], Jaroslav Doležel[e], Ni
and Andreas Houben[b,1]

[a]Institute of Bioinformatics and Systems Biology/Munich Information Center for Protein Sequences, Helmh
Environmental Health, 85764 Neuherberg, Germany; [b]Leibniz Institute of Plant Genetics and Crop Plant R
Centre, Academy of Sciences of the Czech Republic, Institute of Plant Molecular Biology, České Budějovice
University of Zurich, 8008 Zurich, Switzerland; [e]Center of the Region Haná for Biotechnological and Agricult
of Experimental Botany, Olomouc 77200, Czech Republic; and [f]Division of Plant Breeding and Applied Ge
Freising, Germany

Edited by James A. Birchler, University of Missouri, Columbia, MO, and approved July 6, 2012 (received fo

Supernumerary B chromosomes are optional additions to the basic set of A chromosomes, and occur in all eukaryotic groups. They differ from the basic complement in morphology, pairing behavior, and inheritance and are not required for normal growth and

probably a rare event;
variants within related
gle origin.
One of the best-st



nature plants

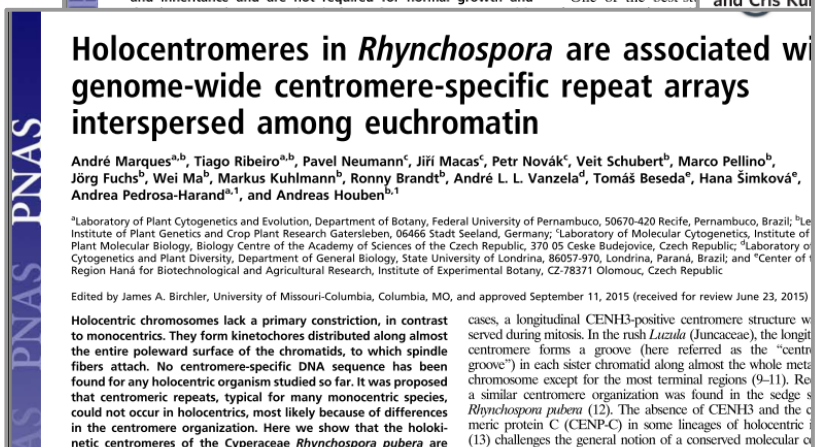
ARTICLES

PUBLISHED: 27 MAY 2016 | ARTICLE NUMBER: 16074 | DOI: 10.1038/NPLANTS.2016.74

OPEN

Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*

Aureliano Bombarely[1†], Michel Moser[2†], Avichai Amrad[2], Manzhu Bao[3], Laure Bapaume[4], Cornelius S. Barry[5], Mattijs Bliek[6], Maaike R. Boersma[7], Lorenzo Borghi[8], Rémy Bruggmann[9], Marcel Bucher[10], Nunzio D'Agostino[11], Kevin Davies[12], Uwe Druege[13], Natalia Dudareva[14], Marcos Egea-Cortines[15], Massimo Delledonne[16], Noe Fernandez-Pozo[17], Philipp Franken[13], Laurie Grandont[18], J. S. Heslop-Harrison[19], J
Diwa Malla[2]
Joëlle Muhle
Didier Reinh
M. Eric Schr
L. Urbanus[6],
Julia Weiss[1]
and Cris Kul



Holocentromeres in *Rhynchospora* are associated wi
genome-wide centromere-specific repeat arrays
interspersed among euchromatin

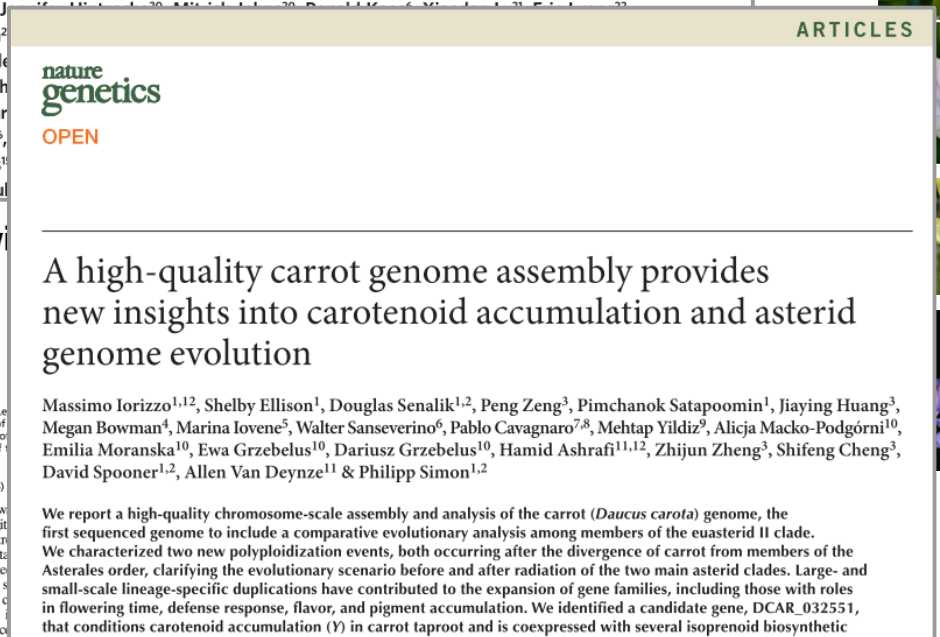
André Marques[a,b], Tiago Ribeiro[a,b], Pavel Neumann[c], Jiří Macas[c], Petr Novák[c], Veit Schubert[b], Marco Pellino[b], Jörg Fuchs[b], Wei Ma[b], Markus Kuhlmann[b], Ronny Brandt[b], André L. L. Vanzela[d], Tomáš Beseda[e], Hana Šimková[e], Andrea Pedrosa-Harand[a,1], and Andreas Houben[b,1]

[a]Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, 50670-420 Recife, Pernambuco, Brazil; [b]Le
Institute of Plant Genetics and Crop Plant Research Gatersleben, 06466 Stadt Seeland, Germany; [c]Laboratory of Molecular Cytogenetics, Institute of
Plant Molecular Biology, Biology Centre of the Academy of Sciences of the Czech Republic, 370 05 Ceske Budejovice, Czech Republic; [d]Laboratory o
Cytogenetics and Plant Diversity, Department of General Biology, State University of Londrina, 86057-970, Londrina, Paraná, Brazil; and [e]Center of
Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, CZ-78371 Olomouc, Czech Republic

Edited by James A. Birchler, University of Missouri-Columbia, Columbia, MO, and approved September 11, 2015 (received for review June 23, 2015)

Holocentric chromosomes lack a primary constriction, in contrast to monocentrics. They form kinetochores distributed along almost the entire poleward surface of the chromatids, to which spindle fibers attach. No centromere-specific DNA sequence has been found for any holocentric organism studied so far. It was proposed that centromeric repeats, typical for many monocentric species, could not occur in holocentrics, most likely because of differences in the centromere organization. Here we show that the holoki-netic centromeres of the Cyperaceae *Rhynchospora pubera* are

cases, a longitudinal CENH3-positive centromere structure w
served during mitosis. In the rush *Luzula* (Juncaceae), the longit
centromere forms a groove (here referred as the "centr
groove") in each sister chromatid along almost the whole meta
chromosome except for the most terminal regions (9–11). Re
a similar centromere organization was found in the sedge s
Rhynchospora pubera (12). The absence of CENH3 and the c
meric protein C (CENP-C) in some lineages of holocentric
(13) challenges the general notion of a conserved molecular c



ARTICLES

nature genetics

OPEN

A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution

Massimo Iorizzo[1,12], Shelby Ellison[1], Douglas Senalik[1,2], Peng Zeng[3], Pimchanok Satapoomin[1], Jiaying Huang[3], Megan Bowman[4], Marina Iovene[5], Walter Sanseverino[6], Pablo Cavagnaro[7,8], Mehtap Yildiz[9], Alicja Macko-Podgórni[10], Emilia Moranska[10], Ewa Grzebelus[10], Dariusz Grzebelus[10], Hamid Ashrafi[11,12], Zhijun Zheng[3], Shifeng Cheng[3], David Spooner[1,2], Allen Van Deynze[11] & Philipp Simon[1,2]

We report a high-quality chromosome-scale assembly and analysis of the carrot (*Daucus carota*) genome, the first sequenced genome to include a comparative evolutionary analysis among members of the euasterid II clade. We characterized two new polyploidization events, both occurring after the divergence of carrot from members of the Asterales order, clarifying the evolutionary scenario before and after radiation of the two main asterid clades. Large- and small-scale lineage-specific duplications have contributed to the expansion of gene families, including those with roles in flowering time, defense response, flavor, and pigment accumulation. We identified a candidate gene, DCAR_032551, that conditions carotenoid accumulation (*Y*) in carrot taproot and is coexpressed with several isoprenoid biosynthetic

Repetitive DNA characterization using RepeatExplorer

Plants

- Over 100 species characterized so far
- Comparative studies
- Whole genome assembly projects

Mammals

Bats, deer

Fish

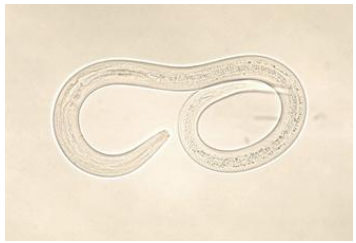
- *Austrolebias charrua, Cynopoecilus melanotaenia*

Insects

- Locust, grasshoppers, kissing bugs

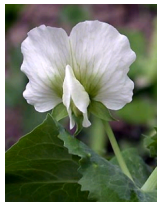
Worms

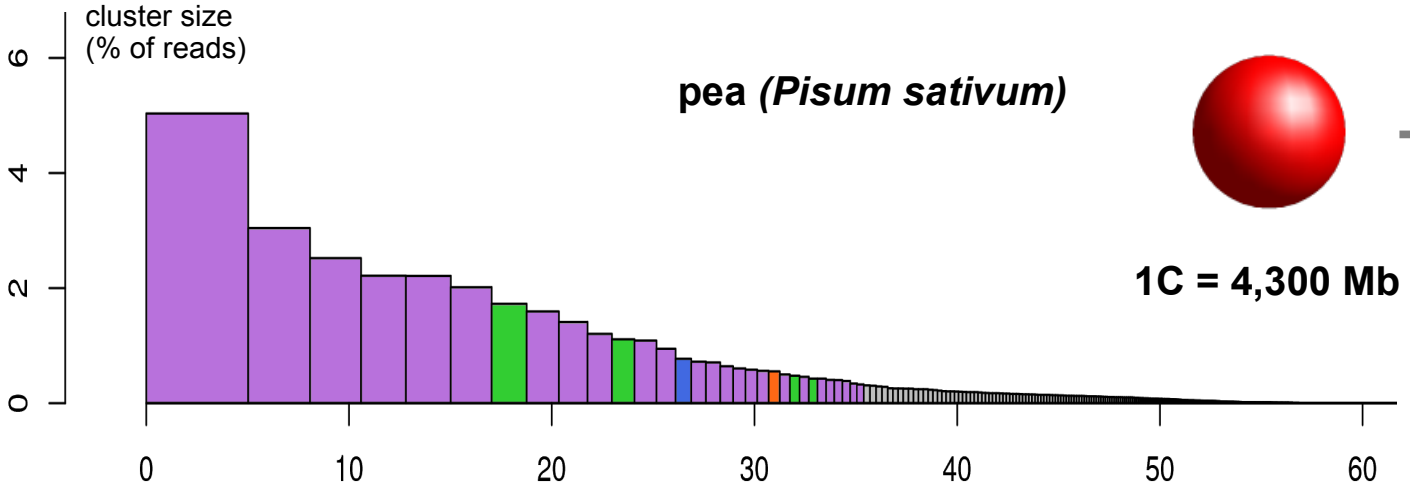
- Soil helmints



APPLICATIONS: Repeat composition of individual species

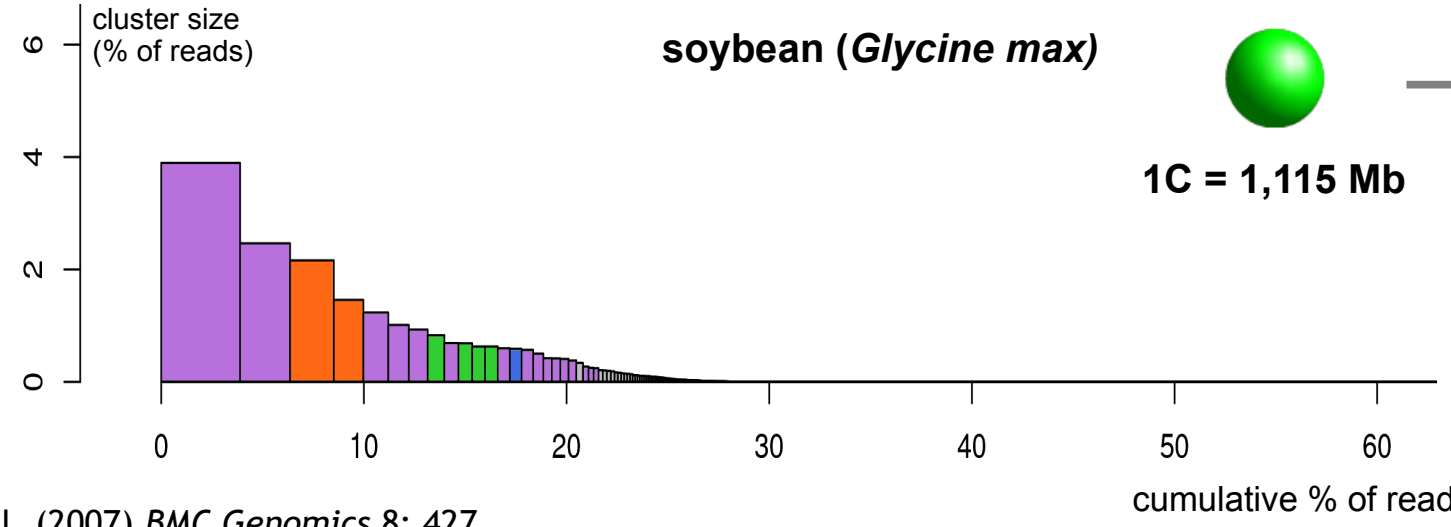
Repeat type: Ty3/gypsy Ty1/copia Satellite DNA rDNA other/unknown





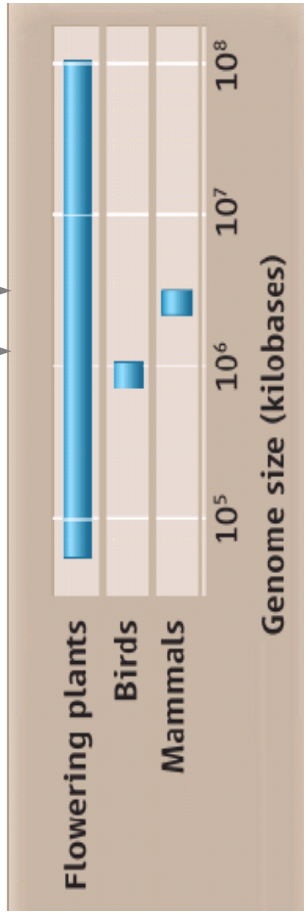
pea (*Pisum sativum*)

1C = 4,300 Mb



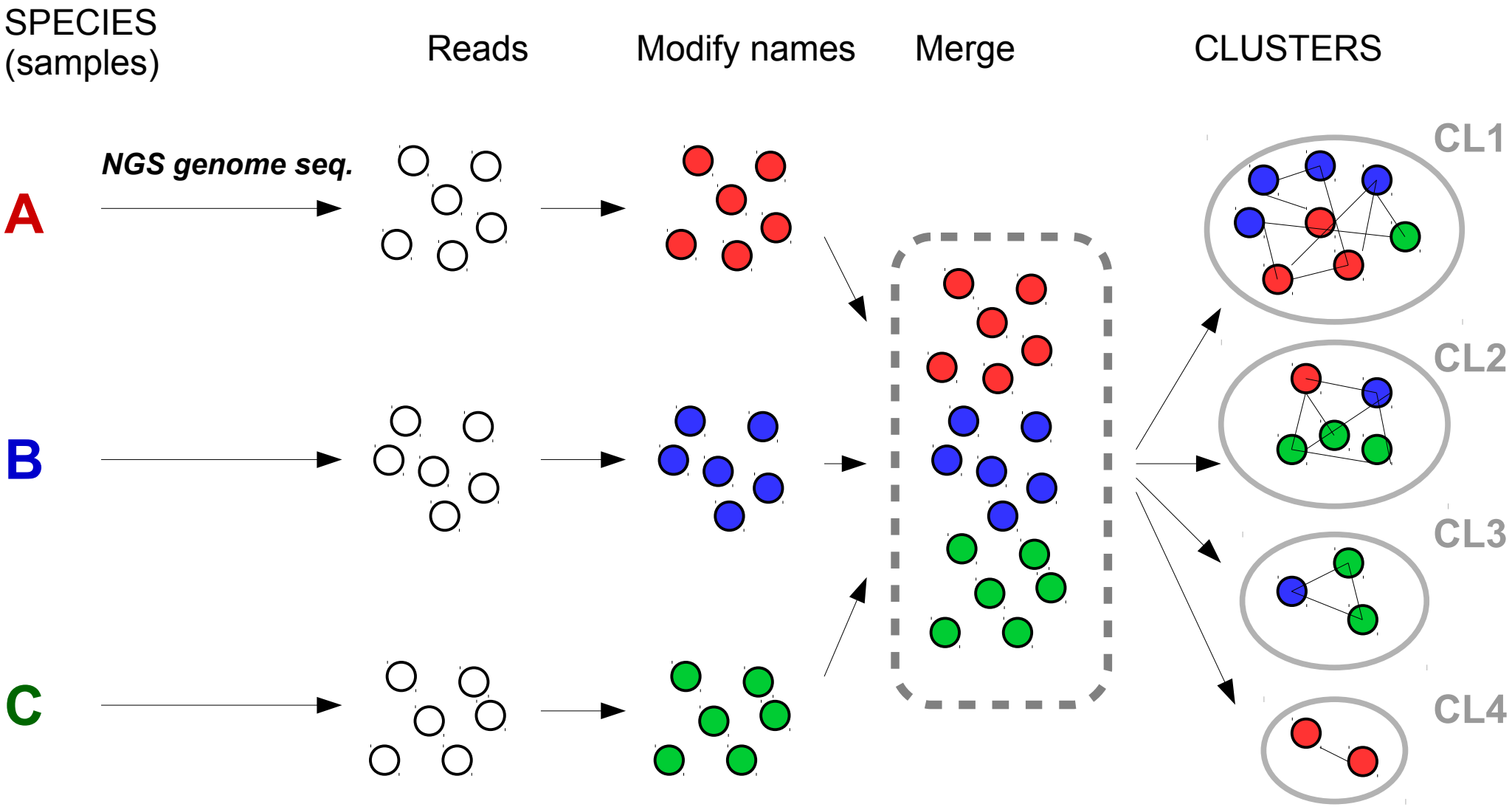
soybean (*Glycine max*)

1C = 1,115 Mb



Macas et al. (2007) *BMC Genomics* 8: 427.

APPLICATIONS: Comparative analysis



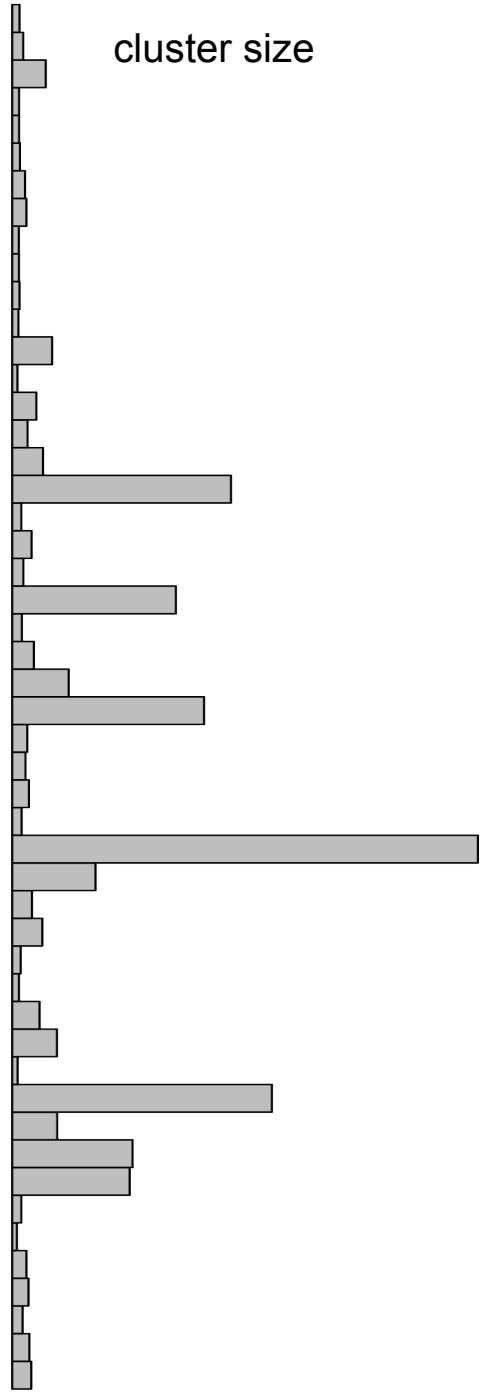
Comparative analysis

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

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

proportion of reads from individual species >>>

cluster size



SPECIES:

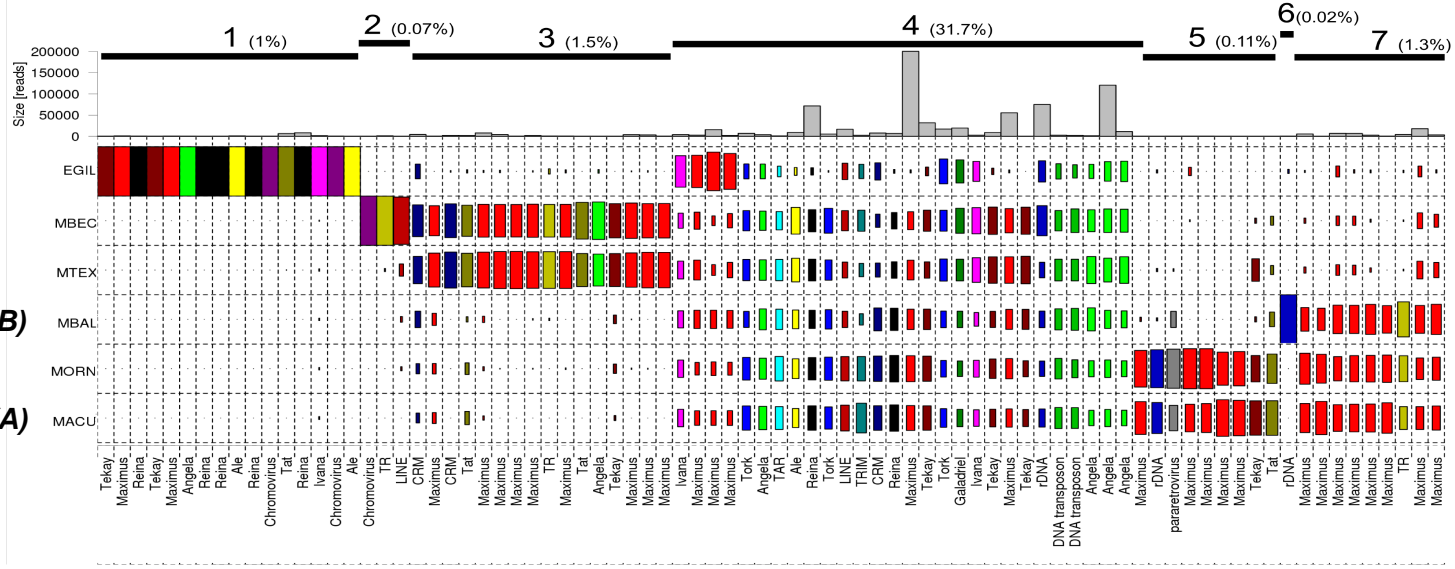
- ***Musa ornata***
- ***M. balbisiana (B)***
- ***M. acuminata (A)***
- ***M. textillis (T)***
- ***M. beccarii***
- ***Ensete gilletii***

Comparative analysis

A

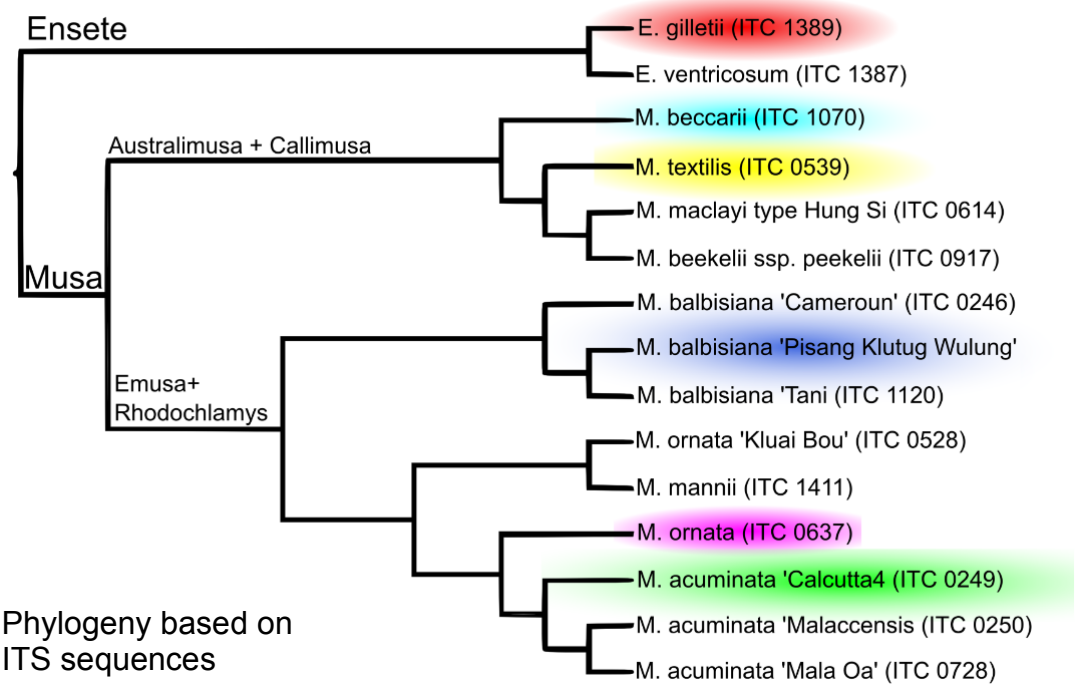
SPECIES:

Ensete gilletii
Musa beccarii
M. textillis (T)
M. balbisiana (B)
M. ornata
M. acuminata (A)



Repeats

- Maximus
- Angela
- Tork
- Ale
- Ivana
- TAR
- Reina
- Tekay
- Galadriel
- CRM
- Tat
- Chromovirus/unspecified
- TRIM
- pararetrovirus
- LINE
- DNA transposon
- rDNA
- TR
- unknown



Phylogeny based on
ITS sequences

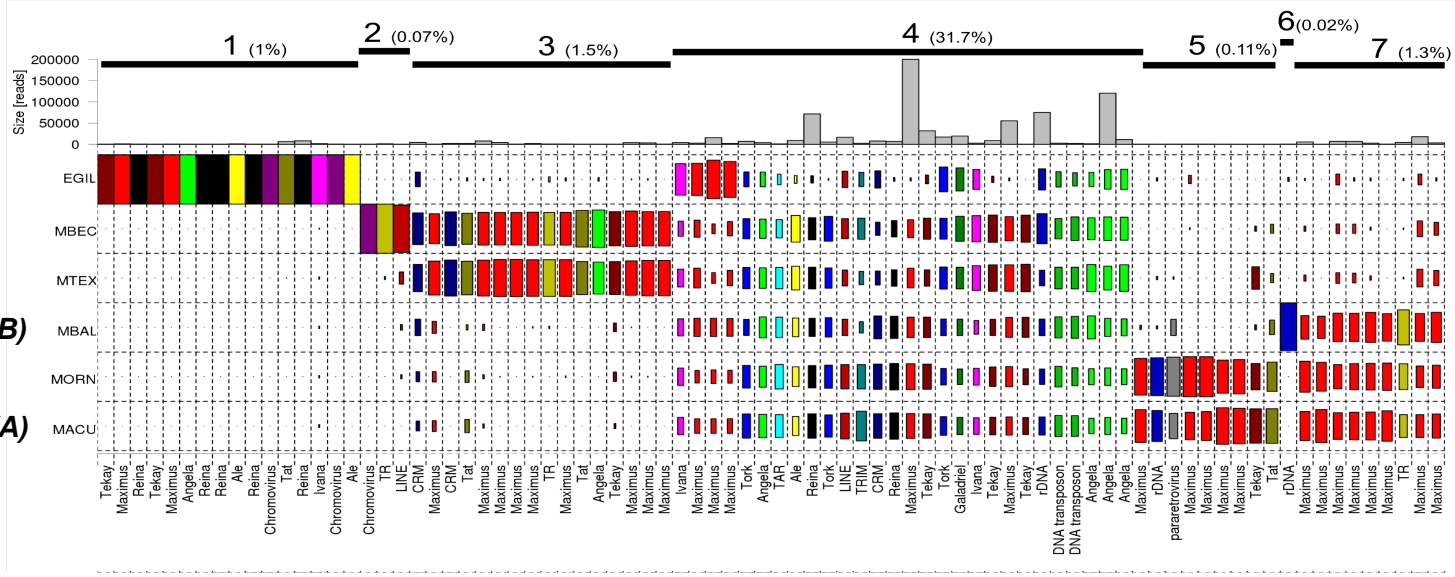
Novak et al. 2014

Comparative analysis

A

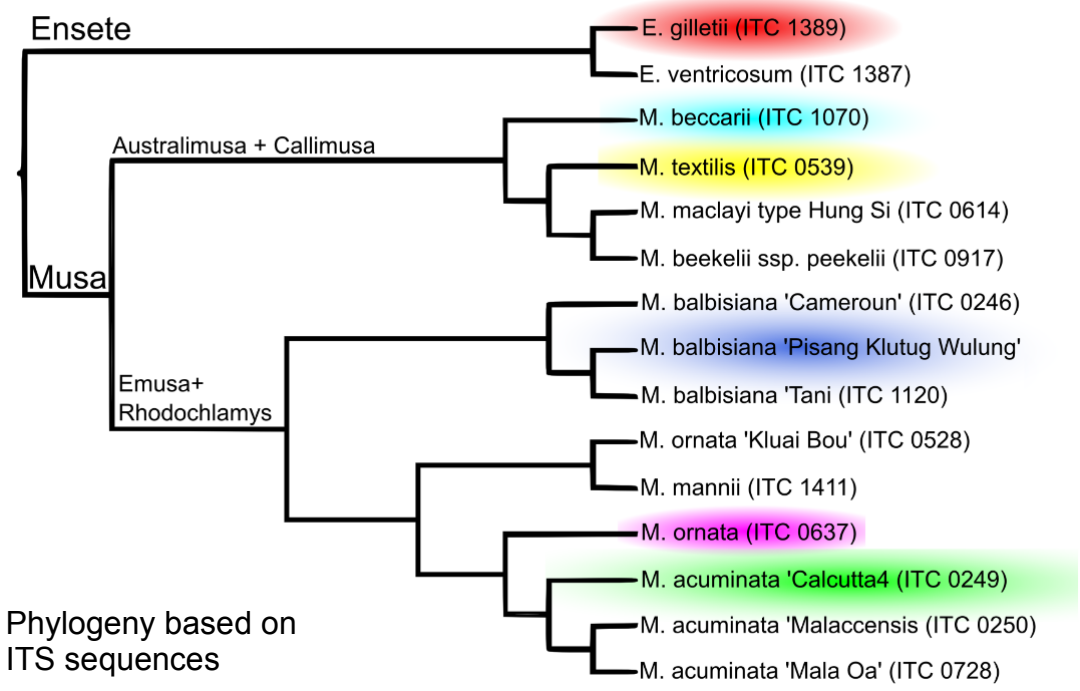
SPECIES:

Ensete gilletii
Musa beccarii
M. textillis (T)
M. balbisiana (B)
M. ornata
M. acuminata (A)



Repeats

- Maximus
- Angela
- Tork
- Ale
- Ivana
- TAR
- Reina
- Tekay
- Galadriel
- CRM
- Tat
- Chromovirus/unspecified
- TRIM
- pararetrovirus
- LINE
- DNA transposon
- rDNA
- TR
- unknown



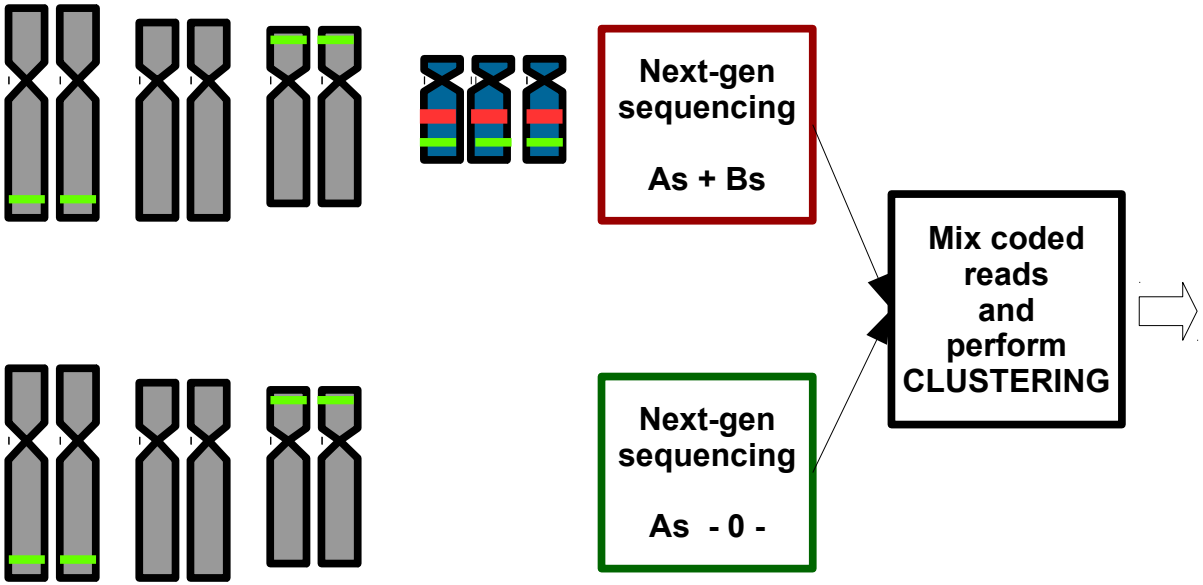
Phylogeny based on ITS sequences

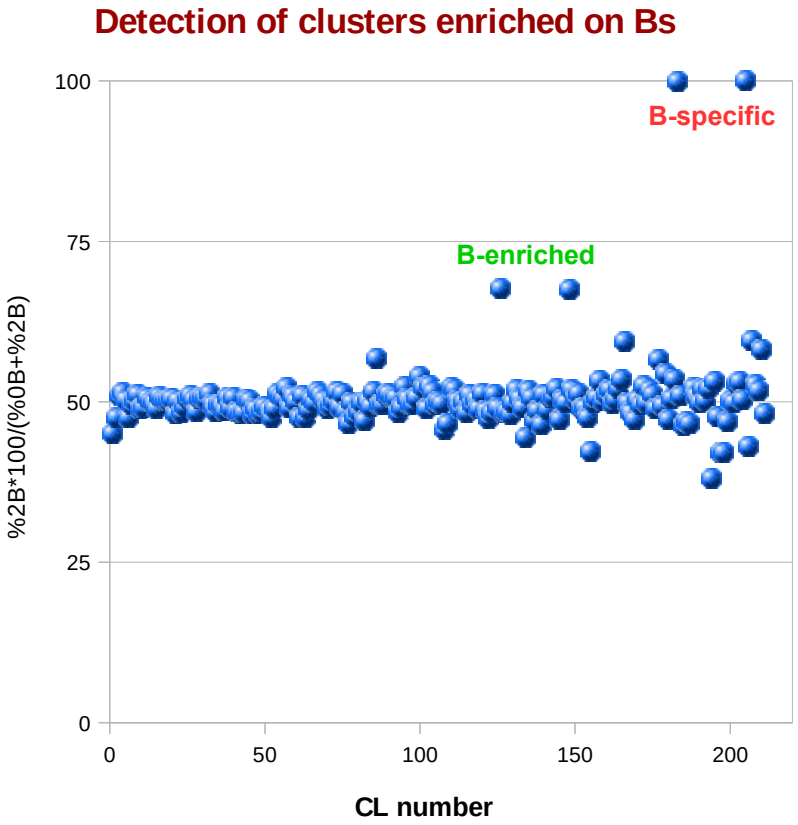
Genomic repeat abundances contain phylogenetic signal

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

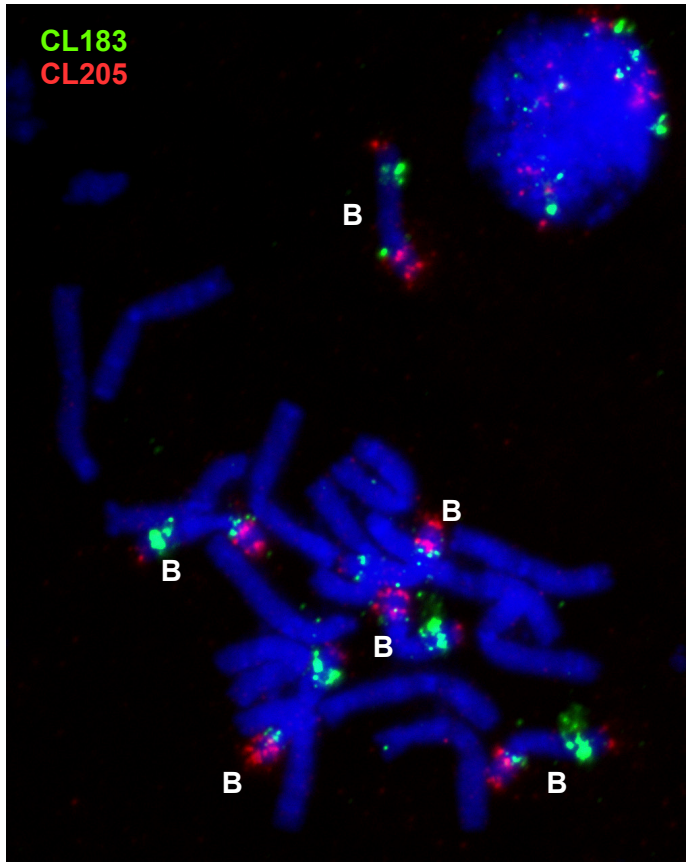
Novak et al. 2014

Comparative analysis (searching for B-specific repeats)

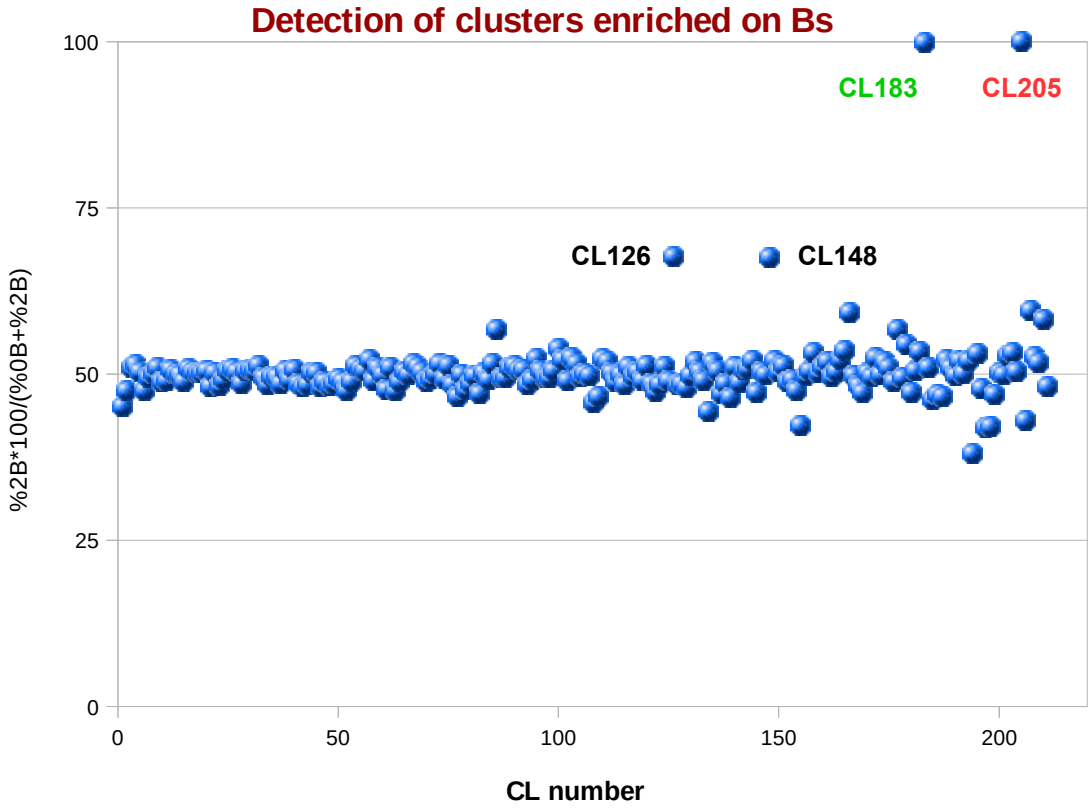




Aegilops speltoides, comparative analysis of 0B / 2B plants



FISH by Alevtina Ruban



Wu et al. (2019) *New Phytol.*, in press

Repeats significantly enriched on Bs

CL	reads	[%]	0B	%	2B	%	%2B*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

Illumina seq. (36 nt reads)

Isolation of nuclei, digestion with micrococcal nuclease

Chromatin immunoprecipitation with **CenH3** antibody

DNA isolation ("**ChIP**")

Sequencing

9.5 mil. reads

DNA isolation ("**INPUT**")

Sequencing

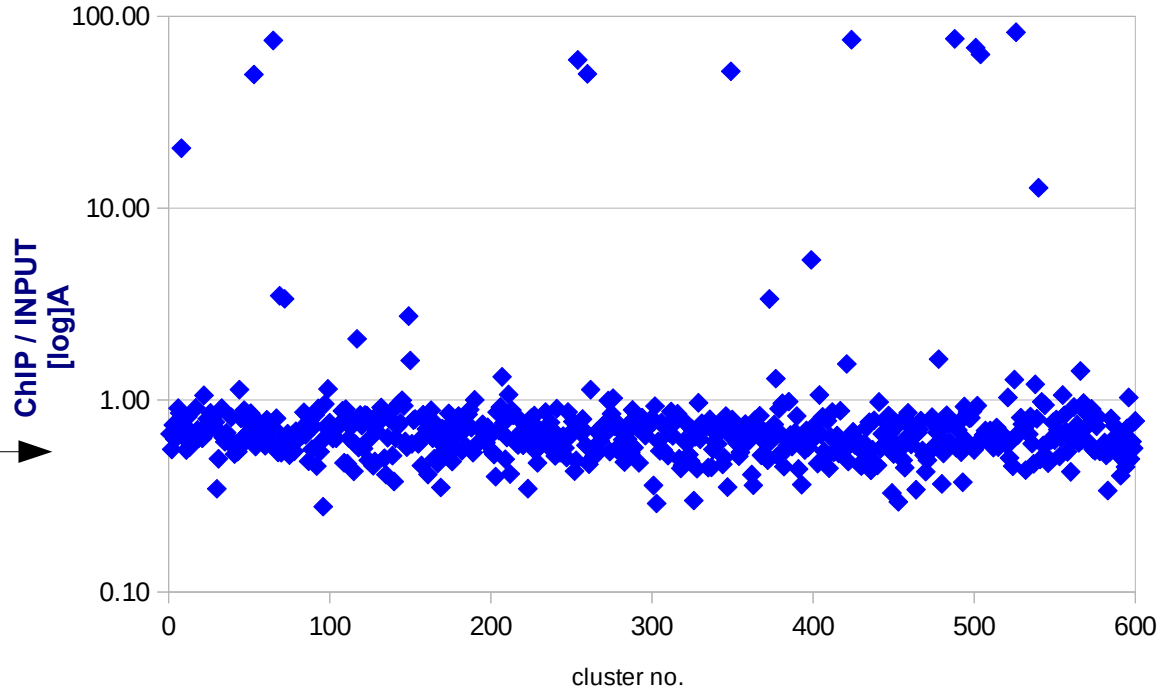
20 mil. reads

Mapping reads to clusters of long (>100nt) reads, calculating ratio of **ChIP / INPUT**

Cluster = repeat family

100-fold enrichment

no enrichment



Identification of centromeric repeats by ChIP-seq

Outline of the experiment

Illumina seq. (36 nt reads)

Isolation of nuclei, digestion with micrococcal nuclease

Chromatin immunoprecipitation with **CenH3** antibody

DNA isolation (“**ChIP**”)

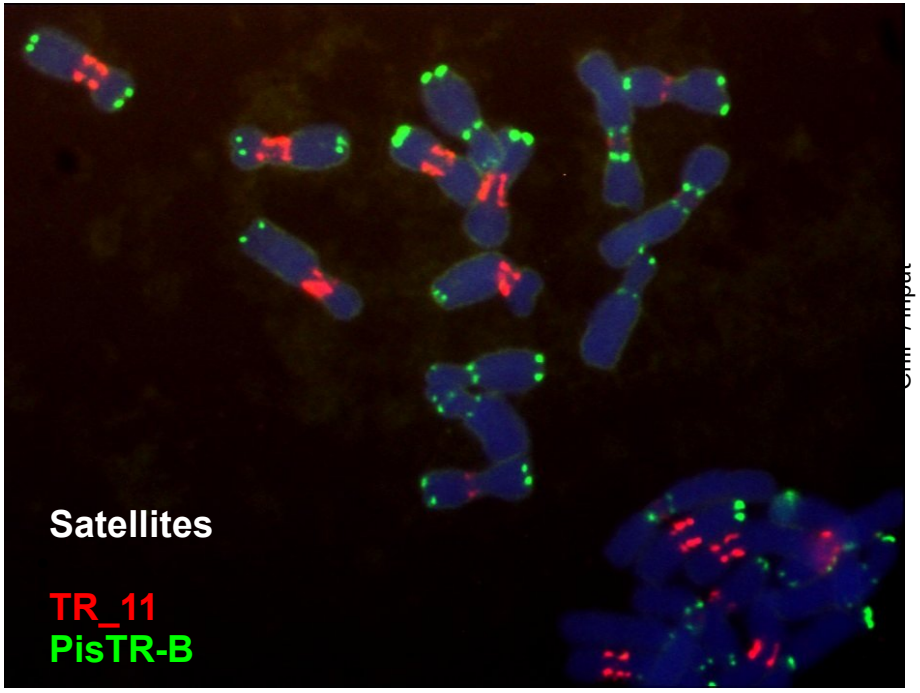
Sequencing **9.5 mil. reads**

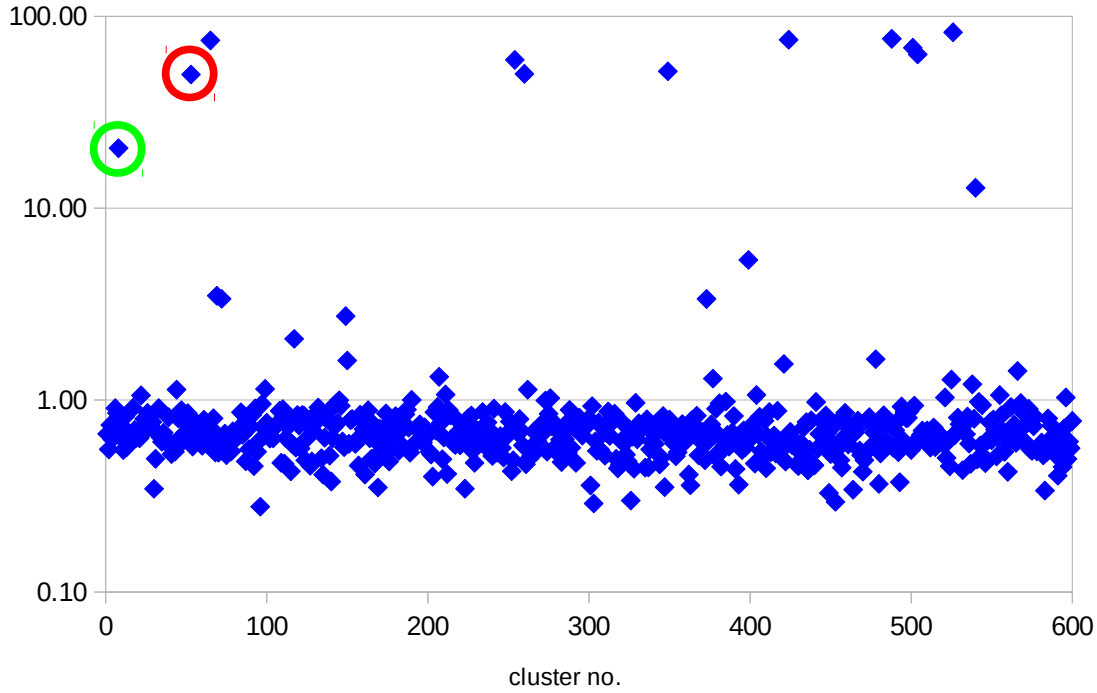
DNA isolation (“**INPUT**”)

Sequencing **20 mil. reads**

Mapping reads to clusters of long (>100nt) reads, calculating ratio of **ChIP / INPUT**

Cluster = repeat family





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

Identification of centromeric repeats by ChIP-seq

Outline of the experiment

Illumina seq. (36 nt reads)

Isolation of nuclei, digestion with micrococcal nuclease

Chromatin immunoprecipitation with **CenH3** antibody

DNA isolation (“**ChIP**”)

Sequencing **9.5 mil. reads**

DNA isolation (“**INPUT**”)

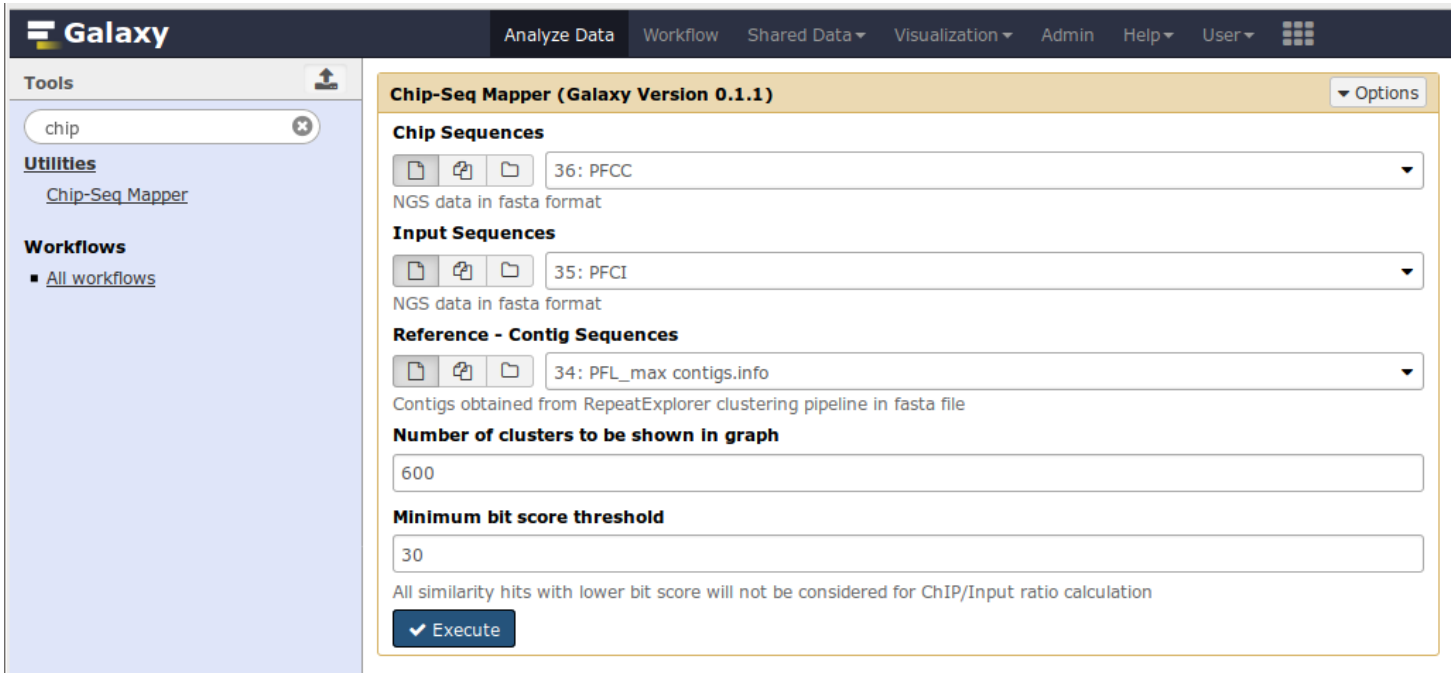
Sequencing **20 mil. reads**

Mapping reads to clusters of long (>100nt) reads, calculating ratio of **ChIP / INPUT**

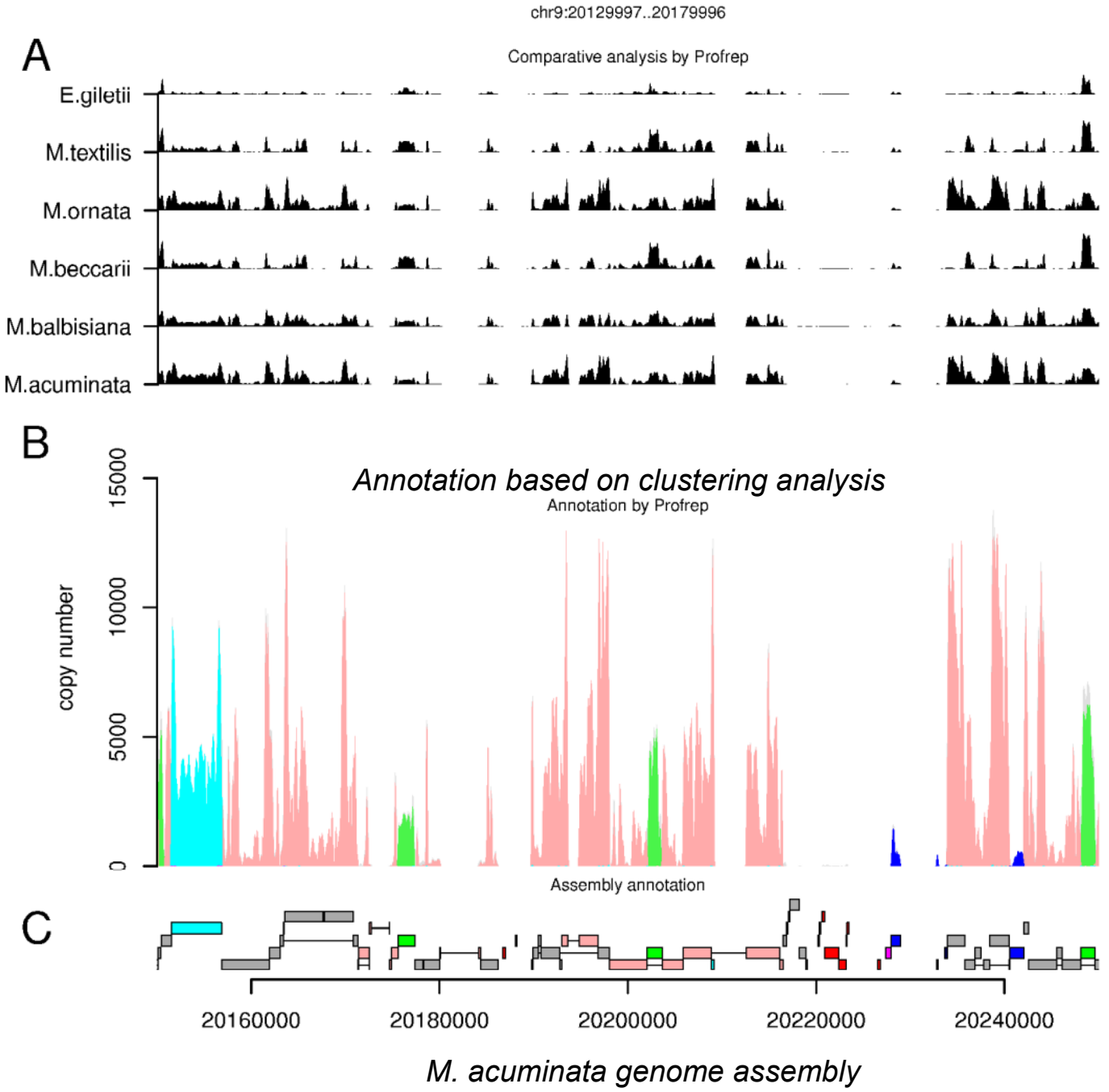
Cluster = repeat family

ChIP-seq Mapper >>>

now available as a tool on our Galaxy server

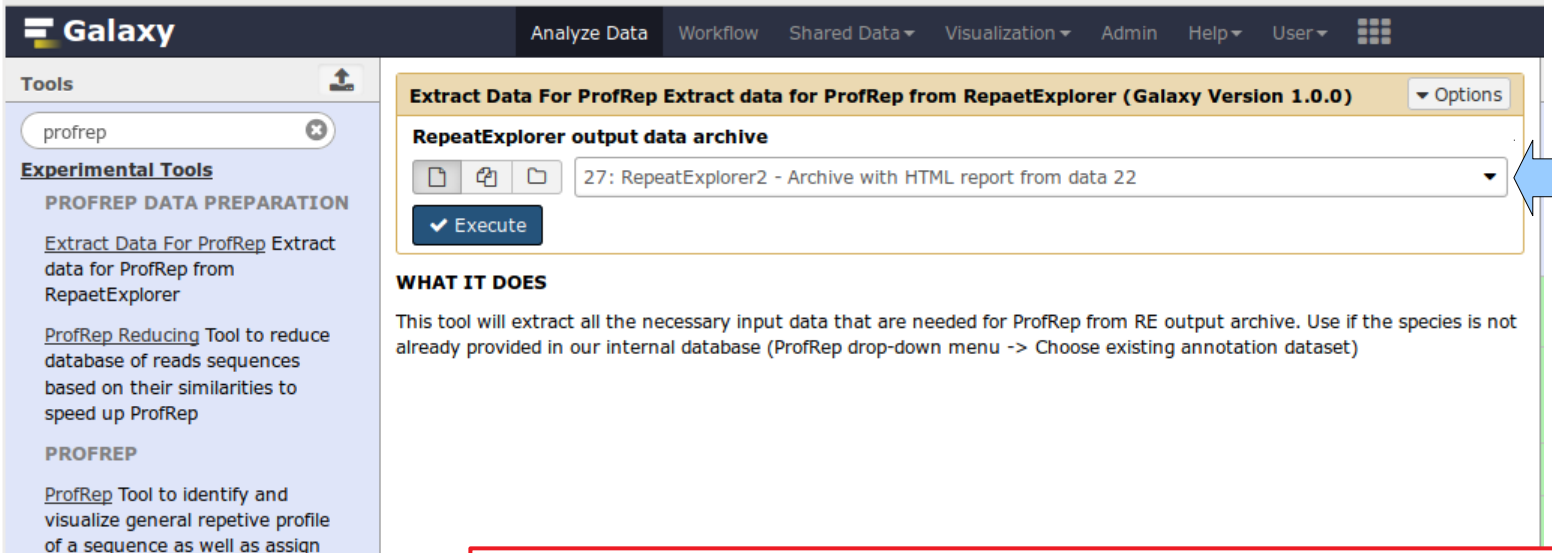


PROFREP - improving repeat annotation in genome assembly



Novak, Hribova, et al.

PROFREP and DANTE



Set of tools available on the Galaxy server

data import from RE2...

… to final annotation