Welcome !

Principles and history of RepeatExplorer

2007

First paper on repeat clustering from NGS data

BMC Genomics

BioMed Central

Research article

Open Access

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

Jiří Macas*, Pavel Neumann and Alice Navrátilová

Address: Biology Centre ASCR, Institute of Plant Molecular Biology, Branišovská 31, České BudEjovice, CZ-37005, Czech Republic

Email: Jiří Macas* - macas@umbr.cas.cz; Pavel Neumann - neumann@umbr.cas.cz; Alice Navrátilová - navratil@umbr.cas.cz

* Corresponding author

Published: 21 November 2007

BMC Genomics 2007, **8**:427 doi:10.1186/1471-2164-8-427

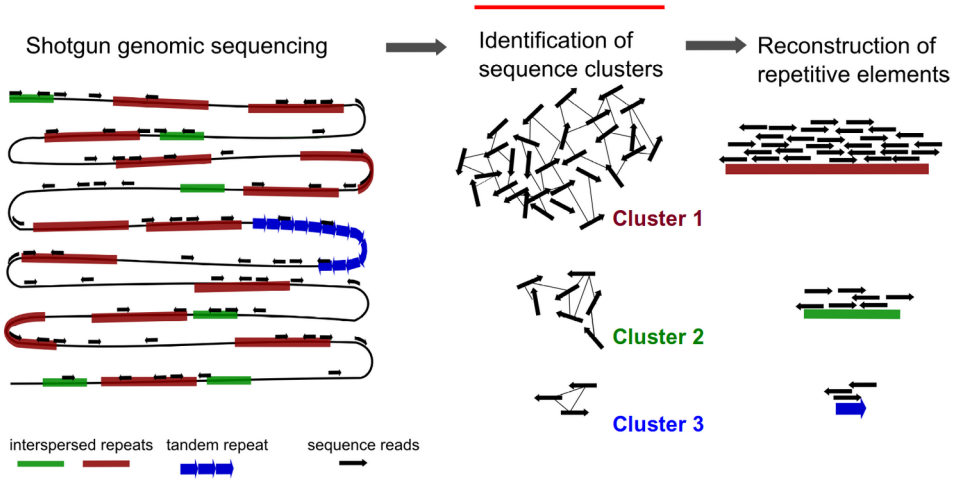
Received: 13 August 2007

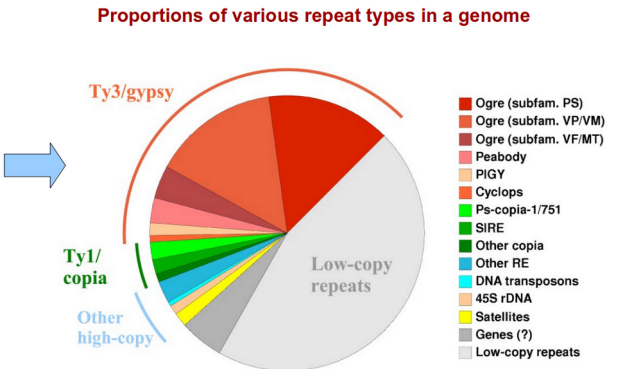
Accepted: 21 November 2007

Principles and history of RepeatExplorer

2007

First paper on repeat clustering from NGS data



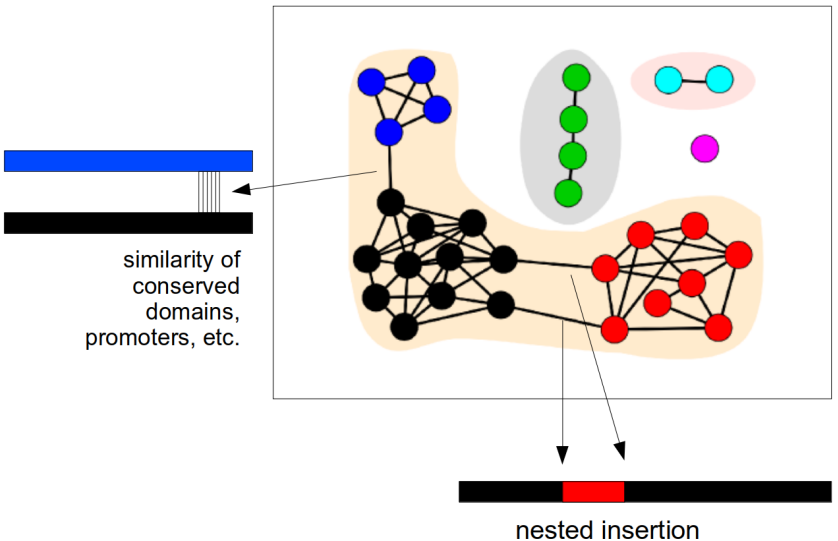


CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

Principles and history of RepeatExplorer

2007

First paper on repeat clustering from NGS data



Chimeric clusters !

Single linkage clustering => *connected components*

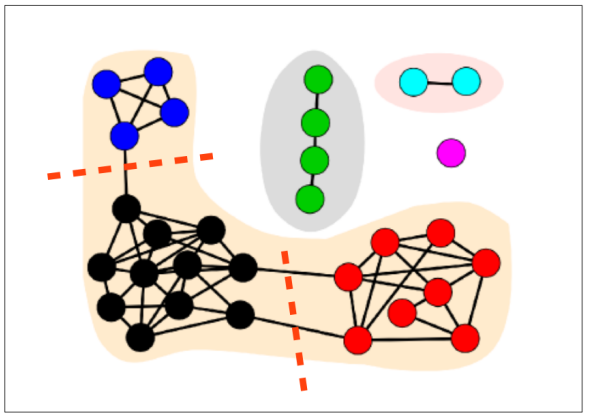
TGICL
(TIGR Gene Indices clustering tool)
Pertea et al., 2003

Principles and history of RepeatExplorer

2007 … **2010**

First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)



Graph-based clustering

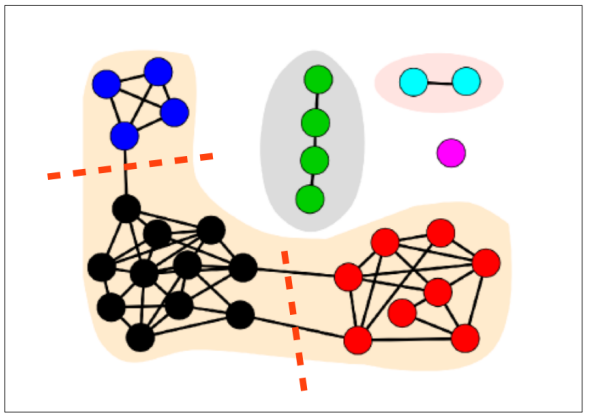
- Sequence overlaps between the reads are transformed to a graph where the reads are represented as nodes and their similarities as edges connecting the nodes
- Graph structure is examined to detect ***communities of frequently connected nodes*** which **are split to separate clusters**

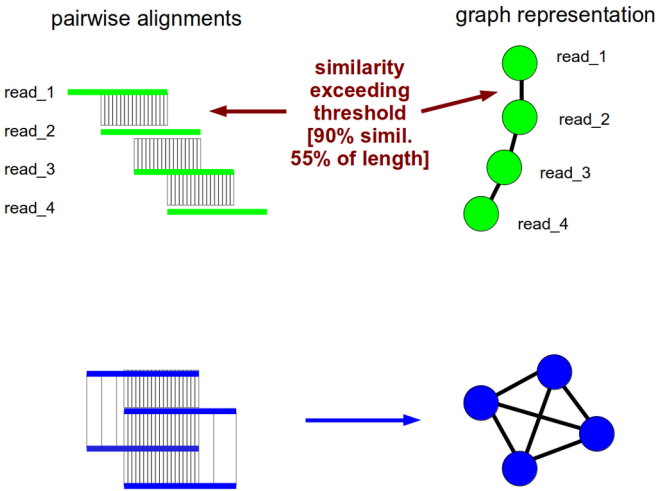
Principles and history of RepeatExplorer

2007 … **2010**

First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)





Graph-based clustering

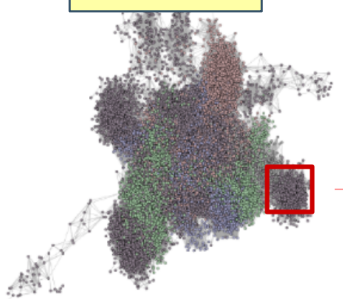
- Sequence overlaps between the reads are transformed to a graph represented as nodes and their similarities as edges connecting t
- Graph structure is examined to detect ***communities of frequently connected nodes*** which **are split to separate clusters**

Principles and history of RepeatExplorer

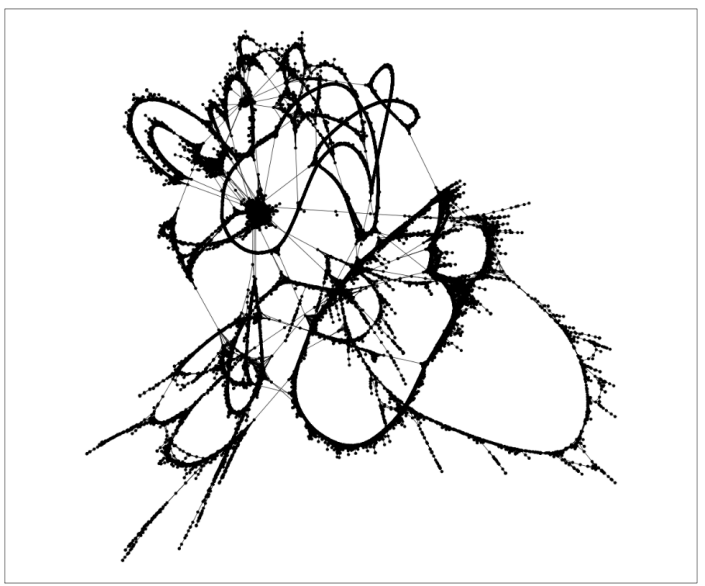
2007 … 2010

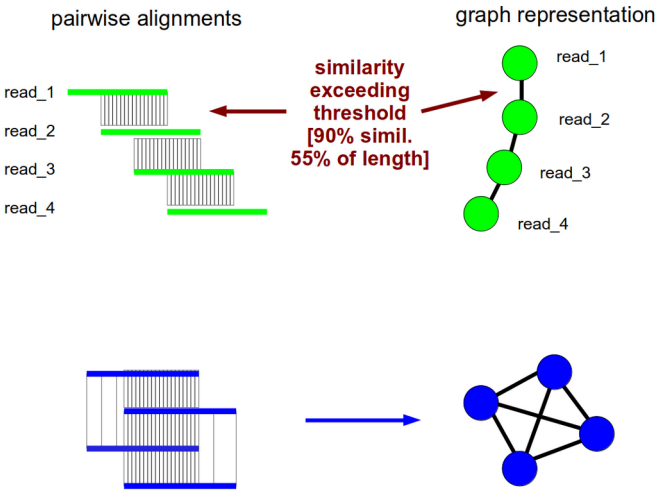
First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)



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



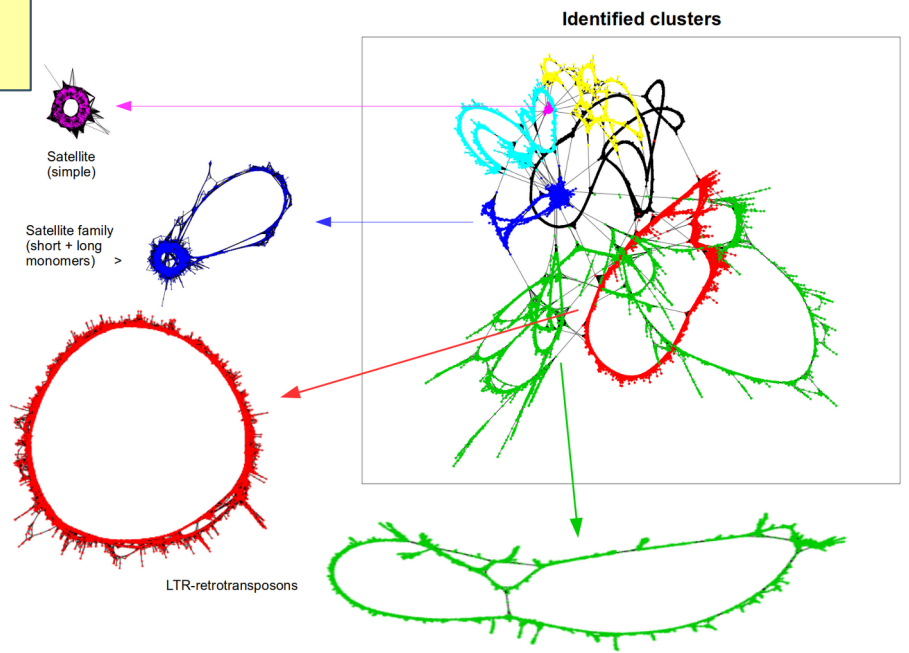


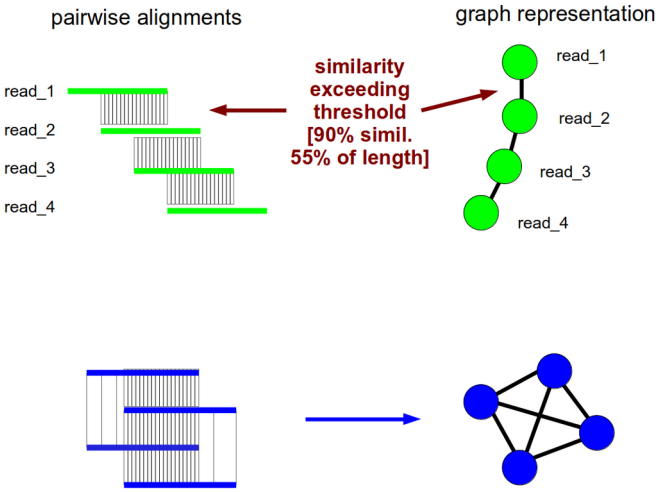
Principles and history of RepeatExplorer

2007 … **2010**

First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)





Principles and history of RepeatExplorer

2007 … **2010** … **2013**

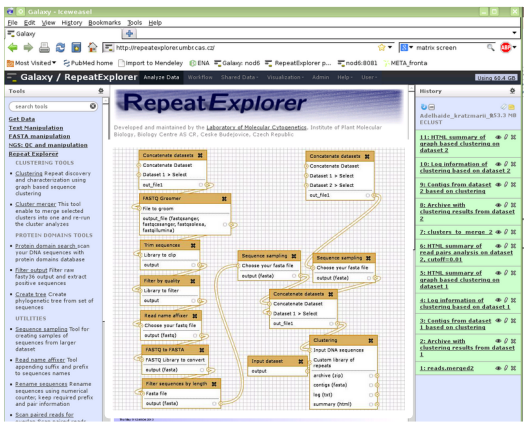
First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)

command-line version

Repeat Explorer in Galaxy (Novak et al. 2013)

Public web-based server



Principles and history of RepeatExplorer

2007 … **2010** … **2013** **2014** … **2016** **2017**

First paper on repeat clustering from NGS data

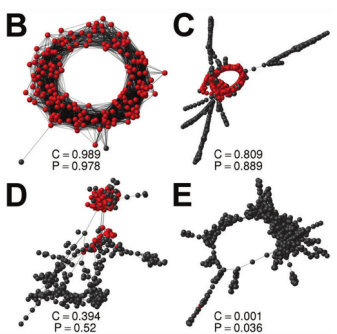
Introduction of **graph-based clustering** (Novak et al. 2010)

Repeat Explorer in Galaxy (Novak et al. 2013)

elixir CZECH REPUBLIC

ELIXIR $$$

TAREAN

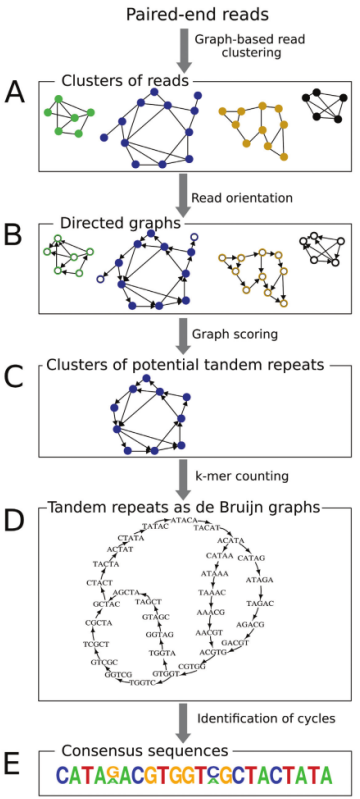


Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx257

TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads

Petr Novák, Laura Ávila Robledillo, Andrea Koblížková, Iva Vrbová, Pavel Neumann and Jiří Macas*

Institute of Plant Molecular Biology, Biology Centre CAS, České Budějovice CZ-37005, Czech Republic



Principles and history of RepeatExplorer

2007 … **2010** … **2013** **2014** … **2016** **2017** **2018**

First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)

Repeat Explorer in Galaxy (Novak et al. 2013)

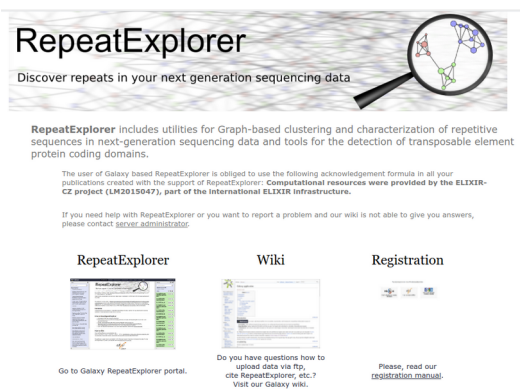
Public web-based server

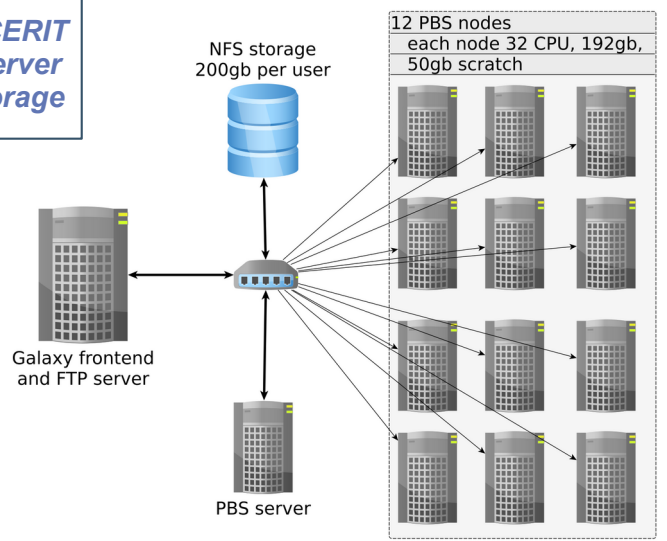
elixir CZECH REPUBLIC

ELIXIR $$$

TAREAN

ELIXIR / CERIT Galaxy server + data storage





Principles and history of RepeatExplorer

2007 … **2010** … **2013** **2014** … **2016** **2017** **2018**

First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)

Repeat Explorer in Galaxy (Novak et al. 2013)

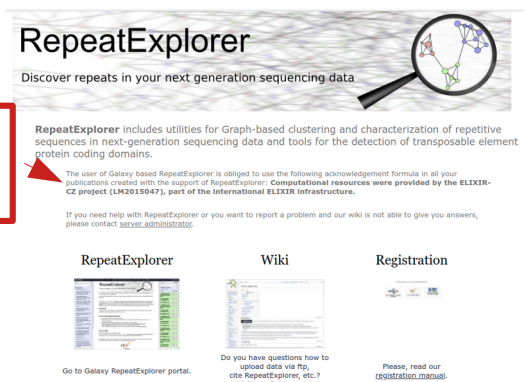
Public web-based server

elixir CZECH REPUBLIC

ELIXIR $$$

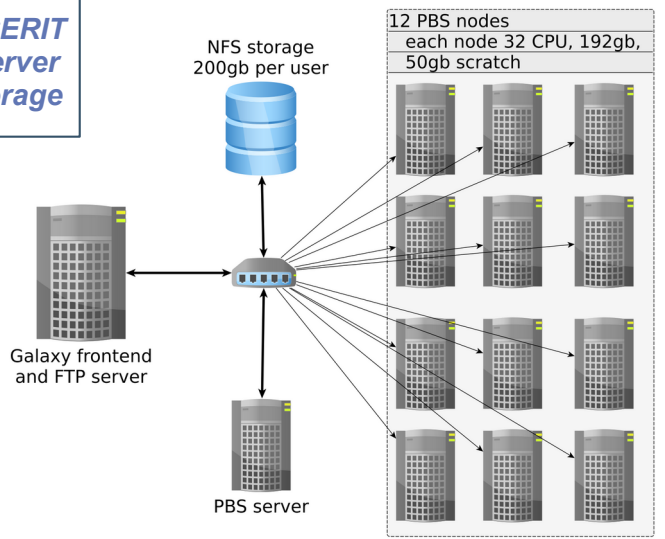
TAREAN

ELIXIR / CERIT Galaxy server + data storage



The user of Galaxy based RepeatExplorer is obliged to use the following acknowledgement formula in all your publications created with the support of RepeatExplorer: **Computational resources were provided by the ELIXIR-CZ project (LM2015047), part of the international ELIXIR infrastructure.**

Please, acknowledge ELIXIR in your publications !



Principles and history of RepeatExplorer

2007 … **2010** … **2013** **2014** … **2016** **2017** **2018** **2019**

First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)

Repeat Explorer in Galaxy (Novak et al. 2013)

elixir CZECH REPUBLIC

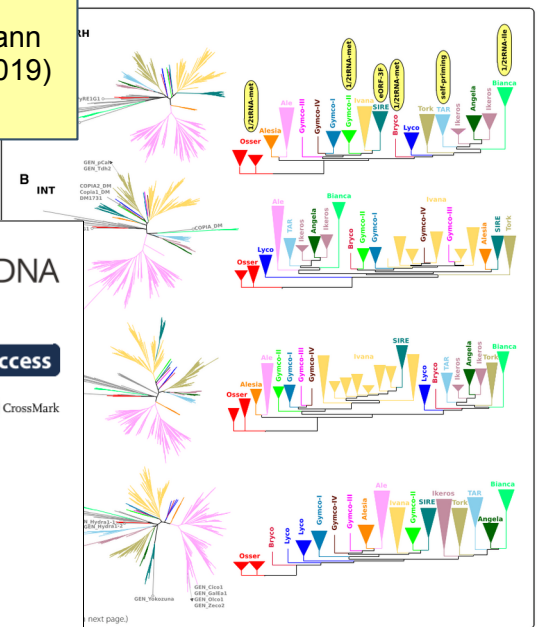
ELIXIR $$$

TAREAN

ELIXIR / CERIT Galaxy server + data storage

REXdb database

(Neumann et al. 2019)



Neumann *et al. Mobile DNA* (2019) 10:1
https://doi.org/10.1186/s13100-018-0144-1

Mobile DNA

RESEARCH Open Access

CrossMark

Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification

Pavel Neumann*, Petr Novák, Nina Hoštáková and Jiří Macas

Principles and history of RepeatExplorer

2007 … **2010** … **2013** **2014** … **2016** **2017** **2018** **2019** **2020**

First paper on repeat clustering from NGS data

Introduction of **graph-based clustering** (Novak et al. 2010)

Repeat Explorer in Galaxy (Novak et al. 2013)

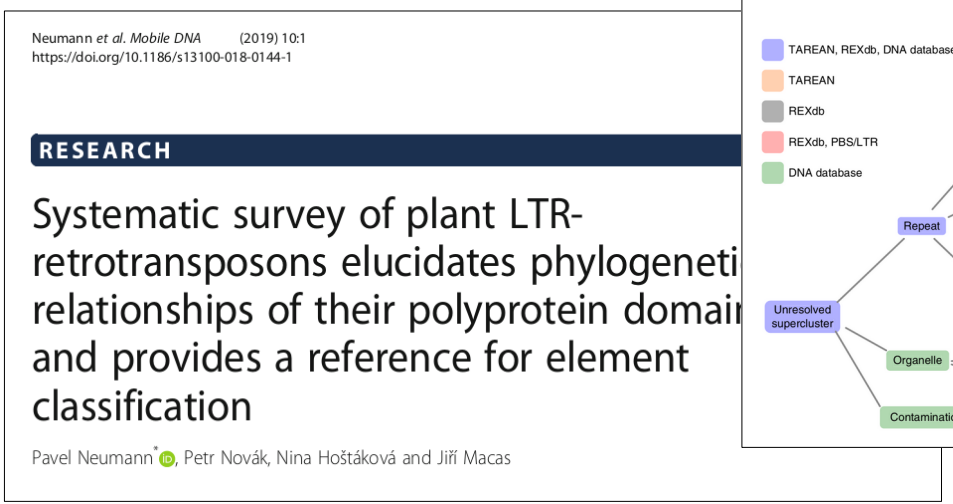
elixir CZECH REPUBLIC

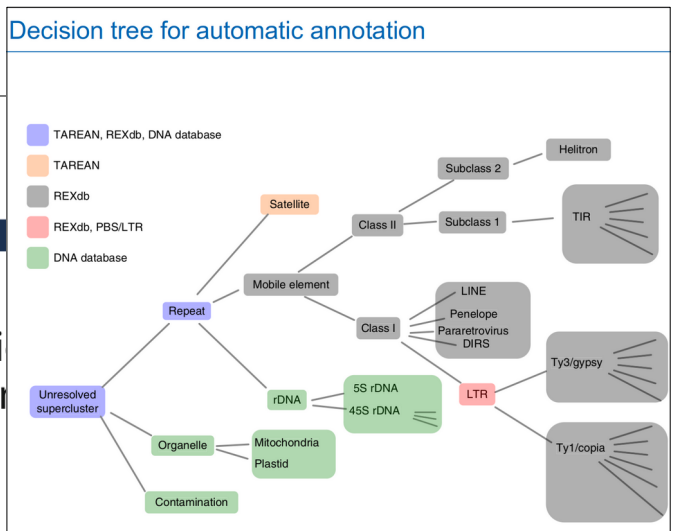
TAREAN

REXdb

(Neumann et al. 2019)

Repeat Explorer ver. 2





Principles and history of RepeatExplorer

2007 … 2010 … 2013 2014 … 2016 **2017** **2018** **2019** **2020**

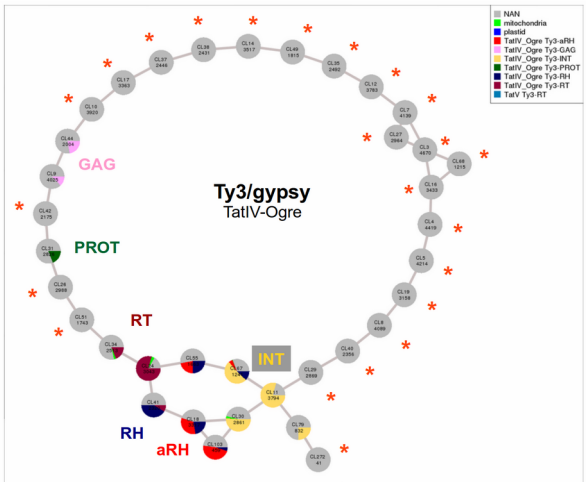
TAREAN

REXdb

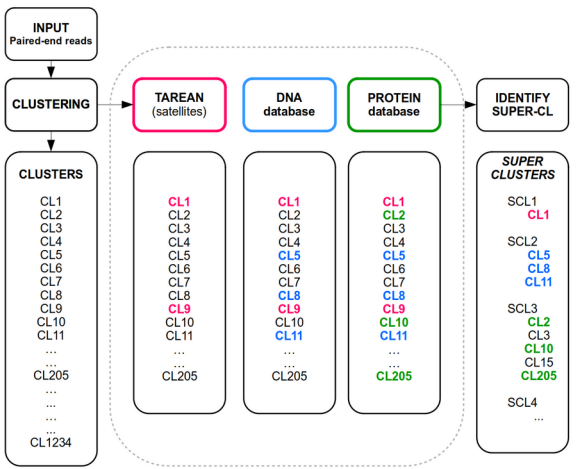
(Neumann et al. 2019)

Repeat Explorer ver. 2

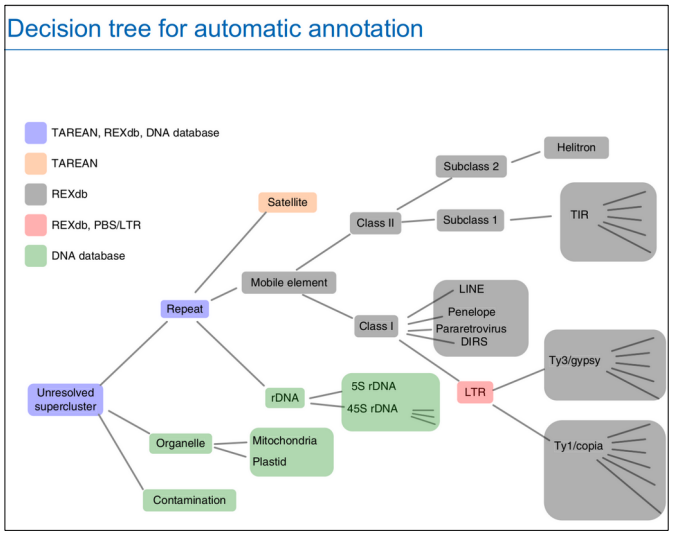
Superclusters provide more complete annotation



RepeatExplorer 2 – automatic detection of superclusters



Decision tree for automatic annotation



Principles and history of RepeatExplorer

2007 … 2010 … 2013 2014 … 2016 **2017** **2018** **2019** **2020** **2023**

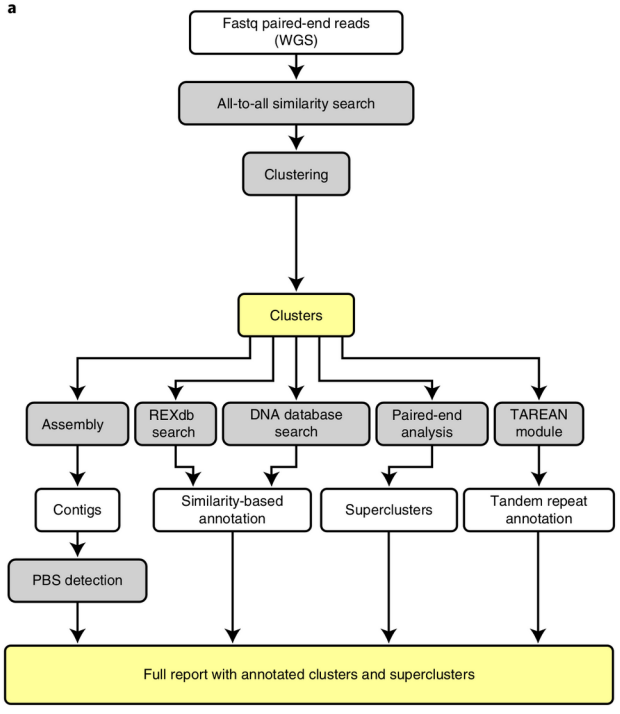
TAREAN

REXdb

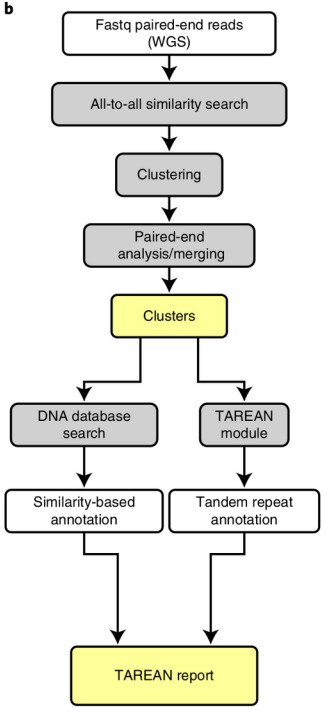
(Neumann et al. 2019)

Repeat Explorer ver. 2

RepeatExplorer pipeline



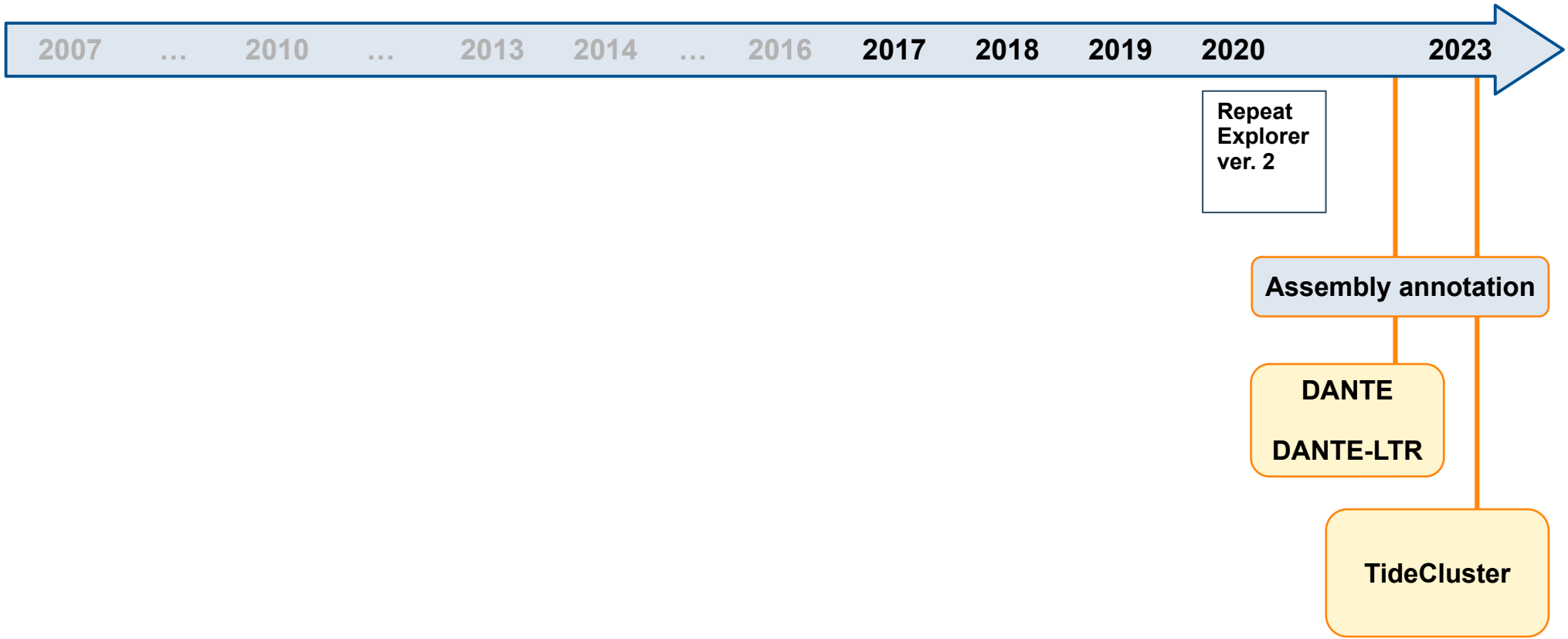
TAREAN pipeline



***Additional tools*:**

- **ChIP-seq Mapper**
  - **Long reads**
    - **DANTE**

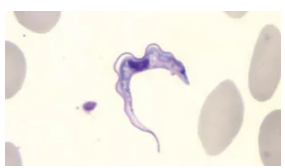
Principles and history of RepeatExplorer

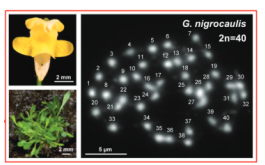


Applications

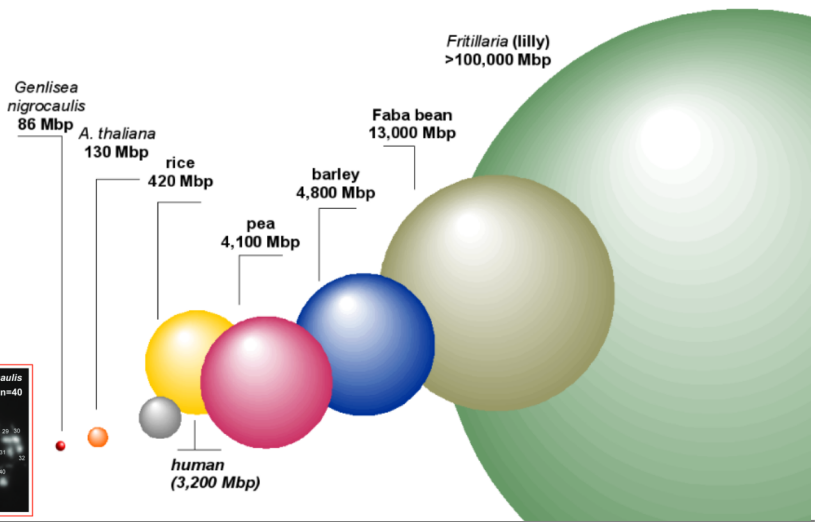
Applications

- Repeat composition
  - single species



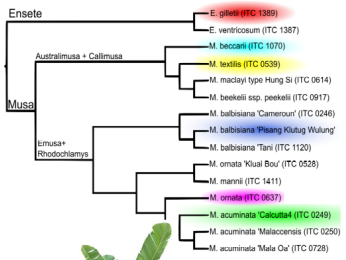


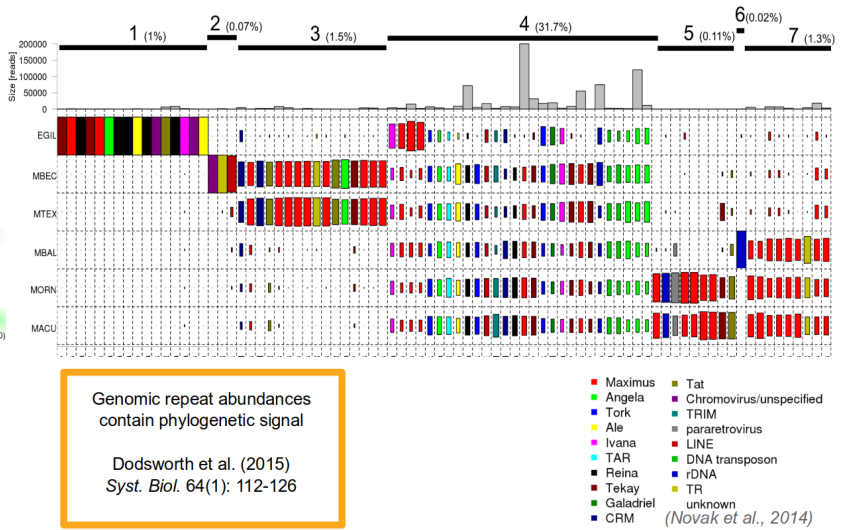
Plant species varying ~2,000-fold
in their genome sizes



Applications

- Repeat composition
  - single species
  - comparative analysis





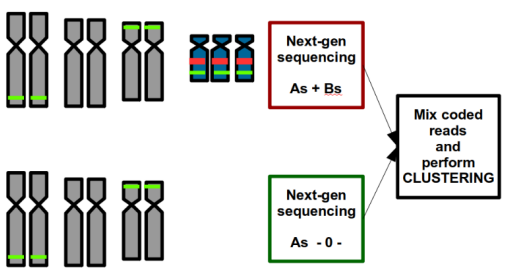
Genomic repeat abundances contain phylogenetic signal

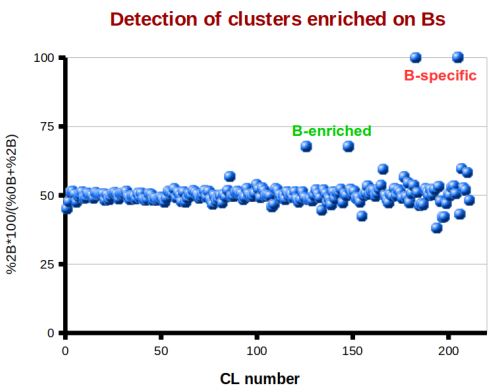
Dodsworth et al. (2015)
Syst. Biol. 64(1): 112-126

(Novak et al., 2014)

Identification of chromosome B-specific repeats

Comparative analysis of B+/- plants



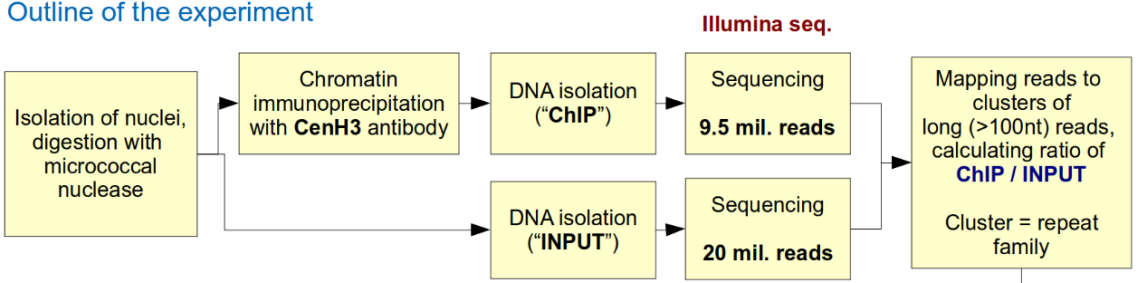


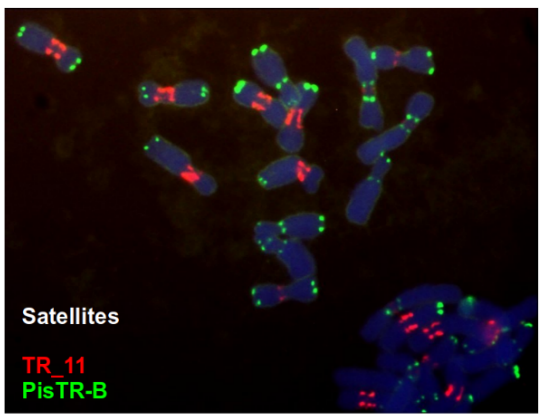
Applications

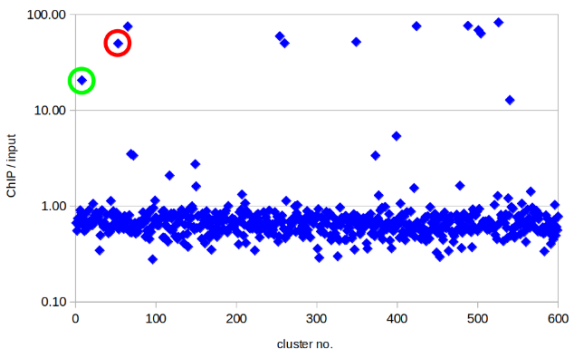
- Repeat composition
  - single species
  - comparative analysis
- Repeat clusters as a reference
  - ChIP-seq

Identification of centromeric repeats by ChIP-seq

Outline of the experiment







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

Applications

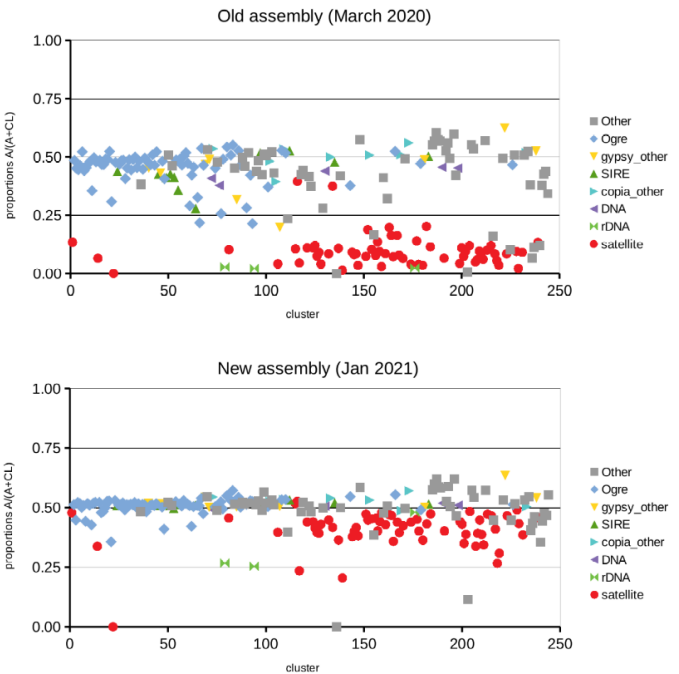
- Repeat composition
  - single species
  - comparative analysis
- Repeat clusters as a reference
  - ChIP-seq
  - assembly

Assessing completeness of genome assemblies

Illumina WGS reads (cluster size)

Mapped fragments from assembly

				sum of hits 133232913	blasted fragments 207900483	
Size_adjusted	Proportion [%]	Final_annotation	CL	hits		
			0	6171084	2.968	
85804	4.187	All/repeat/satellite/FabTR_FokI	1	1343009	0.646	0.13
38954	1.901	All/repeat/mobile_element/Class_I	2	3723933	1.791	0.49
37146	1.813	All/repeat/mobile_element/Class_I	3	3091188	1.487	0.45
37051	1.808	All/repeat/mobile_element/Class_I	4	3014725	1.450	0.45
36830	1.797	All/repeat/mobile_element/Class_I	5	3221042	1.549	0.46
33547	1.637	All/repeat/mobile_element/Class_I	6	3719091	1.789	0.52
32220	1.572	All/repeat/mobile_element/Class_I	7	2574963	1.239	0.44
31149	1.520	All/repeat/mobile_element/Class_I	8	2619726	1.260	0.45
30972	1.511	All/repeat/mobile_element/Class_I	9	2791506	1.343	0.47
30361	1.482	All/repeat/mobile_element/Class_I	10	2857979	1.375	0.48
29548	1.442	All/repeat/mobile_element/Class_I	11	1649483	0.793	0.35
29227	1.426	All/repeat/mobile_element/Class_I	12	2924386	1.407	0.50
27341	1.334	All/repeat/mobile_element/Class_I	13	2544864	1.224	0.48
26964	1.316	All/repeat/satellite/FabTR_53	14	189991	0.091	0.06
26082	1.273	All/repeat/mobile_element/Class_I	15	2496449	1.201	0.49
25651	1.252	All/repeat/mobile_element/Class_I	16	2280990	1.097	0.47
25147	1.227	All/repeat/mobile_element/Class_I	17	2239097	1.077	0.47
23553	1.149	All/repeat/mobile_element/Class_I	18	2304645	1.109	0.49
22914	1.118	All/repeat/mobile_element/Class_I	19	2227830	1.072	0.49
22829	1.114	All/repeat/mobile_element/Class_I	20	2547541	1.225	0.52
22696	1.108	All/repeat/mobile_element/Class_I	21	1024482	0.493	0.31
22481	1.097	All/repeat/satellite/FabTR_83	22	6	0.000	0.00
21364	1.043	All/repeat/mobile_element/Class_I	23	1971842	0.948	0.48
20442	0.998	All/repeat/mobile_element/Class_I	24	1610084	0.774	0.44
19605	0.957	All/repeat/mobile_element/Class_I	25	1753755	0.844	0.47
19289	0.941	All/repeat/mobile_element/Class_I	26	1835955	0.883	0.48
19234	0.939	All/repeat/mobile_element/Class_I	27	1842540	0.886	0.49
18932	0.924	All/repeat/mobile_element/Class_I	28	1318285	0.634	0.41
18345	0.895	All/repeat/mobile_element/Class_I	29	1533938	0.738	0.45
18112	0.884	All/repeat/mobile_element/Class_I	30	1753161	0.843	0.49
17991	0.878	All/repeat/mobile_element/Class_I	31	1457392	0.701	0.44
17167	0.838	All/repeat/mobile_element/Class_I	32	1663803	0.800	0.49
16726	0.816	All/repeat/mobile_element/Class_I	33	1693448	0.815	0.50



Applications

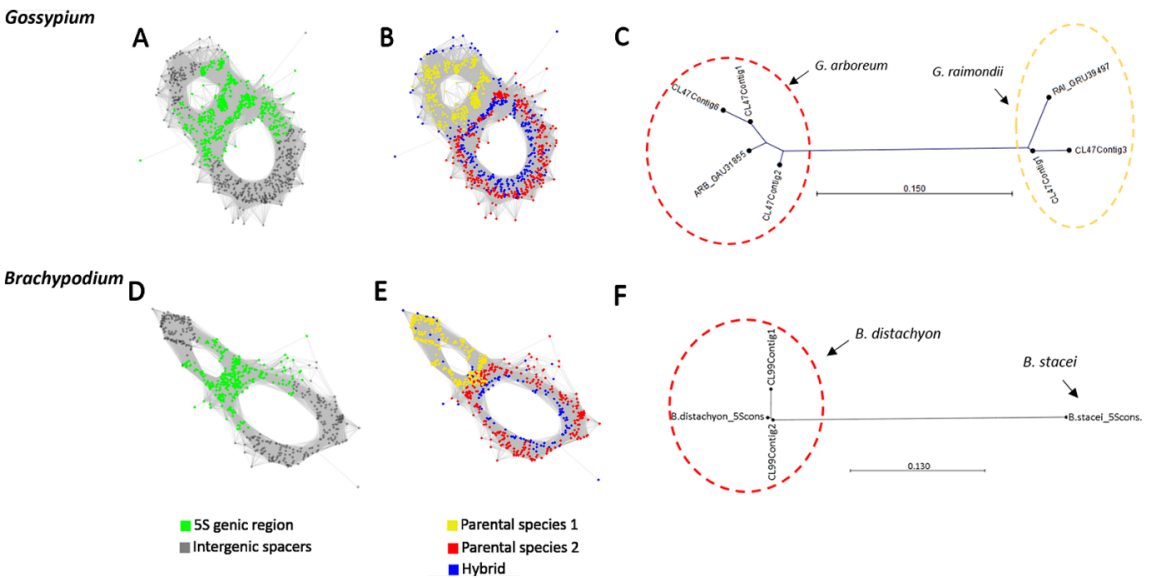
- Repeat composition
  - single species
  - comparative analysis
- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - *graph shapes*

frontiers in Plant Science

ORIGINAL RESEARCH
published: 10 February 2020
doi: 10.3389/fpls.2020.00041

The Utility of Graph Clustering of 5S Ribosomal DNA Homoeologs in Plant Allopolyploids, Homoploid Hybrids, and Cryptic Introgressants

*Sònia Garcia[1,2], Jonathan F. Wendel[3], Natalia Borowska-Zuchowska[4], Malika Aïnouche[5], Alena Kuderova[2] and Ales Kovarik[2]**



Applications

- Repeat composition
  - single species
  - comparative analysis
- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - *graph shapes*
- Satellite DNA
  - cytogenetic studies

Cytogenetic markers (FISH probes)

NGS data (Illumina)

↓

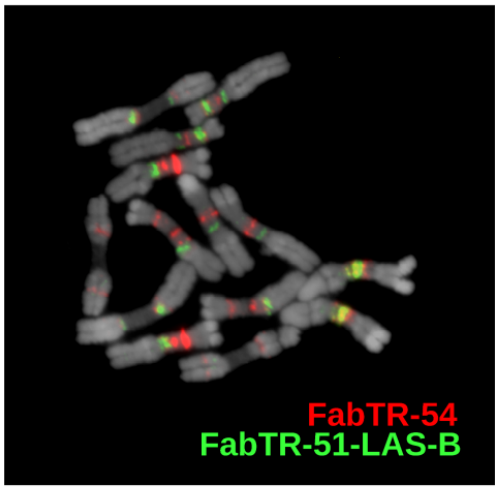
RepeatExplorer / **TAREAN**

↓

Satellite DNA

↓

FISH probe



Lathyrus sativus
(FISH by L. Avila Robledillo)

Applications

- Repeat composition
  - single species
  - comparative analysis
- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - *graph shapes*
- Satellite DNA
  - cytogenetic studies
  - diagnostic markers

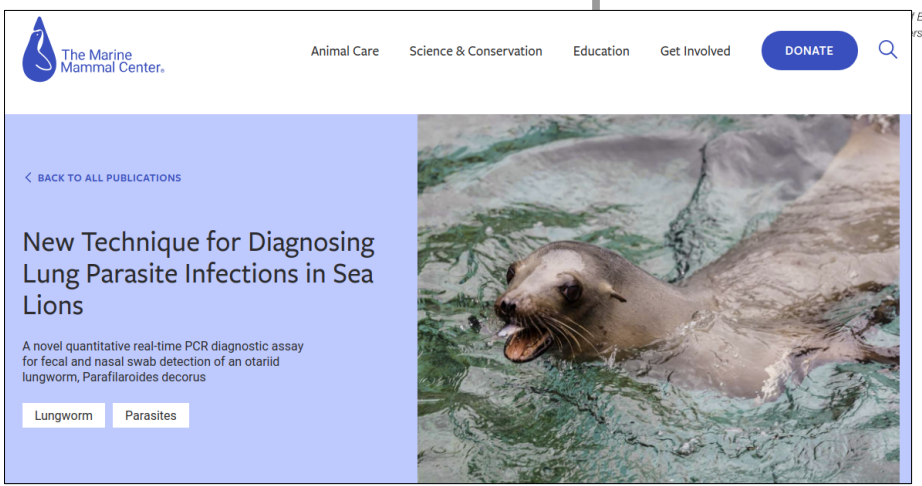

frontiers in Genetics

REVIEW
published: 23 September 2019
doi: 10.3389/fgene.2019.00883

A Case for Using Genomics and a Bioinformatics Pipeline to Develop Sensitive and Species-Specific PCR-Based Diagnostics for Soil-Transmitted Helminths

Jessica R. Grant[1], Nils Pilotte[1,2] and Steven A. Williams[1,2]*

[1] Biological Sciences, Smith College, Northampton, MA, United States, [2] Molecular and Cellular Biology ersity of Massachusetts, Amherst, MA, United States


The Marine Mammal Center

Animal Care | Science & Conservation | Education | Get Involved | DONATE

BACK TO ALL PUBLICATIONS

New Technique for Diagnosing Lung Parasite Infections in Sea Lions

A novel quantitative real-time PCR diagnostic assay for fecal and nasal swab detection of an otariid lungworm, Parafilaroides decorus

Lungworm | Parasites

Applications

- Repeat composition
  - single species
  - comparative analysis
- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - *graph shapes*
- Satellite DNA
  - cytogenetic studies
  - diagnostic markers
- RE utilization in other pipelines


SCIENTIFIC REPORTS

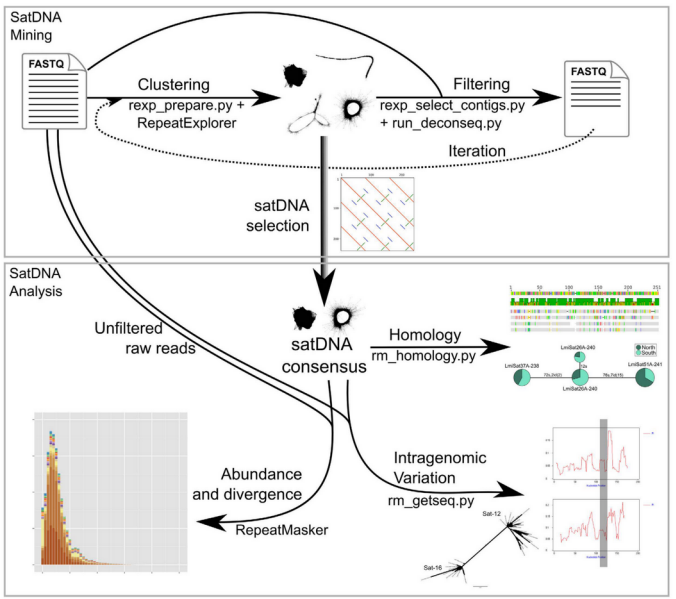
OPEN

High-throughput analysis of the satellitome illuminates satellite DNA evolution

Received: 14 January 2016
Accepted: 02 June 2016

Francisco J. Ruiz-Ruano, María Dolores López-León, Josefa Cabrero & Juan Pedro M. Camacho

satMiner



Applications

- Repeat composition
  - single species
  - comparative analysis
- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - *graph shapes*
- Satellite DNA
  - cytogenetic studies
  - diagnostic markers
- RE utilization in other pipelines

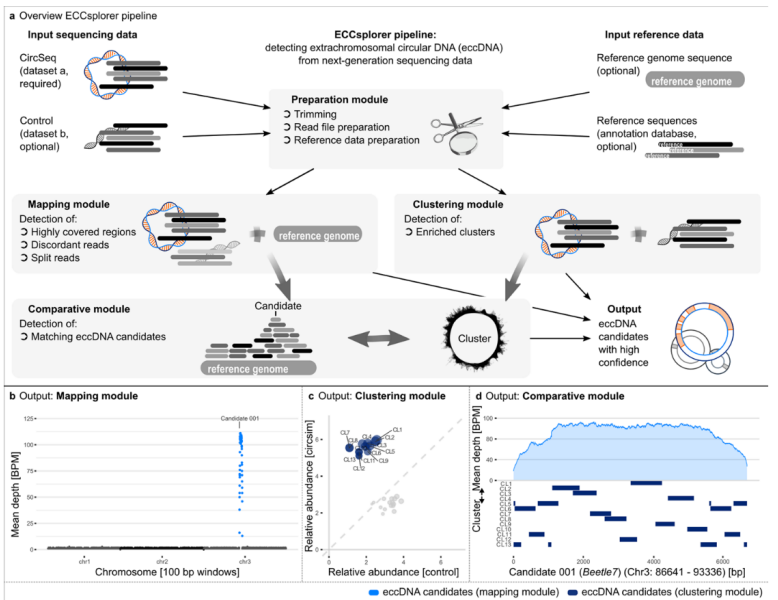
Mann *et al. BMC Bioinformatics* (2022) 23:40
https://doi.org/10.1186/s12859-021-04545-2

BMC Bioinformatics

SOFTWARE Open Access

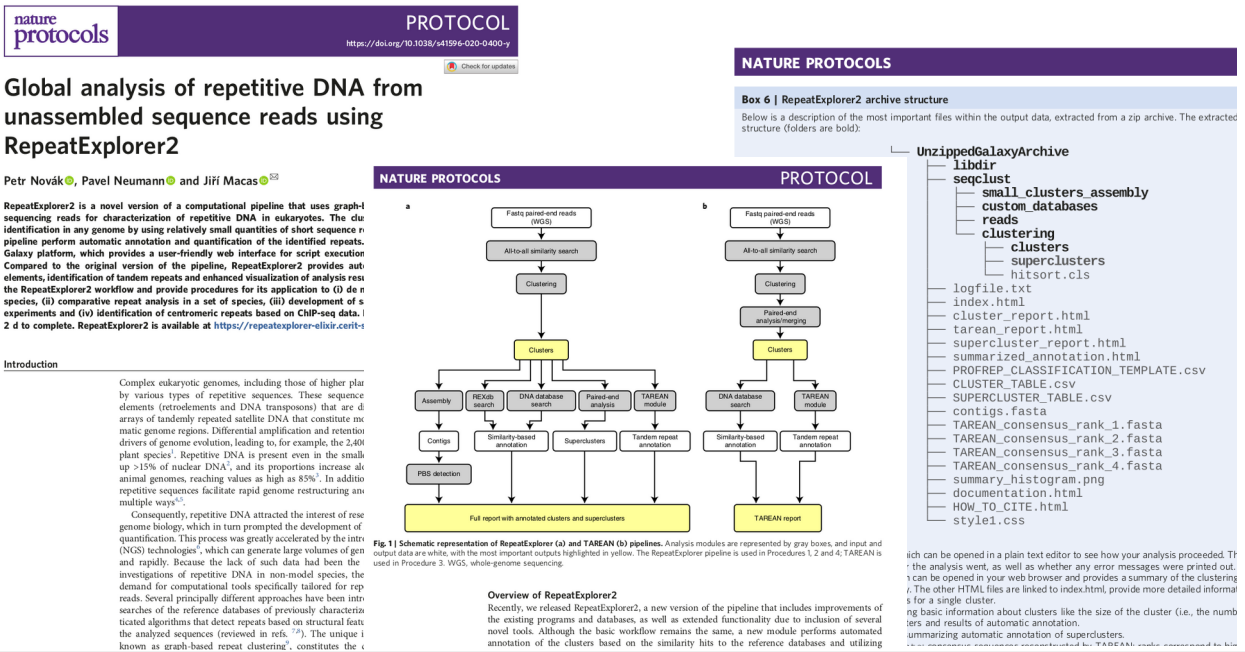
ECCsplorer: a pipeline to detect extrachromosomal circular DNA (eccDNA) from next-generation sequencing data

Ludwig Mann, Kathrin M. Seibt, Beatrice Weber and Tony Heitkam*



Protocols and tutorials

- RepeatExplorer principles and example protocols published
- Corresponding video tutorials available from YouTube



nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-020-0400-y

Check for updates

Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2

Petr Novák, Pavel Neumann and Jiří Macas

RepeatExplorer2 is a novel version of a computational pipeline that uses graph-l sequencing reads for characterization of repetitive DNA in eukaryotes. The clu identification in any genome by using relatively small quantities of short sequence r pipeline perform automatic annotation and quantification of the identified repeats. Galaxy platform, which provides a user-friendly web interface for script execution Compared to the original version of the pipeline, RepeatExplorer2 provides aut elements, identification of tandem repeats and enhanced visualization of analysis resu the RepeatExplorer2 workflow and provide procedures for its application to (i) de n species, (ii) comparative repeat analysis in a set of species, (iii) development of s experiments and (iv) identification of centromeric repeats based on ChIP-seq data. 2 d to complete. RepeatExplorer2 is available at https://repeatexplorer-elixir.cerit-s

Introduction

Complex eukaryotic genomes, including those of higher plar by various types of repetitive sequences. These sequence elements (retroelements and DNA transposons) that are d arrays of tandemly repeated satellite DNA that constitute ma matic genome regions. Differential amplification and retentio drivers of genome evolution, leading to, for example, the 2,40 plant species[1]. Repetitive DNA is present even in the small up >15% of nuclear DNA[3], and its proportions increase ale animal genomes, reaching values as high as 85%[2]. In additio repetitive sequences facilitate rapid genome restructuring an multiple ways[4,5].

Consequently, repetitive DNA attracted the interest of rese genome biology, which in turn prompted the development of quantification. This process was greatly accelerated by the intro (NGS) technologies[6], which can generate large volumes of gen and rapidly. Because the lack of such data had been the investigations of repetitive DNA in non-model species, the demand for computational tools specifically tailored for rep reads. Several principally different approaches have been intr searches of the reference databases of previously characteriz ticated algorithms that detect repeats based on structural feat the analyzed sequences (reviewed in refs. [7,8]). The unique known as graph-based repeat clustering[9], constitutes the

NATURE PROTOCOLS PROTOCOL

Fig. 1 | Schematic representation of RepeatExplorer (a) and TAREAN (b) pipelines. Analysis modules are represented by gray boxes, and input and output data are white, with the most important outputs highlighted in yellow. The RepeatExplorer pipeline is used in Procedures 1, 2 and 4; TAREAN is used in Procedure 3. WGS, whole-genome sequencing.

Overview of RepeatExplorer2

Recently, we released RepeatExplorer2, a new version of the pipeline that includes improvements of the existing programs and databases, as well as extended functionality due to inclusion of several novel tools. Although the basic workflow remains the same, a new module performs automated annotation of the clusters based on the similarity hits to the reference databases and utilizing

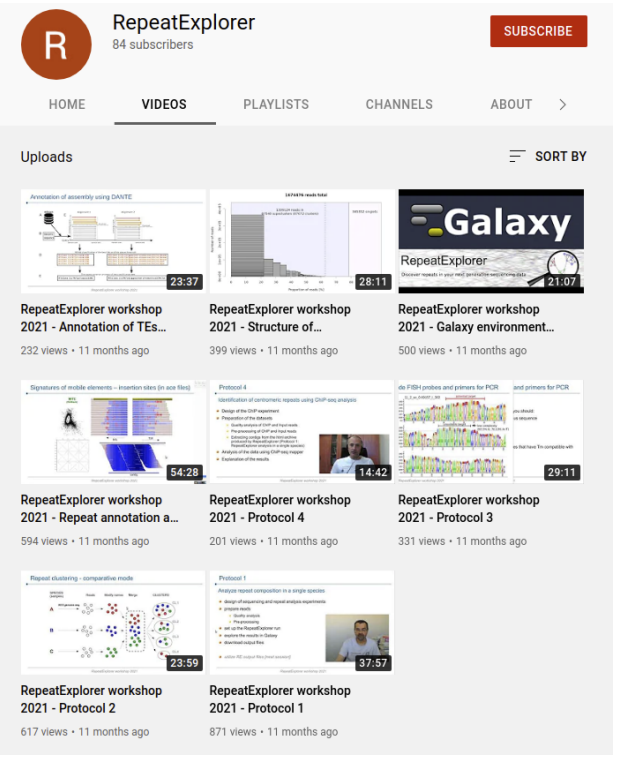
NATURE PROTOCOLS

Box 6 | RepeatExplorer2 archive structure

Below is a description of the most important files within the output data, extracted from a zip archive. The extracted structure (folders are bold):

```
└── UnzippedGalaxyArchive
    ├── libdir
    ├── seqclust
    │   ├── small_clusters_assembly
    │   ├── custom_databases
    │   ├── reads
    │   └── clustering
    │       ├── clusters
    │       ├── superclusters
    │       └── hitsort.cls
    ├── logfile.txt
    ├── index.html
    ├── cluster_report.html
    ├── tarean_report.html
    ├── supercluster_report.html
    ├── summarized_annotation.html
    ├── PROFREP_CLASSIFICATION_TEMPLATE.csv
    ├── CLUSTER_TABLE.csv
    ├── SUPERCLUSTER_TABLE.csv
    ├── contigs.fasta
    ├── TAREAN_consensus_rank_1.fasta
    ├── TAREAN_consensus_rank_2.fasta
    ├── TAREAN_consensus_rank_3.fasta
    ├── TAREAN_consensus_rank_4.fasta
    ├── summary_histogram.png
    ├── documentation.html
    ├── HOW_TO_CITE.html
    └── style1.css
```

which can be opened in a plain text editor to see how your analysis proceeded. Th the analysis went, as well as whether any error messages were printed out. can be opened in your web browser and provides a summary of the clustering The other HTML files are linked to index.html, provide more detailed informa for a single cluster.
ng basic information about clusters like the size of the cluster (i.e., the numb ters and results of automatic annotation.
ummarizing automatic annotation of superclusters.



RepeatExplorer
84 subscribers
SUBSCRIBE
HOME VIDEOS PLAYLISTS CHANNELS ABOUT
Uploads
SORT BY