Welcome!
Principles and history of RepeatExplorer

2007

First paper on repeat clustering from NGS data

BMC Genomics

Research article

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, Braníčská 31, 370 05, Brno, Czech Republic.

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

* Corresponding author

Published: 21 November 2007
Received: 13 August 2007
Accepted: 21 November 2007

doi:10.1186/1471-2164-8-427

RepeatExplorer workshop 2023
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 $\Rightarrow$ 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.
Introduction of graph-based clustering (Novak et al. 2010)

First paper on repeat clustering from NGS data

Graph-based clustering

- Sequence overlaps between the reads are transformed to a graph represented as nodes and their similarities as edges connecting them.
- 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)*

*pairwise alignments* → *graph representation*

- **similarity exceeding threshold** [90% simil. 55% of length]
Principles and history of RepeatExplorer

First paper on repeat clustering from NGS data

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

- 2007...
- 2010...
- 2013
- 2014
- 2016
- 2017
- 2018
- 2019
- 2020...
- 2021
- 2022

RepeatExplorer workshop 2023
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)
- Repeat Explorer in Galaxy (Novak et al. 2013)
- Command-line version
- Public web-based server
## Principles and history of RepeatExplorer

<table>
<thead>
<tr>
<th>Year</th>
<th>Key Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>First paper on repeat clustering from NGS data</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Introduction of graph-based clustering (Novák et al. 2010)</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Repeat Explorer in Galaxy (Novák et al. 2013)</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>TAREAN</td>
</tr>
</tbody>
</table>

**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

*Nucleic Acids Research, 2017*  
doi: 10.1093/nar/gkx257
Principles and history of RepeatExplorer

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>First paper on repeat clustering from NGS data</td>
</tr>
<tr>
<td>2010</td>
<td>Introduction of graph-based clustering (Novak et al. 2010)</td>
</tr>
<tr>
<td>2013</td>
<td>Repeat Explorer in Galaxy (Novak et al. 2013)</td>
</tr>
<tr>
<td>2014</td>
<td>TAREAN</td>
</tr>
<tr>
<td>2016</td>
<td>ELIXIR $$$</td>
</tr>
<tr>
<td>2017</td>
<td>ELIXIR / CERIT Galaxy server + data storage</td>
</tr>
</tbody>
</table>

---

**RepeatExplorer workshop 2023**
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)

TAREAN

ELIXIR $$$

Public web-based server

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>First paper on repeat clustering from NGS data</td>
</tr>
<tr>
<td>2010</td>
<td>Introduction of graph-based clustering (Novak et al. 2010)</td>
</tr>
<tr>
<td>2013</td>
<td>Repeat Explorer in Galaxy (Novak et al. 2013)</td>
</tr>
<tr>
<td>2014</td>
<td>TAREAN</td>
</tr>
<tr>
<td>2016</td>
<td>ELIXIR / CERIT Galaxy server + data storage</td>
</tr>
<tr>
<td>2017</td>
<td>REXdb database (Neumann et al. 2019)</td>
</tr>
</tbody>
</table>

---

![Image](https://via.placeholder.com/150)

**Mobile DNA**

**RESEARCH**

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

---

**RepeatExplorer workshop 2023**
## Principles and history of RepeatExplorer

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>First paper on repeat clustering from NGS data</td>
</tr>
<tr>
<td>2010</td>
<td>Introduction of graph-based clustering (Novak et al. 2010)</td>
</tr>
<tr>
<td>2013</td>
<td>Repeat Explorer in Galaxy (Novak et al. 2013)</td>
</tr>
<tr>
<td>2014</td>
<td>TAREAN</td>
</tr>
<tr>
<td>2016</td>
<td>REXdb (Neumann et al. 2019)</td>
</tr>
<tr>
<td>2017</td>
<td>Repeat Explorer ver. 2</td>
</tr>
</tbody>
</table>

**RESEARCH**

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 Hoftáková and Jiří Macas

---

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

**Decision tree for automatic annotation**
Principles and history of RepeatExplorer

Superclusters provide more complete annotation

RepeatExplorer 2 – automatic detection of superclusters

Decision tree for automatic annotation
Principles and history of RepeatExplorer


RepeatExplorer pipeline
- Fastq paired-end reads (WGS)
- All-to-all similarity search
- Clustering
- Clusters
  - Assembly
  - REXdb search
  - DNA database search
  - Paired-end analysis
  - TAREAN module
  - Contigs
    - Similarity-based annotation
    - Superclusters
    - Tandem repeat annotation
    - PBS detection
- Full report with annotated clusters and superclusters

TAREAN pipeline
- Fastq paired-end reads (WGS)
- All-to-all similarity search
- Clustering
  - Paired-end analysis/merging
- Clusters
  - DNA database search
  - TAREAN module
  - Similarity-based annotation
  - Tandem repeat annotation
  - TAREAN report

Additional tools:
- ChIP-seq Mapper
- Long reads
- DANTE

TAREAN
- REXdb
  (Neumann et al. 2019)
- Repeat Explorer ver. 2

RepeatExplorer workshop 2023
Principles and history of RepeatExplorer

2007 ... 2010 ... 2013  2014 ... 2016  2017  2018  2019  2020  2023

Repeat Explorer

ver. 2

Assembly annotation

DANTE
DANTE-LTR

TideCluster
Applications
Applications

- Repeat composition
  - single species

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

- Repeat composition
  - single species
  - comparative analysis

Identification of chromosome B-specific repeats

Comparative analysis of B+/- plants

Detection of clusters enriched on Bs

Genomic repeat abundances contain phylogenetic signal

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

(Novak et al., 2014)
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:
- Isolation of nuclei, digestion with micrococcal nuclease
- Chromatin immunoprecipitation with CenH3 antibody
- DNA isolation ("ChIP")
- DNA isolation ("INPUT")
- Sequencing
- 9.5 mil. reads
- 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
Applications

- Repeat composition
  - single species
  - comparative analysis

- Repeat clusters as a reference
  - ChIP-seq
  - assembly
Applications

- Repeat composition
  - single species
  - comparative analysis

- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - graph shapes
Applications

- Repeat composition
  - single species
  - comparative analysis

- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - *graph shapes*

- Satellite DNA
  - cytogenetic studies
Applications

- Repeat composition
  - single species
  - comparative analysis

- Repeat clusters as a reference
  - ChIP-seq
  - assembly
  - graph shapes

- Satellite DNA
  - cytogenetic studies
  - diagnostic markers
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
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
Protocols and tutorials

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