Welcome !
## Program at a glance

### Tuesday (May 24)

**Introduction and update on RepeatExplorer**
- Principles and history of RepeatExplorer  
  (*J. Macas*)
- Diagnostic features of repetitive elements – Part I  
  (*P. Neumann*)
- Coffee break
- Diagnostic features of repetitive elements – Part II  
  (*P. Neumann*)
- RepeatExplorer tools for genome annotation  
  (*P. Novák*)

**Lunch**

**Practical training I**
- Basic Protocols 1-4: troubleshooting & advanced data analysis

**Dinner at CITYgastro restaurant**

### Wednesday (May 25)

**Short talks**
- Nicola Schmidt
- Ludwig Mann
- Zirlane Portugal da Costa

**Coffee break**

**Practical training II**
- using RE output for annotating genome assemblies
- REXdb and DANTE
- structure-based annotation of complete LTR-retrotransposons

### Thursday (May 26)

**Short talks**
- Matej Lexa
- Monika Čechová
- Sophie Maiwald
- Vratislav Peška

**Coffee break**

**Practical training III**
- reconstructing phylogenetic relationships
- local installation and running RE tools from a command line
- topics proposed by the participants & individual consultations

**Lunch**
Protocols and tutorials

- RepeatExplorer principles and example protocols published
- Corresponding video tutorials available from YouTube
Principles and history of RepeatExplorer
**Principles and history of RepeatExplorer**

**First paper on repeat clustering from NGS data**

- **2007**

**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, 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


Received: 13 August 2007

Accepted: 21 November 2007

**RepeatExplorer workshop 2022**
Principles and history of RepeatExplorer

First paper on repeat clustering from NGS data

2007

- Shotgun genomic sequencing
- Identification of sequence clusters
- Reconstruction of repetitive elements

CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

Proportions of various repeat types in a genome

Ty3/gypsy
Ty1/copia
Low-copy repeats

RepeatExplorer workshop 2022
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

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.

First paper on repeat clustering from NGS data

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

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

First paper on repeat clustering from NGS data
Principles and history of RepeatExplorer

2007 ... 2010

- First paper on repeat clustering from NGS data
- Introduction of graph-based clustering (Novak et al. 2010)

command-line version

FIRST WORKSHOP!
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 Range</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007 … 2010</td>
<td>First paper on repeat clustering from NGS data</td>
</tr>
<tr>
<td>2010 … 2013</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 … 2016</td>
<td>Public web-based server</td>
</tr>
</tbody>
</table>

ELIXIR $$$
Principles and history of RepeatExplorer


First paper on repeat clustering from NGS data
Introduction of graph-based clustering (Novák et al. 2010)
Repeat Explorer in Galaxy (Novák et al. 2013)

TAREAN

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

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

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

RepeatExplorer workshop 2022
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

TAREAN

ELIXIR $$$

ELIXIR / CERIT Galaxy server + data storage

RepeatExplorer

Discover repeats in your next-generation sequencing data

RepeatExplorer includes utilities for graph-based clustering and characterization of repetitive sequences in next-generation sequencing data and tools for the detection of transposable element protein coding contents.

The use of software RepeatExplorer is subject to the following terms of use: you must not use the software for any purpose other than to research the subject matter of the software. You must not use the software for any purpose that is not authorized by the software provider. You must not use the software for any purpose that is not authorized by the software publisher. You must not use the software for any purpose that is not authorized by the software distributor. You must not use the software for any purpose that is not authorized by the software developer. You must not use the software for any purpose that is not authorized by the software owner. You must not use the software for any purpose that is not authorized by the software author. You must not use the software for any purpose that is not authorized by the software creator. You must not use the software for any purpose that is not authorized by the software creator.

RepeatExplorer workshop 2022
# 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>
<tr>
<td>2022</td>
<td>Public web-based server</td>
</tr>
</tbody>
</table>

---

**Please, acknowledge ELIXIR in your publications!**

---

*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.*
Principles and history of RepeatExplorer

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 / CERIT Galaxy server + data storage

REXdb database (Neumann et al. 2019)

Mobile DNA

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 Hoffáková and Jiří Macas
# 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>

## Introduction

- **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)

## Decision tree for automatic annotation

- TAREAN, REXdb, DNA database
- TAREAN
- REXdb
- REXdb, PRG/LTR
- DNA database

## 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

Principles and history of RepeatExplorer

Superclusters provide more complete annotation

RepeatExplorer 2 – automatic detection of superclusters

Decision tree for automatic annotation

RepeatExplorer workshop 2022
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
  - Full report with annotated clusters and superclusters

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

TAREAN
(REXdb (Neumann et al. 2019))
Repeat Explorer ver. 2
RepeatExplorer workshop 2022

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

TAREAN

REXdb

(Neumann et al. 2019)

Repeat Explorer ver. 2

Additional tools:

• ChIP-seq Mapper
  • DANTE
  • Long reads
  • Assembly annotation

RepeatExplorer workshop 2022
Applications

- Repeat composition
  - single species

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

RepeatExplorer workshop 2022
Applications

- Repeat composition
  - single species
  - comparative analysis

Genomic repeat abundances contain phylogenetic signal
Dodsworth et al. (2015)
Syst. Biol. 64(1): 112-126

RepeatExplorer workshop 2022
Applications

- Repeat composition
  - single species
  - comparative analysis

Identification of chromosome B-specific repeats

Comprehensive analysis of B+/- plants

Detection of clusters enriched on Bs

Genomic repeat abundances contain phylogenetic signal


(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
- Illumina seq.
- Mapping reads to clusters of long (>100nt) reads, calculating ratio of ChIP / INPUT
- Cluster = repeat family

Satellites

TR_11
PisTR-B

Neumann et al. (2012) PLoS Genetics e1002777
Applications

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

Assessing completeness of genome assemblies
Applications

- Repeat composition
  - single species
  - comparative analysis

- Repeat clusters as a reference
  - ChIP-seq
  - assembly
    - reference databases for repeat annotation
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

SOFTWARE

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

Ludwig Mann, Kathrin M. Seib, Beatrice Weber, and Tony Herkaint
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

Enjoy the workshop!