RepeatExplorer pipeline

RepeatExplorer

Discover repeats in your next generation sequencing data
What is RepeatExplorer?

Implementation of principles described in:

- Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula* (BMC Genomics 2007, 8:427)

- Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data (BMC Bioinformatics 2010, 11:378)


Available Tools:

- NGS data preprocessing
- RepeatExplorer2 pipeline
- TAREAN pipeline
- Chip-Seq analysis
- Domain based ANnotation of Transposable Element – DANTE
- Profrep
- Visualization

Protocols

Principle of RepeatExplorer

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

Cluster 1

Cluster 2

Cluster 3

interspersed repeats  tandem repeat  sequence reads
Principle of RepeatExplorer

**Cluster 1**

**Cluster 2**

**Cluster 3**

CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY
Graph Based Representation of Sequence Reads

(edges, vertices, nodes)
Graph Based Representation of Sequence Reads

pairwise alignments

[diagram showing pairwise alignments for read_1 to read_4]

graph representation

[diagram showing graph representation with nodes and edges indicating similarity exceeding threshold (90% simil. 55% of length)]
Principle of RepeatExplorer

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

Clusters as connected components:
- Cluster 1
- Cluster 2
- Cluster 3

interspersed repeats  tandem repeat  sequence reads
Principle of RepeatExplorer

Shotgun genomic sequencing

Identification of sequence clusters

Reconstruction of repetitive elements

clusters as communities

interspersed repeats  tandem repeat  sequence reads
A **community**, with respect to graphs, can be defined as a subset of nodes that are densely connected to each other and loosely connected to the nodes in the other communities in the same graph.

*community ~ cluster ~ repeat family*
Graph Based Representation of Sequence Reads

- Informative graphical representation
- Graph layout
- Vertex coloring
RepeatExplorer Pipeline

Preprocessing
- NGS short reads
- Quality control
- Trimming
- Adapter removing

Prerun
- High quality short reads
- Estimating Genome repetitiveness and data size limit
- all reads or subset

All-to-All Comparison
- mgblast
- HSP filtering

Clustering
- Sequence Graph

Clusters
RepeatExplorer Pipeline

- Paired-end analysis
- Superclusters
- Assembly
- Contigs
- Tandem repeats annotation
- Results synthesis
- Similarity based annotation
- Comparing with known repeats
- Built in database
- Custom database
- Report

RepeatExplorer workshop 2021
RepeatExplorer pipeline

Input data
- Short reads – hundreds of nt
- Paired-end, interleaved
- Single-end
- Pre-processed
- Uniform length
- FASTA format

Preprocessing
- Quality control
- Trimming
- Adapter removing

High quality short reads

Prerun
- Estimating data size limit
- All reads or subset

All-to-All comparison
- mgblast hit filtering

Sequence Graph

Clustering

Clusters
Preprocessing

RepeatExplorer utilities:

Preprocessing of FASTQ reads

1. Trimming (optional)
2. Filter by quality
4. Cutadapt filtering
5. Discard single reads, keep complete pairs
6. Sampling (optional)
7. Interlacing
RepeatExplorer pipeline

Preprocessing
- short reads
- Quality control
- Trimming
- Adapter removing

High quality short reads

Prerun
- Estimating data size limit
- all reads or subset

All-to-All comparison
- mgblast
- hit filtering

Clustering
- Sequence Graph

Clusters

RepeatExplorer workshop 2021
Prerun

All-to-all sequence comparison on small sample of input data

\[ N = 10 \]
\[ E = 12 \]
\[ k_g = 0.27 \]

\[ N = 10 \]
\[ E = 45 \]
\[ k_g = 1.0 \]

**Graph density -** $k_g$ is genome specific coefficient and depends on the repetitive content and genome size.

Density corresponds to probability that two randomly taken sequences from genome will be similar.

$k_g$ is used to estimate maximum number of processable reads:

\[ k_g = \frac{2E}{N(N-1)} \]
**Prerun**

All-to-all sequence comparison on small sample of input data

All-to-all sequence comparison on small sample of NGS reads

\[ k_g = \frac{2E}{N(N-1)} \]

\(N\) .. 20,000 sample reads
\(E\) .. number of identified similarity hits

\(k_g\) is used to estimate maximum number of reads \(N_{max}\) providing that we can process with available RAM (M)

\[ N_{max} = \sqrt{\frac{M}{k_g}} \]
All-to-all sequence comparison on small sample of input data

\[ k_g = \frac{2E}{N(N-1)} \]

\( N \) .. 20,000 sample reads

\( E \) .. number of identified similarity hits

\( k_g \) is used to estimate maximum number of reads \( N_{\text{max}} \) providing that we can process with available RAM (M)

\[ N_{\text{max}} = \sqrt{m \frac{M}{k_g}} \]

### Table: Species and Read Information

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of reads</th>
<th>Genome Size (1C)</th>
<th>Coverage [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musa acuminata</td>
<td>3,046,164</td>
<td>623 Mbp</td>
<td>48.9</td>
</tr>
<tr>
<td>Lasiurus borealis</td>
<td>4,256,140</td>
<td>2,526 Mbp</td>
<td>16.8</td>
</tr>
<tr>
<td>Pisum sativum</td>
<td>3,011,839</td>
<td>4,300 Mbp</td>
<td>7.0</td>
</tr>
<tr>
<td>Vicia panonica</td>
<td>1,039,442</td>
<td>5,730 Mbp</td>
<td>1.8</td>
</tr>
<tr>
<td>Silene latifolia</td>
<td>2,943,062</td>
<td>5,850 Mbp</td>
<td>5.0</td>
</tr>
<tr>
<td>Secale cereale</td>
<td>1,899,753</td>
<td>7,917 Mbp</td>
<td>2.4</td>
</tr>
<tr>
<td>Lathyrus latifolius</td>
<td>1,464,940</td>
<td>9,980 Mbp</td>
<td>1.5</td>
</tr>
<tr>
<td>Fritilaria imperialis</td>
<td>12,220,382</td>
<td>42,400 Mbp</td>
<td>2.9</td>
</tr>
<tr>
<td><strong>Fritilaria affinis</strong></td>
<td><strong>1,168,248</strong></td>
<td><strong>45,000 Mbp</strong></td>
<td><strong>0.3</strong></td>
</tr>
</tbody>
</table>

Number of reads which can be processed with 16GB RAM
Prerun

All-to-all sequence comparison on small sample of input data

graph density = 1
Prerun

All-to-all sequence comparison on small sample of input data

- Graph density = 1
- Graph density = 0.54
Prerun

All-to-all sequence comparison on small sample of input data

Number of reads (Vertices) 44,772
Number of similarity hits (Edges) 542,348,907

Input data (All reads) 2,000,000
Total number of similarity hits 1,394,970,205

Approx 1/3 of stored similarity hits originate from satellite which represent only 2% of genome
Prerun

Satellite filtering (optional)

- Clusters composed from satellite reads can be scaled down without losing information.
- Sample of 10% of reads of is kept in analysis to keep track of this satellite
RepeatExplorer pipeline

- **Preprocessing**
  - short reads
  - Quality control
  - Trimming
  - Adapter removing

- **Prerun**
  - High quality short reads
  - Estimating data size limit

- **All-to-All Comparison**
  - all reads or subset
  - mgblast
  - hit filtering

- **Clustering**
  - Sequence Graph
  - Clusters
All-to-all comparison

- Similarity search using **mgblast**
- Default threshold:
  - overlap: 55 nt and 55% of the length
  - minimal similarity 90%
- By default mgblast is using **DustMasker** (low complexity repeat filter)
  - simple repeats are underestimated or not detected (e.g. telomeric motifs, microsatellites)
  - Masking of low complexity can be disabled → long running time and increased memory usage

≥ 55 nt
≥ 55% of read length
All-to-all comparison

- Similarity search using **mgblast**
- Default threshold:
  - overlap: 55 nt and 55% of the length
  - minimal similarity 90%
- By default mgblast is using **DustMasker** (low complexity repeat filter)
  - simple repeats are underestimated or not detected (e.g. telomeric motifs, microsatellites)
  - Masking of low complexity can be disabled → long running time and increased memory usage
- Adapters in sequence can slow down all-to-all search
RepeatExplorer pipeline

- Graph is divided into subgraphs (clusters/communities)
- Clusters have dense connections between the nodes within the clusters but sparse connections between nodes in different clusters
RepeatExplorer pipeline

Clustering

- Clusters
- Top clusters
- Small clusters
- Singlets
RepeatExplorer pipeline

Cluster centered analysis

- All-to-All Comparison
- Clustering
  - Sequence Graph
  - mgblast hit filtering

- Clusters
  - Paired-end analysis
  - Superclusters
  - Assembly
  - Contigs
  - TANDER
  - Tandem repeats annotation
  - Comparing with known repeats
    - Built in database
    - Custom database
  - Similarity based annotation

- Results synthesis
- Report

RepeatExplorer workshop 2021
RepeatExplorer pipeline

- Paired-end analysis
- Clusters
  - Assembly
  - TAREAN
  - Comparing with known repeats
    - Built in database
    - Custom database
- Superclusters
  - Contigs
  - Tandem repeats annotation
  - Similarity based annotation
- Results synthesis
- Report
  - Superclusters
  - Paired-end analysis
Clusters and Superclusters

Sometimes (often) reads which belong to single repeat family are split into multiple clusters.
Clusters and Superclusters

Sometimes (often) reads which belong to single repeat family are split into multiple clusters.

We need to identify such false splits.

Supercluster
Clusters and Superclusters

Identification of supercluster using paired-end reads

\[ k_{x,y} = \frac{2W}{n_x + n_y} \]

\( W \) number of reads pairs shared between clusters \( x \) and \( y \)
\( n_x \) and \( n_y \) is number of reads in cluster \( x \) and cluster \( y \) with absent read mate within the same cluster respectively

Suitable \( k_{x,y} \) cutoff 0.05 – 0.2
- full connection: \( k_{x,y} = 1 \)
- no connection: \( k_{x,y} = 0 \)
Clusters and Superclusters

Identification of supercluster using paired-end reads

 paired-end read analysis
RepeatExplorer pipeline

- Paired-end analysis
- Superclusters
- Assembly
- Contigs
- Results synthesis
- Tandem repeats annotation
- Clusters
- Comparing with known repeats
  - Built in database
  - Custom database
- Similarity based annotation
- Report

RepeatExplorer workshop 2021
Tandem Repeat Analyzer - TAREAN

TAREAN calculates **graph layout** and provide automatic analysis of **graph topology** with the aim to identify **tandem repeats**
Tandem Repeat Analyzer - TAREAN

Principle

- All reads within cluster are set to have same 3’to5’ orientation with hypothetical tandem repeat monomer
- Directed graph is constructed
- Larges circular structures are detected (a.k.a strongly connected component)
- Connected Component index
In clusters which are derived from tandem repeat, most of the paired-end reads should be complete.

Pair completeness = fraction of complete pairs in cluster
Tandem Repeat Analyzer - TAREAN

**Principle**

- **Putative satellite (high confidence)**
  
  High P and C score

- **Putative satellite (low confidence)**
  
  P and C score lower

- **Putative LTR element**
  
  Primer binding site detected, presence of long ORF

- **rDNA**
  
  Tandem organization + similarity to known rDNA sequences

- **Other clusters**
  
  No tandem repeat like structure
RepeatExplorer pipeline

Principle

Reconstruction of tandem repeat monomer

- k-mer based approach
- multiple variants reported
- sorted based on significance

TAREAN limitation

- paired end reads required
- limited sensitivity to TR with very short monomer
RepeatExplorer pipeline

Clusters
- Paired-end analysis
- Assembly
- Comparing with known repeats
- Integration into built-in and custom databases

Contigs
- TAREAN
- Tandem repeats annotation
- Similarity-based annotation

Superclusters
- Results synthesis
- Report

Paired-end analysis
- Contigs
- Results synthesis
- Report
Contig assembly

**Reads** are assembled by CAP3 program, each cluster separately:

\[
\begin{align*}
\text{ACTGTGTCGTCGTCGTCGTGTG} \\
\text{CGTCGTCG - CGTGTGGGT} \\
\text{GTCGTGTG - TTGTCGTCTGA} \\
\text{ACTGTGTCGTCGTCGTCGTCGTGTGGTTGTCGTCTGA}
\end{align*}
\]

High confidence putative satellite clusters are not assembled by CAP3, instead TAREAN generate **k-mer based** consensus:

\[
\begin{align*}
\text{TATAACACATCCATTATTTAAGACAAATGTTGCAATTTCTTTGTAAATTTTACCCATTTCTTAATATAATGCGTTAAAAACATGTATACAATACCACTTTTT} \\
\text{AGCAATATGTGCTACATTATTTAAGACATTTACCCATTTTTCAATATTACCTGAAAAACAT}
\end{align*}
\]
RepeatExplorer pipeline

All **reads** are compared with:

- Database of protein domains (REXdb)
- DNA database
  - rDNA, tRNA
  - Organelle DNA
  - potential contaminants
- Custom database (optional)
### Similarity search

<table>
<thead>
<tr>
<th>Database sequence classification</th>
<th>Protein domain</th>
<th>Number of reads with similarity hit</th>
<th>Proportion No of reads / cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td>mitochondria</td>
<td></td>
<td>25</td>
<td>0.0023</td>
</tr>
<tr>
<td>Ogre Ty3-RH</td>
<td>Ty3-RH</td>
<td>2977</td>
<td>0.27402</td>
</tr>
<tr>
<td>Retand Ty3-RH</td>
<td>Ty3-RH</td>
<td>2</td>
<td>0.00018</td>
</tr>
<tr>
<td>Ogre Ty3-RT</td>
<td>Ty3-RT</td>
<td>3473</td>
<td>0.31968</td>
</tr>
<tr>
<td>Ogre Ty3-aRH</td>
<td>Ty3-aRH</td>
<td>1713</td>
<td>0.15768</td>
</tr>
</tbody>
</table>
RepeatExplorer pipeline

Reporting:

- HTML reports
- Visualization
- Automatic classification
### Reporting

#### Table

<table>
<thead>
<tr>
<th>nhits</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3181</td>
</tr>
<tr>
<td>*-repeat</td>
<td>3181</td>
</tr>
<tr>
<td>*-mobile element</td>
<td>3181</td>
</tr>
<tr>
<td>*-Class I</td>
<td>3181</td>
</tr>
<tr>
<td>*-LTR</td>
<td>3181</td>
</tr>
<tr>
<td>*-Tyl_copia</td>
<td>3181</td>
</tr>
<tr>
<td>-Ate</td>
<td>64</td>
</tr>
<tr>
<td>-Alessa</td>
<td>5</td>
</tr>
<tr>
<td>-Angela</td>
<td>1</td>
</tr>
<tr>
<td>-Bianca</td>
<td>1</td>
</tr>
<tr>
<td>-Bryco</td>
<td>14</td>
</tr>
<tr>
<td>-Gynmo-I</td>
<td>1</td>
</tr>
<tr>
<td>-Gynmo-II</td>
<td>3</td>
</tr>
<tr>
<td>-Ikeros</td>
<td>3</td>
</tr>
<tr>
<td>-Ivana</td>
<td>3962</td>
</tr>
<tr>
<td>-SIRE</td>
<td>8</td>
</tr>
<tr>
<td>-TAR</td>
<td>4</td>
</tr>
</tbody>
</table>

Ivana

870

CL6

3021

CL5

3572

RepeatExplorer workshop 2021
Automatic annotation

- Ali
  - contamination
  - organelle
    - plastic
    - mitochondria
  - repeat
    - rDNA
    - satellite
    - mobile_element
      - 45S_rDNA
      - 5S_rDNA
      - Class_I
        - pararetrovirus
        - DIRS
        - Penelope
        - LINE
      - 18S_rDNA
      - 25S_rDNA
      - 5.8S_rDNA
      - SINE
      - LTR
  - Ty3_gypsy
    - chromovirus
      - MITE
      - EnSpm_CACTA
      - hAT
      - Kol
    - non-chromovirus
      - nonchromo-outgroup
      - Phgy
      - Selgy
      - OTA
      - Chlamyvir
      - Tcn1
      - CRM
      - Galadriel
      - Tekay
      - Reina
      - chomo-outgroup
  - Ty1_copia
    - Ikeros
    - Ivana
    - Osse
    - SIRE
    - TAR
    - Tork
    - Ty1-outgroup
    - non-chromovirus
Automatic annotation
Automatic annotation

- Ali
  - contamination
  - organelle
    - plastid
  - repeat
    - rDNA
      - 45S_rDNA
      - 5S_rDNA
    - satellite
      - Class_I
        - pararetrovirus
      - DIRS
      - Penelope
      - LINE
    - LTR
  - mobile_element
  - 5.8S_rDNA
  - SINE

- Ty1_copia
  - Ikeros
  - Ivana
  - Osse
  - SIRE
  - TAR
  - Tork
  - Ty1-outgroup
    - non-chromovirus
      - nonchromo-outgroup
      - Phygy
      - Selgy
      - OTA
      - Chlamyvir
      - Tcn1
      - CRM
      - Galadriel
      - Tekay
      - Reina
  - chromovirus
    - MITE
    - EnSpm_CACTA
    - hAT
    - Koh
  - Athila
  - Tat
  - Retand

- Ty3_gypsy

- Second best hit
- Best hit
Automatic annotation

Lowest common ancestor

Second best hit

Best hit
Automatic annotation
Automatic annotation

Spurious hits
## Automatic annotation

<table>
<thead>
<tr>
<th>Repeat Type</th>
<th>Proportion [%]</th>
<th>Ns@cl.</th>
<th>Nclusters</th>
<th>Nreads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclassified_repeat (conflicting evidences)</td>
<td>4.06</td>
<td>2</td>
<td>5</td>
<td>67995</td>
</tr>
<tr>
<td>-rDNA</td>
<td>0.29</td>
<td>2</td>
<td>4</td>
<td>4823</td>
</tr>
<tr>
<td></td>
<td>-18S_rDNA</td>
<td>0.04</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-25S_rDNA</td>
<td>0.02</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-5.8S_rDNA</td>
<td>0.12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>satellite</td>
<td>8.78</td>
<td>33</td>
<td>33</td>
<td>147033</td>
</tr>
<tr>
<td>-mobile_element</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Class I</td>
<td>0.77</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-SINE</td>
<td>0.14</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-LYTR</td>
<td>0.18</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Gymco-III</td>
<td>0.26</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Gymco-I</td>
<td>0.36</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Tyr1-outgroup</td>
<td>9.57</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Ty3_gypsy</td>
<td>0.26</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-non-chromovirus</td>
<td>0.36</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-non-chrom-outgroup</td>
<td>9.57</td>
</tr>
</tbody>
</table>

RepeatExplorer workshop 2021
Automatic annotation

ONE DOES NOT SIMPLY ANNOTATE REPEATOME
RepeaExplorer

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

TAREAN

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

RepeatExplorer workshop 2021
RepeatExplorer Related Tools

- **DANTE** – Domain based **AN**notation of **T**ransposable **E**lements
  - assembly annotation using REXdb
  - same TE classification system as RepeatExplorer based on REXdb
- **Profrep**
  - assembly annotation based on RE results
- **ChIP-Seq Mapper**
  - Identification of repeats associated with CENH3 or with epigenetic marks
Availability

RepeatExplorer Galaxy Server

https://repeatexplorer-elixir.cerit-sc.cz/
regalaxy@rt.cesnet.cz

Support:

Martina Macháč
Zdeněk Salvet
Miroslav Ruda
Ivana Křenková
Availability

Command line tools

https://bitbucket.org/repeatexplorer/  ChIP-Seq Mapper, RepeatExplorer utilities
https://bitbucket.org/petrnovak/repex_tarean  RepeatExplore with TAREAN
https://github.com/kavonrtep/dante  DANTE
https://github.com/kavonrtep/SeqGrapheR/  SeqGrapheR

Contributors:

Jiri Macas  Georg Hermanutz
Pavel Neumann  Nina Hostakova
Jaroslav Steinhaisel  Tihana Vodrak
Jiri Pech  Petr Novak
Karsten Klein
Thank you!

Questions?