7th Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis in Plants

May 22-24, 2018

Institute of Plant Molecular Biology, České Budějovice, Czech Republic
RepeatExplorer Server

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
- Graph-based clustering
  - Characterization of repeats:
    - Identification, Annotation, Quantification
- Satellite identification
- Chip-Seq analysis
- Domain based ANnotation of Transposable Element – DANTE
- Profrep

Contributors:
- Jiri Macas
- Pavel Neumann
- Jaroslav Steinhaisel
- Jiri Pech
- Karsten Klein
- Georg Hermanutz
- Nina Hostakova
- Tihana Vodrak
- Petr Novak
Graph Based Representation of Sequence Reads

What is Graph?

(edges)

(vertices, nodes)
Graph Based Representation of Sequence Reads

There are two approaches how to use a graph to describe and analyze sequences reads:

- **Overlap-Layout-Consensus**
- **De-Bruijn Graph**
Graph Based Representation of Sequence Reads

There are two approaches how to use a graph to describe and analyze sequences reads:

- **Overlap-Layout-Consensus**
- **De-Bruijn Graph**
Graph Based Representation of Sequence Reads

The are two approaches how to use a graph to describe and analyze sequences reads:

- **Overlap-Layout-Consensus**
- **De-Bruijn Graph**

Sequence read:

```
AAAGCTCAGTTTTCGAGCCAGAGACCCACGAAAGTGTGGGAGCCTTACAGCGCAACTTCAGCAAGAGCGGAG
```

```
AAAGCTCAGTT
AAGCTCAGTTT
AGCTCAGTTTC
GCTCAGTTTCG
CTCAGTTTCGA
TCAGTTTCGAG
CAGTTTCGAGC
AGT TTCGAGCC
...........
```
Graph Based Representation of Sequence Reads

Why to use graph representations:

The are number of available algorithms for graph analysis

- Robust partitioning/classification of reads based on mutual similarities
- Informative graphical representation (layouts)
- Path in graph can be converted to contigs
Graph Based Representation of Sequence Reads

Why to use graph representations:

There are number of available algorithms for graph analysis:
- Robust partitioning/classification of reads based on mutual similarities
- Informative graphical representation (layouts)
- Path in graph can be converted to contigs
Graph Based Representation of Sequence Reads

Why to use graph representations:

There are a number of available algorithms for graph analysis:

- Robust partitioning/classification of reads based on mutual similarities
- Informative graphical representation (layouts)
- Path in graph can be converted to contigs
Graph Based Representation of Sequence Reads

Why to use graph representations:

The are number of available algorithms for graph analysis

- Robust partitioning/classification of reads based on mutual similarities
- Informative graphical representation (layouts)
- Path in graph can be converted to contigs

```
sequence variant
GCCCA → CCCAC → CCACA → CACAG → ACAGG
GCCCT → CCCTC → CCTCA → TCTAG → TCAGG

Insertion/Deletion
CTCTG → CCTGC → CTGGT → TGCTG → GCTGG
CTCTC → CCTCT
CTCTT → TCTTG → CTTGT → TTGTG
CTTGG → TTGGT → TGGTC
AGGTC → GTGTC
```

“bubbles” → sequence variants  shorter paths → deletions
RepeatExplorer workflow

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

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

All-to-All Comparison
- mgblast hit filtering

Clustering
- Sequence Graph
- Clusters
RepeatExplorer workflow

- Paired-end analysis
- Superclusters
- Assembly
- Contigs
- TAREAN
- Tandem repeats annotation
- Comparing with known repeats
- Similarity based annotation
- Result synthesis
- Report

Built in database
Custom database
Preprocessing

Short reads
- Quality control
- Trimming
- Adapter removing

High quality short reads

Prerun
- All reads or subset
- Estimating data size limit

All-to-All comparison
- mgblast
- Hit filtering

Clustering
- Sequence Graph
- Clusters
Preprocessing

RepeatExplorer operates under GIGO principle:

Garbage In

RepeatExplorer

Garbage Out
Preprocessing

- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identifiers
Preprocessing

- **Quality control**
  - Trimming, filtering, adapter removing
  - Convert fastq to fasta
  - Interlacing, sampling
  - Modification of sequence identifiers

**FastQC program**

- Galaxy
- GUI based
- Command line

[Graph showing quality scores across all bases (Illumina >v1.3 encoding)]

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>fastq_data.fastq</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Encoding</td>
<td>Sanger / Illumina 1.9</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>16772669</td>
</tr>
<tr>
<td>Sequence length</td>
<td>75</td>
</tr>
<tr>
<td>GC</td>
<td>45</td>
</tr>
</tbody>
</table>

Bioinformatics Group at the Babraham Institute.
Preprocessing

- **Quality control**
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identifiers

**Dotter** - graphical dotplot program for detailed comparison of two sequences
Preprocessing

- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identifiers

Visual inspection!
Preprocessing

- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identifiers

Tool:
Preprocessing of fastq paired-reads

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

Default filtering:

- quality cut-off: 10
- percent above cut-off: 95%
- Maximum Ns: 0
Preprocessing

- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identifiers

Tool:
Preprocessing of fastq paired-reads

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

cutadapt
https://cutadapt.readthedocs.io

A tool that removes adapter sequences from high throughput sequence reads

- Sequence trimming
  or
- Complete removal

Adapter presence greatly affect all-to-all comparison computational time and contig assembly
Preprocessing

- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identifiers

Tool:
Preprocessing of fastq paired-reads

1.Trimming (optional)
2.Filter by quality
3.Discard single reads, keep complete pairs
4.Cutadapt filtering
5.Discard single reads, keep complete pairs
6.Sampling (optional)
7.Interlacing two fasta files
Preprocessing

- Quality control
-Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling

**Modification of sequence identifiers**

Tool: affixer

**Comparative analysis:**

<table>
<thead>
<tr>
<th>Genome AB</th>
<th>Genome XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;FUXEH4V01EAG68.... acgacagctgactaatgc &gt;FUXEH4V01BKPDK cttcgaggctacacgagct &gt;FUXEH4V01AJAJV actatcgacactgccggcg cg</td>
<td>&gt;F0X5OLU02GZ8YF.... gccccgtcgccgtccgtgtgc &gt;F0X5OLU02I1AMY tgtgtgccccgtctgcgcgcgcgcgccccc &gt;F0X5OLU02HYN8U atatgctatgcgcgc</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AB1</th>
<th>AB2</th>
<th>AB3</th>
</tr>
</thead>
<tbody>
<tr>
<td>acgacagctgactaatgc</td>
<td>cttcgaggctacacgagct</td>
<td>Actatcgacactgccggcg cg</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>XY1</th>
<th>XY2</th>
<th>XY3</th>
</tr>
</thead>
<tbody>
<tr>
<td>gccccgtcgccgtccgtgtgc</td>
<td>tgtgtgccccgtctgcgcgcgcgcgc</td>
<td>-</td>
</tr>
</tbody>
</table>
Prerun

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

Prerun:
- High quality short reads
  - Estimating data size limit
- all reads or subset

All-to-All comparison:
- mgblast
  - hit filtering

Clustering:
- Sequence Graph
  - Clusters

High quality short reads
Prerun: all-to-all sequence comparison on small sample of NGS reads

\[ N = 10 \]
\[ E = 12 \]
\[ k_g = 0.27 \]
\[ N = 10 \]
\[ E = 45 \]
\[ k_g = 1.0 \]

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

- **Graph density** depends on 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.
Prerun: all-to-all sequence comparison on small sample of NGS reads

Number of reads which can be processed with 16GB RAM in various plant species
Pre-clustering analysis

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

\( k_g \) genome specific coefficient - **graph density**  
depends on 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 reads providing that **we can process ~ 340 \cdot 10^6** of similarity hits on machine with 16GB of RAM
Prerun – optional filtering of abundant satellite sequences

Example of dense satellite cluster:

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

Density 0.54

Approx 1/3 of stored similarity hits originate from satellite which represent approx 2% of genome
Prerun – optional filtering of abundant satellite sequences

Example of dense satellite cluster:

Such clusters can be filtered out from the clustering

**Filtering criteria:**
- cluster must be classified by TAREAN as satellite
- cluster consist of at least 1000 reads
- reads in cluster generate at least 3% of total similarity hits

However, 10% of the reads of affected clusters is kept in the analysis – to keep track of such clusters
All-to-All Comparison

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

High quality short reads

Prerun
- Estimating data size limit

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

Sequence Graph

Clustering

Clusters
All-to-All Comparison

All pairs of reads with similarity above threshold are found using megablast:

**Default threshold:**
Minimal overlap: 55 nt and 55% of length of shorter sequence
**Minimal similarity:** 90%

Stringency affects maximum number of reads!

**Length of overlap** must be adjusted for reads shorter than 100 nt

**Alternative threshold – Illumina short:**
Minimal overlap 20 nt and 40% of length, minimal similarity :90%
All-to-All Comparison

All pairs of reads with similarity above threshold are found using megablast:

**Default threshold:**
- Minimal overlap: 55 bp and 55% of length of shorter sequence
- **Minimal similarity:** 90%

Stringency affects maximum number of reads!

**Length of overlap** must be adjusted for reads shorter than 100 nt

Presence of unfiltered adapter sequence does not pass similarity threshold, but all-to-all comparison becomes extremely slow!
All pairs of reads with similarity above threshold are found using megablast:

**Default threshold:**
Minimal overlap: 55 bp and 55% of length of shorter sequence  
**Minimal similarity:** 90%

Stringency affects maximum number of reads!

**Length of overlap** must be adjusted for reads shorter than 100 nt

Low complexity repeat – **DustMasker**

Simple repeats are underestimated or not detected at all: e.g.
- Telomeric motifs
- Microsatellites
...
Clustering

**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
Clustering

- Graph is divided into subgraphs (clusters/communities)
- Quality of division is measured using **modularity**
- Modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random
- Clusters have dense connections between the nodes within the clusters but sparse connections between nodes in different clusters
Cluster centered analysis

Clusters
- Paired-end analysis
- Assembly
- TAREAN
- Comparing with known repeats
  - Built in database
  - Custom database

Superclusters
- Contigs
- Tandem repeats annotation
- Similarity based annotation

Results synthesis

Report

TAREAN Assembly

Clusters Built in database Custom database

Results synthesis

Report
Cluster centered analysis

Clusters

- Paired-end analysis
- TAREAN
- Comparing with known repeats
- Built in database
- Custom database

Superclusters

- Assembly
- Results synthesis
- Contigs
- Tandem repeats annotation
- Similarity based annotation

Report
Supercluster Identification
Reads which originate from single repeat are frequently split into multiple cluster during clustering phase – we need to identify such clusters
Identification of related clusters from presence of paired reads

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

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

Suitable cutoff for \( k_{x,y} \) is 0.05 – 0.2:
- Full connection: \( k_{x,y} = 1 \)
- No connection: \( k_{x,y} = 0 \)
Supercluster Identification

paired-end read analysis
In the absence of paired-end reads, clusters are equivalent to superclusters.
Cluster centered analysis

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

Read length $\ll$ monomer

Read length $\geq$ monomer

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

Paired-end reads

Graph-based read clustering

Clusters of reads

Directed graphs

Read orientation
Tandem Repeat Analyzer - TAREAN

Original reads

Oriented reads
A directed graph is called **strongly connected** if every vertex is reachable from every other vertex.
Tandem Repeat Analyzer - TAREAN

\[ C = \frac{\text{size of the largest strongly connected components}}{\text{Total graph size}} \]

\( C \) – Connected component index
Paired-End Sequencing

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

Clusters of potential tandem repeats

k-mer counting

Tandem repeats as de Bruijn graphs

Identification of cycles

Consensus sequences: CATAGACGGTGTTGCTACTATA
TAREAN sort clusters into five groups

- **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**
Cluster centered analysis

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

Reads are assembled by CAP3 program, each clusters separately:

ACTGTGTCGTCGTCGTCGTGTG
CGTCGTCG–CGTGTGGT
GTCGTGTG–TTGTCGTCTGA
ACTGTGTCGTCGTCGTCGTCGTGTGTTGTCGTCTGA

Putative satellite clusters are not assembled by CAP3, instead TAREAN generate k-mer based consensus:
Cluster centered analysis

Clustering

Paired-end analysis
Assembly
TAREAN

Superclusters

Contigs
Tandem repeats annotation

Results synthesis

Comparing with known repeats
Built in database
Custom database

Similarity based annotation

Report

Synthesis
Similarity based annotation

All reads are compared with:

- Database of protein domains
- DNA database
- Custom database (optional)
Similarity based annotation

All reads are compared with:

- Database of protein domains
- DNA database
- Custom database (optional)

Protein domains are derived from coding sequences of transposable elements
Similarity based annotation

All reads are compared with:

- Database of protein domains
- DNA database (Viridiplatae specific!)
- Custom database (optional)

- rDNA
- tRNA
- Plastid DNA
- Mitochondria DNA
- Sequences of potential contaminants
Similarity based annotation

All reads are compared with:

- Database of protein domains
- DNA database
- Custom database (optional)

Library of repeats as DNA sequences in fasta format. The required format for IDs in a custom library is:

>`reapeatname#class/subclass`
Reporting

Clusters

- Paired-end analysis
- Assembly
- TAREAN
- Comparing with known repeats
  - Built in database
  - Custom database

Superclusters

- Contigs
- Tandem repeats annotation
- Similarity based annotation

Results synthesis

Report
Reporting

109610 read total

101061 reads in
1051 supercluster (1090 clusters)

8540 singlets
Reporting

10610 read total

101061 reads in
1051 supercluster (1090 clusters)

cluster
cluster
cluster
SUPERCLUSTER

number of reads [nqc]
150000
150000
100000
50000
0

number of reads [%]
0
10
20
30

annotations_summary
24.02% Class_I/LTR/Ty1_copia/Ivana:Ty1-RT
14.36% Class_I/LTR/Ty1_copia/Ivana:Ty1-RH
1.04% Class_I/LTR/Ty1_copia/Ale:Ty1-RT
0.17% Class_I/LTR/Ty1_copia/SIRE:Ty1-RT
0.11% Class_I/LTR/Ty1_copia/Ale:Ty1-RH
0.08% Class_I/LTR/Ty1_copia/TAR:Ty1-RT
0.03% Class_I/LTR/Ty1_copia/Gymco-II:Ty1-RH
0.03% Class_I/LTR/Ty1_copia/Tork:Ty1-RT
0.03% Class_I/LTR/Ty1_copia/Ikeros:Ty1-RH
Reporting

cluster

SUPERCLUSTER

cluster

cluster

number of reads [pairs]

0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150

CL13

CL48

CL5

CL6

2440

870

3572

3021

Ivana Ty1-GAG
Ivana Ty1-PROT
Ivana Ty1-INT
Ivana Ty1-RH
Ivana Ty1-RT
### Reporting

#### Cluster Summary

<table>
<thead>
<tr>
<th>Cluster</th>
<th>nhits</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3181</td>
<td>0.32</td>
</tr>
<tr>
<td>--repeat</td>
<td>3181</td>
<td>0.32</td>
</tr>
<tr>
<td>--mobile_element</td>
<td>3181</td>
<td>0.32</td>
</tr>
<tr>
<td>--Class I</td>
<td>3181</td>
<td>0.32</td>
</tr>
<tr>
<td>--LTR</td>
<td>3181</td>
<td>0.32</td>
</tr>
<tr>
<td>--Tyl_copia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-AAle</td>
<td>64</td>
<td>0.0065</td>
</tr>
<tr>
<td>-Alesia</td>
<td>5</td>
<td>5e-04</td>
</tr>
<tr>
<td>-Angela</td>
<td>1</td>
<td>1e-04</td>
</tr>
<tr>
<td>-Bianca</td>
<td>1</td>
<td>1e-04</td>
</tr>
<tr>
<td>-Bryco</td>
<td>14</td>
<td>0.0014</td>
</tr>
<tr>
<td>-Gymco-I</td>
<td>1</td>
<td>1e-04</td>
</tr>
<tr>
<td>-Gymco-II</td>
<td>3</td>
<td>3e-04</td>
</tr>
<tr>
<td>-Ikeros</td>
<td>3</td>
<td>3e-04</td>
</tr>
<tr>
<td>-Ivana</td>
<td>3062</td>
<td>0.31</td>
</tr>
<tr>
<td>-SIRE</td>
<td>8</td>
<td>0.00081</td>
</tr>
<tr>
<td>-TAR</td>
<td>4</td>
<td>4e-04</td>
</tr>
<tr>
<td>-Tork</td>
<td>15</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

#### Domain String

- 3 (Ty1-INT), 1 (Ty1-PROT), 4 (Ty1-RH), 56 (Ty1-RT),
- 5 (Ty1-INT), 1 (Ty1-INT), 1 (Ty1-RT),
- 14 (Ty1-INT), 1 (Ty1-INT),
- 2 (Ty1-INT), 1 (Ty1-RH),
- 2 (Ty1-INT), 1 (Ty1-RH),
- 288 (Ty1-GAG), 985 (Ty1-INT), 189 (Ty1-PROT), 513 (Ty1-RH), 1087 (Ty1-RT),
- 2 (Ty1-INT), 6 (Ty1-RT),
- 1 (Ty1-INT), 3 (Ty1-RT),
- 2 (Ty1-GAG), 12 (Ty1-RT), 1 (Ty1-RT),

#### Graph Representation

- Nodes labeled CL6, CL5, and Ivana with respective counts 3021, 3572, and 870.
- Edges connecting the nodes represent relationships or interactions.

#### Additional Calculations

- Number of reads (peaks) is calculated as follows:
  - CL6: 3021 reads
  - CL5: 3572 reads
  - Ivana: 870 reads

---

**Note:** The table and graph provide insights into the clustering and domain distribution within the dataset.
### Repeat annotation summary

<table>
<thead>
<tr>
<th>Genome_proportion[%]</th>
<th>Nsuperclusters</th>
<th>Nclusters</th>
<th>Nreads</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unclassified_repeat</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--rDNA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--45S_rDNA</td>
<td>14.71</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>--18S_rDNA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>--25S_rDNA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--5.8S_rDNA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--5S_rDNA</td>
<td>2.3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>--mobile_element</td>
<td>20.11</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>°--mobile_element</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--SINE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--LTR</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--Tyl_copia</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--Ale</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--Alesia</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--OTA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--Ty3_gypsy</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--Ty1_outgroup</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--SINE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--LTR</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--organelle</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--plastid</td>
<td>8.57</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>°--mitochondria</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--plastid</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--mitochondria</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>--Unclassified</td>
<td>6.92</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>contamination</td>
<td>6.42</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>°--TatII</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--TatIII</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>°--TatIV Ogre</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
What is the best hit?

Top-down classification

[Diagram of a phylogenetic tree showing relationships between different classes and species, such as Ty1_copia and Ty3_gypsy.]
What is the best hit?

Top-down classification

Diagram of a hierarchical classification system, showing various categories and subcategories related to genetic elements and mobile genetic elements.
What is the best hit?

**Parent** (LTR)

- **Child_node1** (Pararetrovirus)
- **Child_node2** (Ty1/Copia)
  - SIRE
  - Angela
- **Child_node3** (Ty3/Gypsy)

Best child selection criteria

- **best hit proportion**: $\frac{H_{c,1}}{H_p} > 0.7$
- **best hit to second best hit**: $\frac{H_{c,1}}{H_{c,1} + H_{c,2}} > 0.9$
- **overall hits proportion**: $\frac{H_{c,1}^2}{N} > 2.5$

* $N$ number of reads in supercluster
* $H_p$ number of hits in parent node
* $H_{c,x}$ number of hits in children node $x$, nodes are sorted by number of hits, largest is the first
* $H_{c,1}$ number of hits in child node with the highest number of hits (best child)
* $H_{c,2}$ number of hits in child node with the second highest number of hits
Clustering can be run in two different modes:

- **Full Repeat Analysis**
  Focus on all types of repeat but less sensitive satellite detection

- **Tandem Repeat Analysis**
  Focus on tandem repeat detection only
  Better sensitivity of satellite identification
Full Repeat Analysis

Clusters

- Paired-end analysis
- Clustering
- Assembly
- TAREAN

Comparing with known repeats

Built in database
Custom database

Results synthesis

Superclusters

Contigs

Tandem repeats annotation

Similarity based annotation

Custom database

Report
- Satellite with longer monomer tend to split onto multiple clusters
- Merging before running TAREAN analysis will improve detection of such satellites
RepeatExplorer2 availability:

source code
https://bitbucket.org/petrnovak/repex_tarean

• Galaxy server – Graphical user interface
http://repeatexplorer-elixir.cerit-sc.cz/

www.repeatexplorer.org - Manuals

• Your custom Galaxy instance
• Command line
Collaboration

Abbott

Costello
RepeatExplorer
Discover repeats in next generation sequencing data
POWERED BY
CAFFEINE