8th Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis in Plants

May 21-23, 2019

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?

(vertices, nodes)

(edges)
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
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**
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:

```
AAAGCTCAGTTTCGAGCCAGAGACCAGAAAGTGTTGGGAGCTTACAGCACAAGACTTTACAGCAAGAGCCGAG
```

```
AAAGCTCAGTT
AAGCTCAGTTT
AGCTCAGTTTC
GCTCAGTTTCG
CTCAGTTTTCGA
TCAGTTTTCGAG
CAGTTTTCGAGC
AGTTTTCGAGCC
```

```
AAAGCTCAGTTT
AAGCTCAGTTTC
AGCTCAGTTTC
GCTCAGTTTCG
CTCAGTTTTCGA
TCAGTTTTCGAG
CAGTTTTCGAGC
```

```
GCTCAGTTTCG
CTCAGTTTTCGA
TCAGTTTTCGAG
CAGTTTTCGAGC
```

```
......
```

```
AAAGCTCAGTT
AAGCTCAGTTT
AGCTCAGTTTC
GCTCAGTTTCG
CTCAGTTTTCGA
TCAGTTTTCGAG
CAGTTTTCGAGC
AGTTTTCGAGCC
```

```
AAAGCTCAGTT
AAGCTCAGTTT
```

```
AGCTCAGTTTC
GCTCAGTTTCG
CTCAGTTTTCGA
TCAGTTTTCGAG
CAGTTTTCGAGC
```

```
GCTCAGTTTCG
CTCAGTTTTCGA
TCAGTTTTCGAG
CAGTTTTCGAGC
```

```
CAGTTTTCGAGC
```

```
CAGTTTTCGAGC
```

```
CAGTTTTCGAGC
```

```
CAGTTTTCGAGC
```

```
CAGTTTTCGAGC
```

```
CAGTTTTCGAGC
```

```
CAGTTTTCGAGC
```
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:

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
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 representation diagram]

- "bubbles" → sequence variants
- shorter paths → deletions
RepeatExplorer workflow

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

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

Sequence Graph
- Clustering
  - Sequence Graph
  - mgblast
  - HSP filtering

Clusters
RepeatExplorer workflow

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

- Short reads
- High quality short reads
- Quality control
- Trimming
- Adapter removing

Prerun
- Estimating data size limit
- All reads or subset

All-to-All comparison
- mgblast hit filtering

Clustering
- Sequence Graph
- Clusters
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 names
Preprocessing

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

FastQC program

- Galaxy
- GUI based
- Command line
Preprocessing

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

**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 names**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Sequence</th>
<th>Trimming</th>
<th>Adapter Removing</th>
</tr>
</thead>
<tbody>
<tr>
<td>@DRR013373.1</td>
<td>HWI-ST1034:55:C0BMWACX:1:1101:1149:2090/1</td>
<td>ATGGATAACGCCCTATGATGCCAGAAACTGCGCTTGGAAGAATCATGGATCCCGTGAGCGAGGAGGTCGTCGA</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>@DRR013373.2</td>
<td>HWI-ST1034:55:C0BMWACX:1:1101:1205:2126/1</td>
<td>AACTTCATCGAATGCTTCGCTACCGGATTAGTTCGGAGAATCCAATCACCGATGCCATCTTTACAGCCGATAACACCACCACCTTCTTGTCTTCTTACGT</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>@DRR013373.3</td>
<td>HWI-ST1034:55:C0BMWACX:1:1101:1300:2065/1</td>
<td>ATCCGGTTGCCGAGAGTCGTTTTAGACTTTATATCGCAGCACAGCACCCGCGCACACACCGTCTCCGGGGAGGCGAATGCTAGCCGCTCGTTTGCTTCTTCC</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>@DRR013373.4</td>
<td>HWI-ST1034:55:C0BMWACX:1:1101:1352:2147/1</td>
<td>ATCCGGTTGCCGAGAGTCGTTTTAGACTTTATATCGCAGCACAGCACCCGCGCACACACCGTCTCCGGGGAGGCGAATGCTAGCCGCTCGTTTGCTTCTTCC</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>@DRR013373.5</td>
<td>HWI-ST1034:55:C0BMWACX:1:1101:1374:2154/1</td>
<td>ATGCATTTAAACACTTAGCTAGCTAAAGAAACAGAGGATTCCACTAACAACCTATCCCATCCTTAAGAACTTTGCAGAACTACTCAAGAACACTGAAGAAC</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>@DRR013373.6</td>
<td>HWI-ST1034:55:C0BMWACX:1:1101:1430:2218/1</td>
<td>TGGCACCTCGATGTCGGCTCTTCGCCACCTGGGGCTGTAGTATGTTCCAAGGGTTGGGCTGTTCGCCCATTAAAGCGGTACGTGGGCTGGGTTCAGAACGC</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>@DRR013373.7</td>
<td>HWI-ST1034:55:C0BMWACX:1:1101:1469:2227/1</td>
<td>AATCAATATAGATTGATCCGAAATATGATTCAAATCCAATTCCAATATAGTCCCTATGGGTACATAAGAAATGTATTGAATCGATTCTTTTTAATGAAGAG</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

### Visual inspection!
Preprocessing

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

**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 names

Tool: affixer

Genome AB
>FUXEH4V01EAG68....
acgacagctgactaatgc
>FUXEH4V01BKPDK
cctcggaggctacacgagct
>FUXEH4V01AJAJV
actatcgacactgccggcgcg
...

Genome XY
>F0X50LU02GZ8YF....
gccccgtcgccgctccgtgctgcg
>F0X50LU02I1AMY
tgtgtgccccgtcgcgcggccccc
>F0X50LU02HYN8U
atatgctatgcggtcg
...

comparative analysis:
>AB1
acgacagctgactaatgc
>AB2
cctcggaggctacacgagct
>AB3
actatcgacactgccggcgcg
...
>XY1
gccccgtcgccgctccgtgctgcg
>XY2
tgtgtgccccgtcgcgcggccccc
>XY3
atatgctatgcggtcg
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

Clusters
Prerun: all-to-all sequence comparison on small sample of NGS reads

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

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

\[ k_g \] genome specific coefficient - **graph density** depends on repetitive content and genome size

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

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
**Prerun:** all-to-all sequence comparison on small sample of NGS reads

### 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**

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

**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 pairs of reads with similarity above threshold are found using mgblast:

**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
...

By default, DustMasker is on but it can be disabled to increase sensitivity of simple repeats detection.

Beware: Disabling dust can significantly increase computation time and memory usage!
Clustering

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

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

All-to-All Comparison

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

Paired-end analysis

Assembly

TAREAN

Clusters

Comparing with known repeats

Built in database
Custom database

Superclusters

Contigs

Results synthesis

Tandem repeats annotation

Similarity based annotation

Report

Paired-end analysis

Assembly

TAREAN

Clusters

Comparing with known repeats

Built in database
Custom database

Superclusters

Contigs

Results synthesis

Tandem repeats annotation

Similarity based annotation

Report
Cluster centered analysis

Paired-end analysis

Superclusters

Clusters

Assembly

TAREAN

Contigs

Results synthesis

Tandem repeats annotation

Comparing with known repeats

Built in database

Custom database

Similarity based annotation

Report

Synthesis

Annotation

Clusters
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

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

- \( W \) 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 \( k_{x,y} = 0 \)
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

Clusters

Comparing with known repeats

Built in database

Custom database

Superclusters

Contigs

TAREAN

Tandem repeats annotation

Results synthesis

Similarity based annotation

Report
TAREAN calculates graph layout and provide automatic analysis of graph topology with the aim to identify tandem repeats.
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

Vertex = sequence read

k-mer counting

Tandem repeats as de Bruijn graphs

Vertex = kmer

Identification of cycles

Consensus sequences

CATAGACGTGGTACGCTACTATA
TAREAN sorts 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
- Superclusters
- TAREAN
- Contigs
- Results synthesis
- Tandem repeats annotation
- Similarity based annotation
- Comparing with known repeats
- Built in database
- Custom database
- Report

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

ACTGTGTCGTCGTCGTCGTCGTGTG
CGTCGTTCG-CGTGTGGTG
GTCGTGTG-TTGTCGTCTGA
ACTGTGTCGTCGTCGTCGTCGTCGTGTGTTGTCGTCTGA

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

Clusters

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
Similarity based annotation

All reads are compared with:

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

All reads are compared with:

- Database of protein domains (REXdb)
- 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

Results synthesis

- Superclusters
- Contigs
  - Tandem repeats annotation
  - Similarity based annotation

Report
Reporting

109610 read total

101061 reads in 1051 supercluster (1090 clusters)

cluster

cluster

cluster

SUPERCLUSTER

cluster

cluster

cluster

SUPERCLUSTER

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
## Reporting

### Cluster SuperRead

<table>
<thead>
<tr>
<th>nhits</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>3181</td>
<td>0.32</td>
</tr>
</tbody>
</table>

**Clusters**

- **-repeat**
- **-mobile element**
- **-Class I**
- **-LTR**
- **-Ty1_copia**
  - **-Ale** 64 0.0065
  - **-Alesia** 5 5e-04
  - **-Angela** 1 1e-04
  - **-Bianca** 1 1e-04
  - **-Bryco** 14 0.0014
  - **-Gymco-I** 1 1e-04
  - **-Gymco-II** 3 3e-04
  - **-Ikeros** 3 3e-04
  - **-Ivana** 3062 0.31
  - **-SIRE** 8 0.00081
  - **-TAR** 4 4e-04
  - **-Tork** 15 0.0015

**Domains**

- 3 (Ty1-INT), 1 (Ty1-PROT), 4 (Ty1-RH), 56 (Ty1-RT), 5 (Ty1-INT), 1 (Ty1-RT), 14 (Ty1-INT), 1 (Ty1-INT), 2 (Ty1-INT), 1 (Ty1-RH), 2 (Ty1-INT), 1 (Ty1-RT), 188 (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-INT), 1 (Ty1-RT), 1 (Ty1-RT)
<table>
<thead>
<tr>
<th>Genome_proportion[%]</th>
<th>Nsuperclusters</th>
<th>Nclusters</th>
<th>Nreads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclassified_repeat</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- rDNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45S_rDNA</td>
<td>14.71</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>18S_rDNA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25S_rDNA</td>
<td>0</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>Satellite</td>
<td>20.11</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Mobile_element</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class_I</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>Ty1_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>- Alysa</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- Angela</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>mitochondria</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>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>
<tr>
<td>Genome proportion</td>
<td>6.92</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>- Unclassified</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- contamination</td>
<td>6.42</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
What is the best hit?

Top-down classification

Diagram showing the classification of various DNA elements and viruses, including Ty1_copia, Ty3_gypsy, non-chromovirus, chromovirus, and various subcategories like 18S_rDNA, 25S_rDNA, 5.8S_rDNA, SINE, LTR, pararetrovirus, DIRS, Penelope, LINE, Athila, Ogre, Tail, Taili, TailII, TailV_Ogre, and TailV.
What is the best hit?

Top-down classification
What is the best hit?

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
What is the best hit?

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: \( \frac{H^2_{c,1}}{N} > 2.5 \)

\( N \) 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 from highest to lowest
\( H_{c,1} \) number of hits in child node with the highest number of hits (best hit)
\( H_{c,2} \) 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

Clustering

Paired-end analysis

TAREAN

Assembly

Clusters

Comparing with known repeats

Built in database

Custom database

Superclusters

Results synthesis

Contigs

Tandem repeats annotation

Similarity based annotation

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 2.0
Discover repeats in your next generation sequencing data

POWERED BY
CAFFEINE

POWERED BY
ICE CREAM