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)
- TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res., doi:10.1093/nar/gkx257(2017)

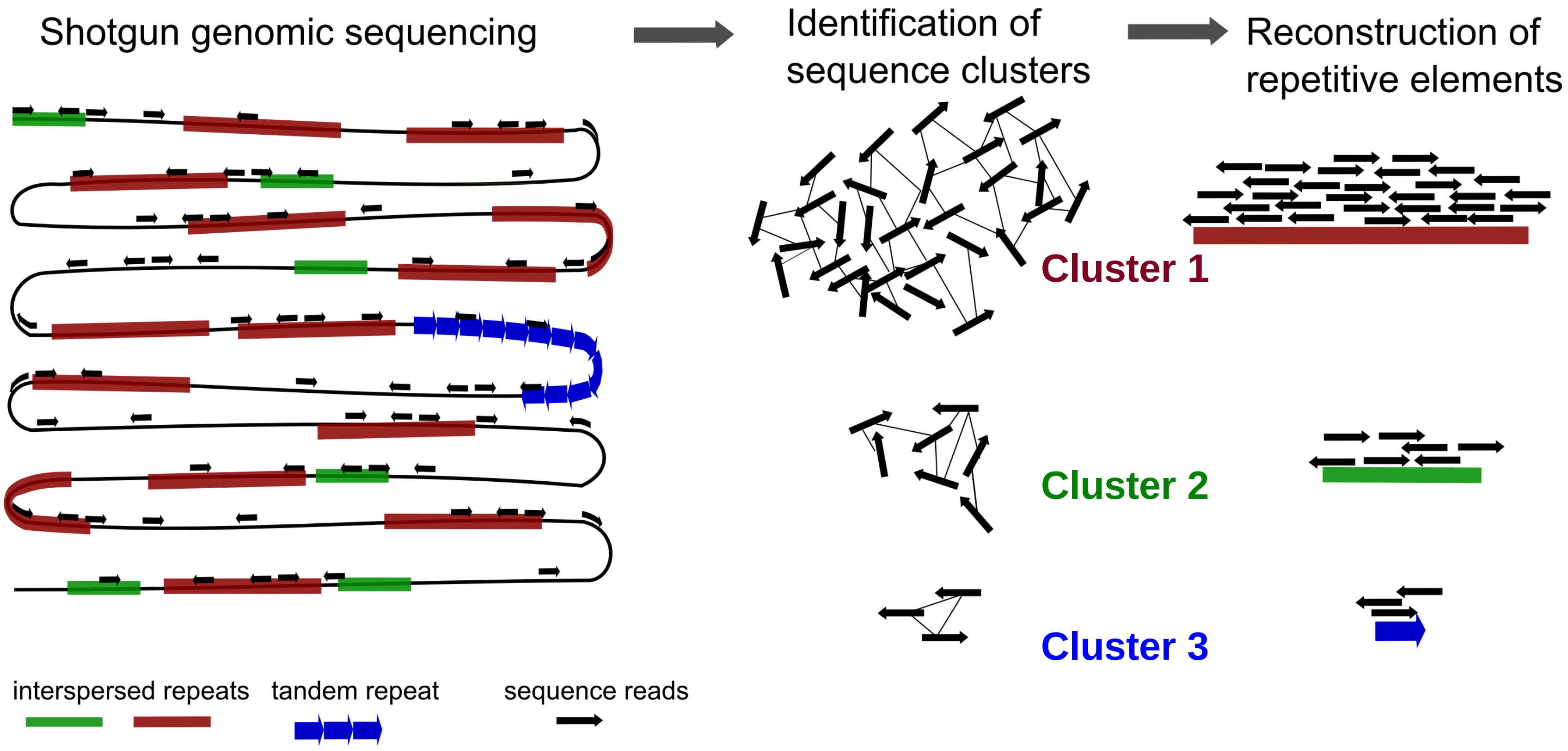
Protocols

- Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nature Protocols 15:3745–3776.

Available Tools:

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

Principle of RepeatExplorer



Shotgun genomic sequencing

Identification of sequence clusters

Reconstruction of repetitive elements

Cluster 1

Cluster 2

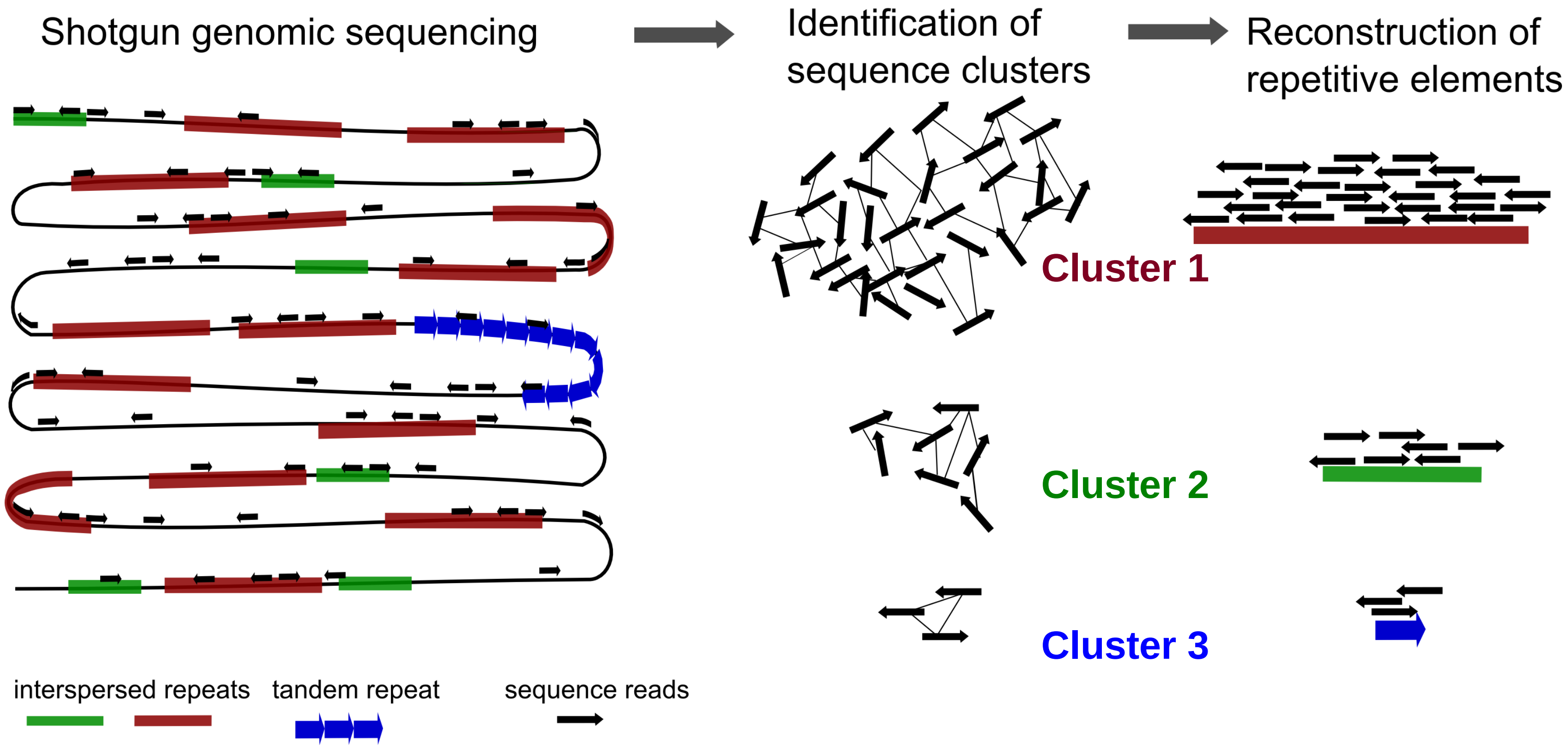
Cluster 3

interspersed repeats 🟩 🟥

tandem repeat

sequence reads

Principle of RepeatExplorer



Shotgun genomic sequencing

Identification of sequence clusters

Reconstruction of repetitive elements

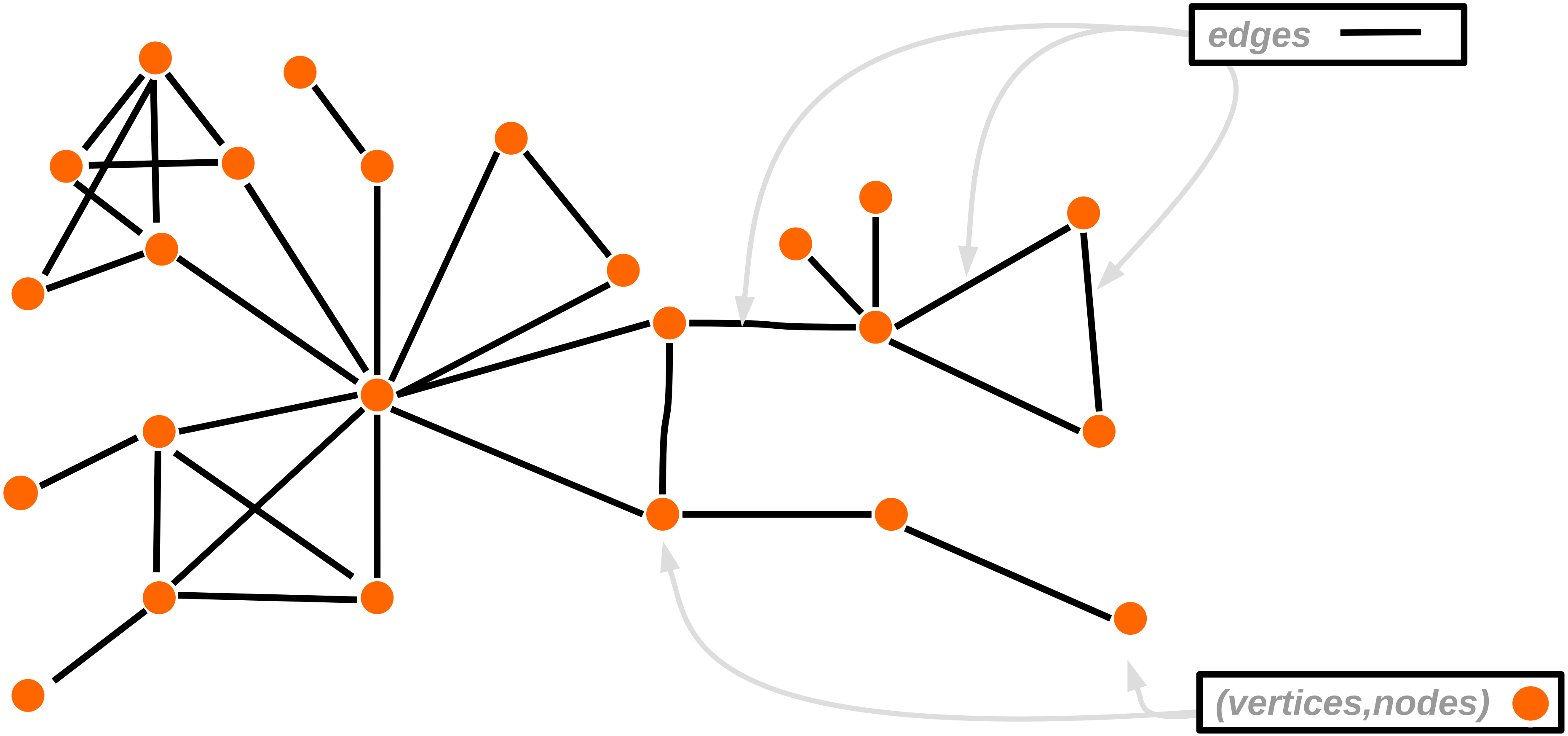
Cluster 1

Cluster 2

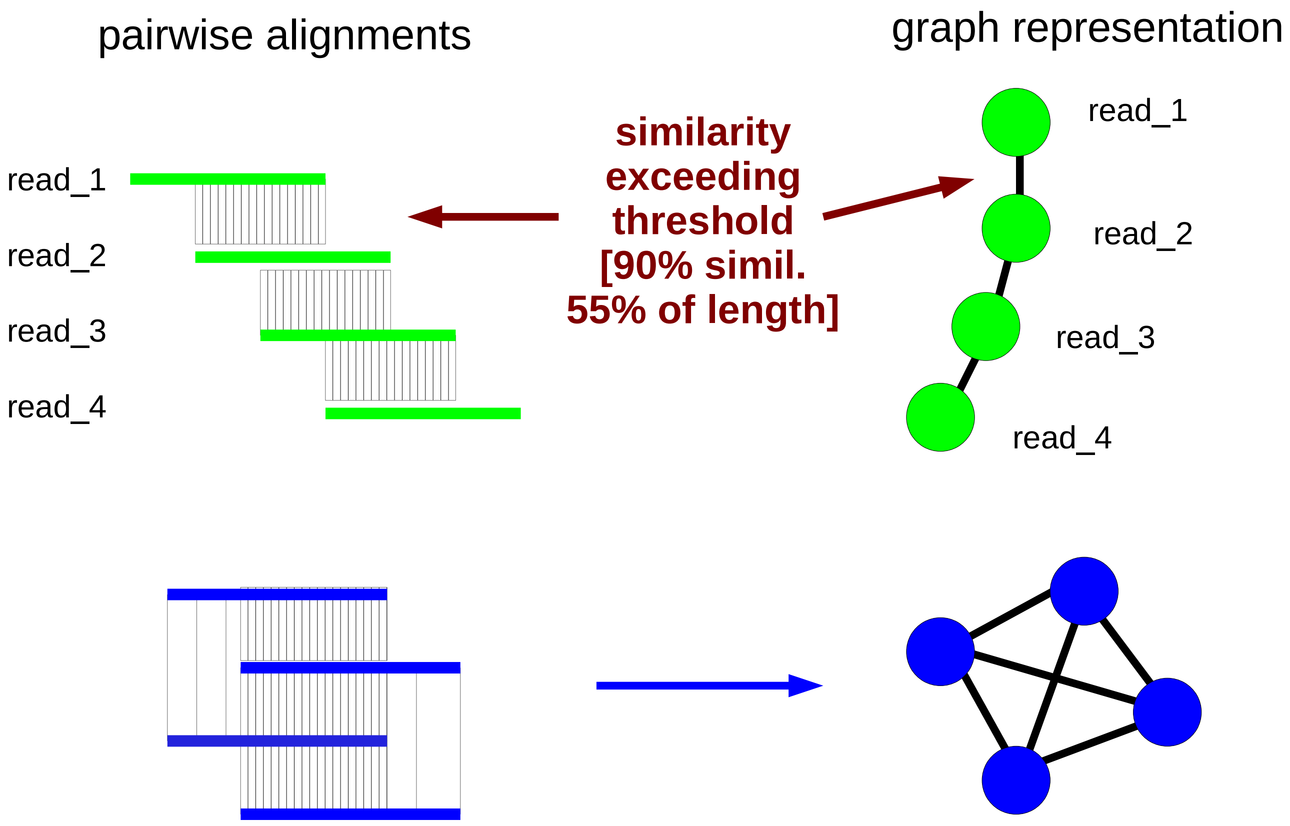
Cluster 3

CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

Graph Based Representation of Sequence Reads



Graph Based Representation of Sequence Reads

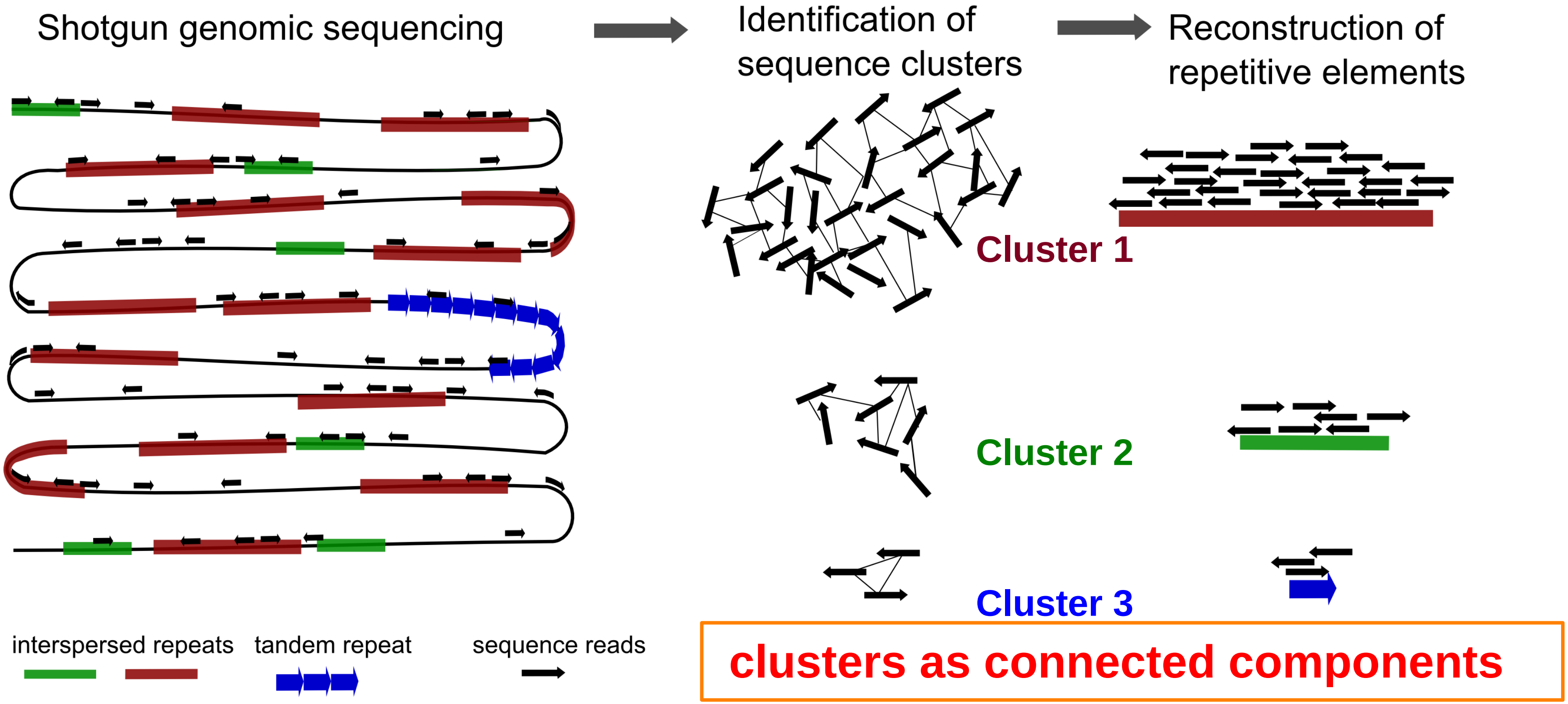


pairwise alignments

graph representation

similarity exceeding threshold [90% simil. 55% of length]

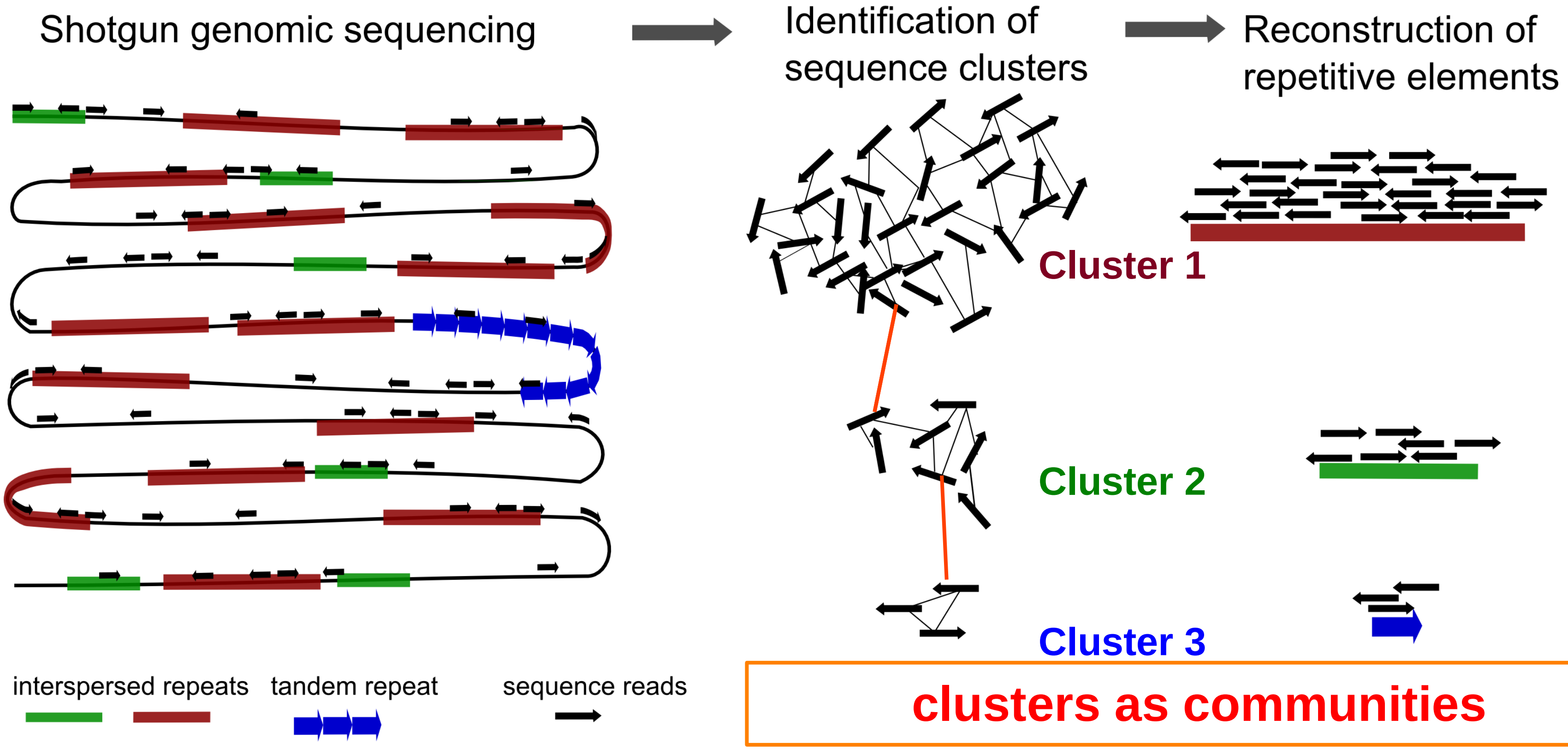
Principle of RepeatExplorer



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

clusters as connected components

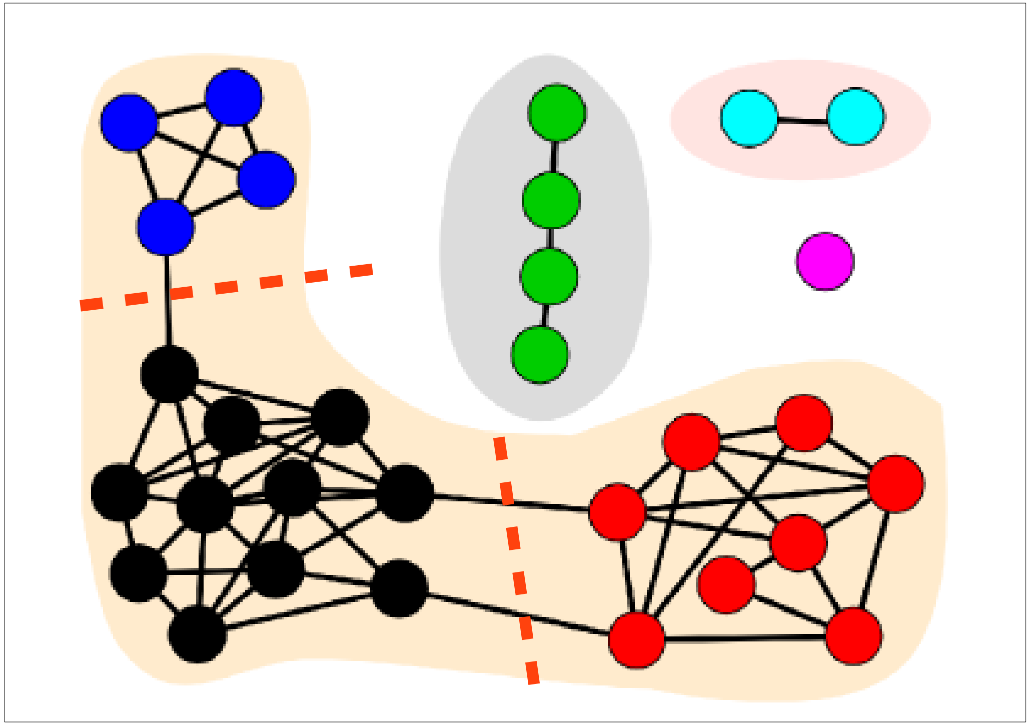
Principle of RepeatExplorer



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

clusters as communities

Graph Based Clustering

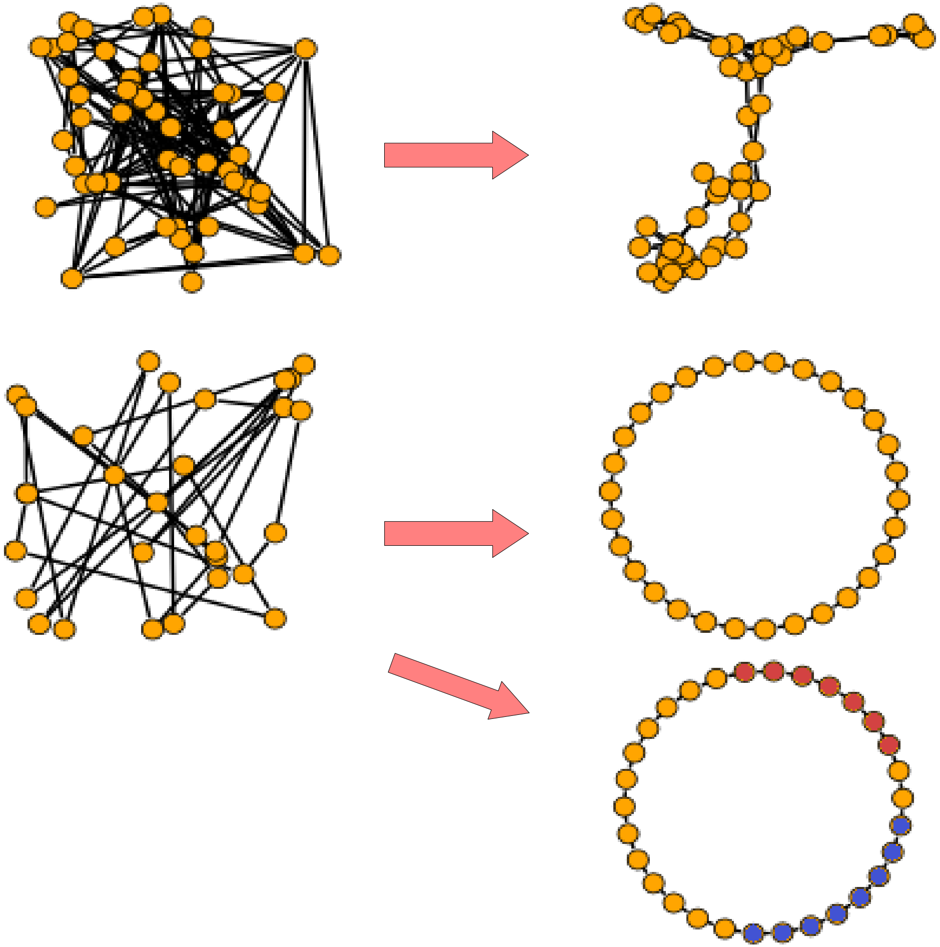


A **community**, with respect to graphs, can be defined as a subset of nodes that are densely connected to each other and loosely connected to the nodes in the other communities in the same graph

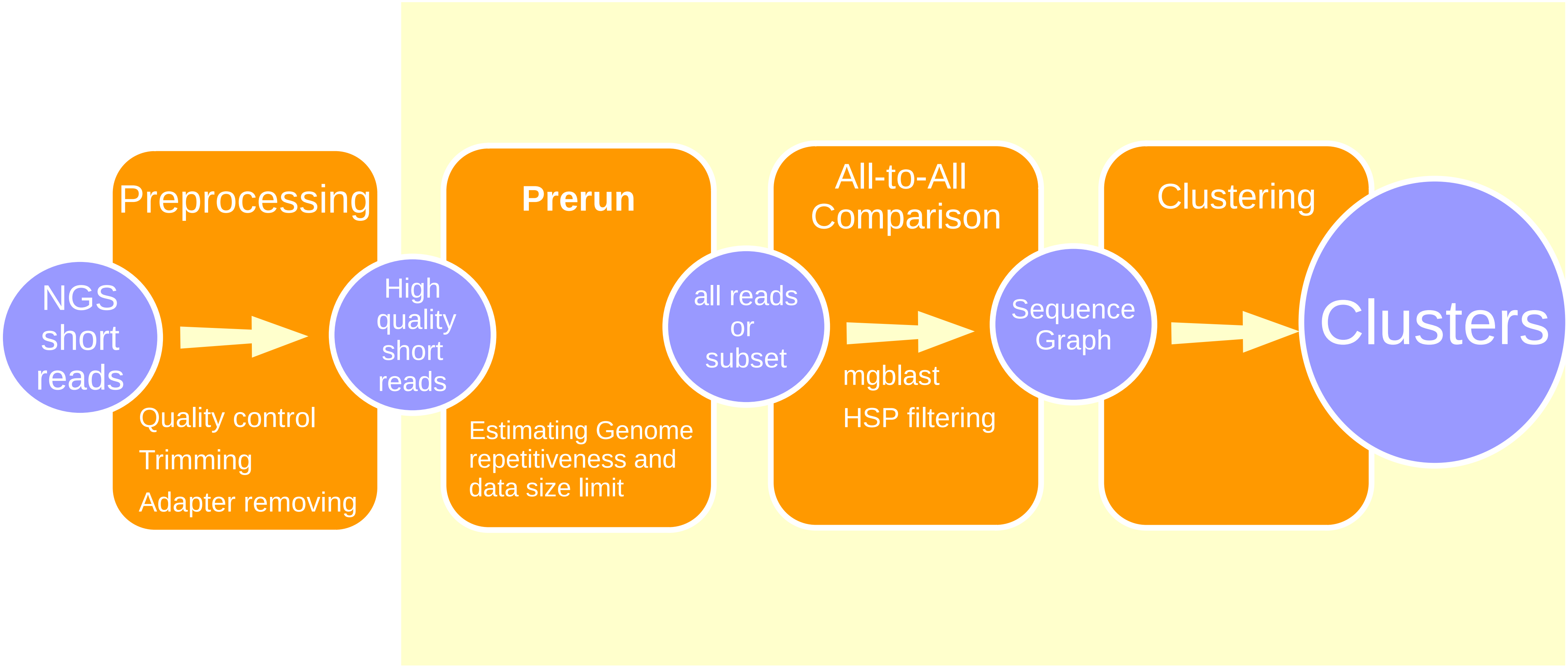
community ~ cluster ~ repeat family

Graph Based Representation of Sequence Reads

- Informative graphical representation
- Graph layout
- Vertex coloring



RepeatExplorer Pipeline



NGS short reads

Preprocessing

Quality control

Trimming

Adapter removing

High quality short reads

Prerun

Estimating Genome repetitiveness and data size limit

all reads or subset

All-to-All Comparison

mgblast

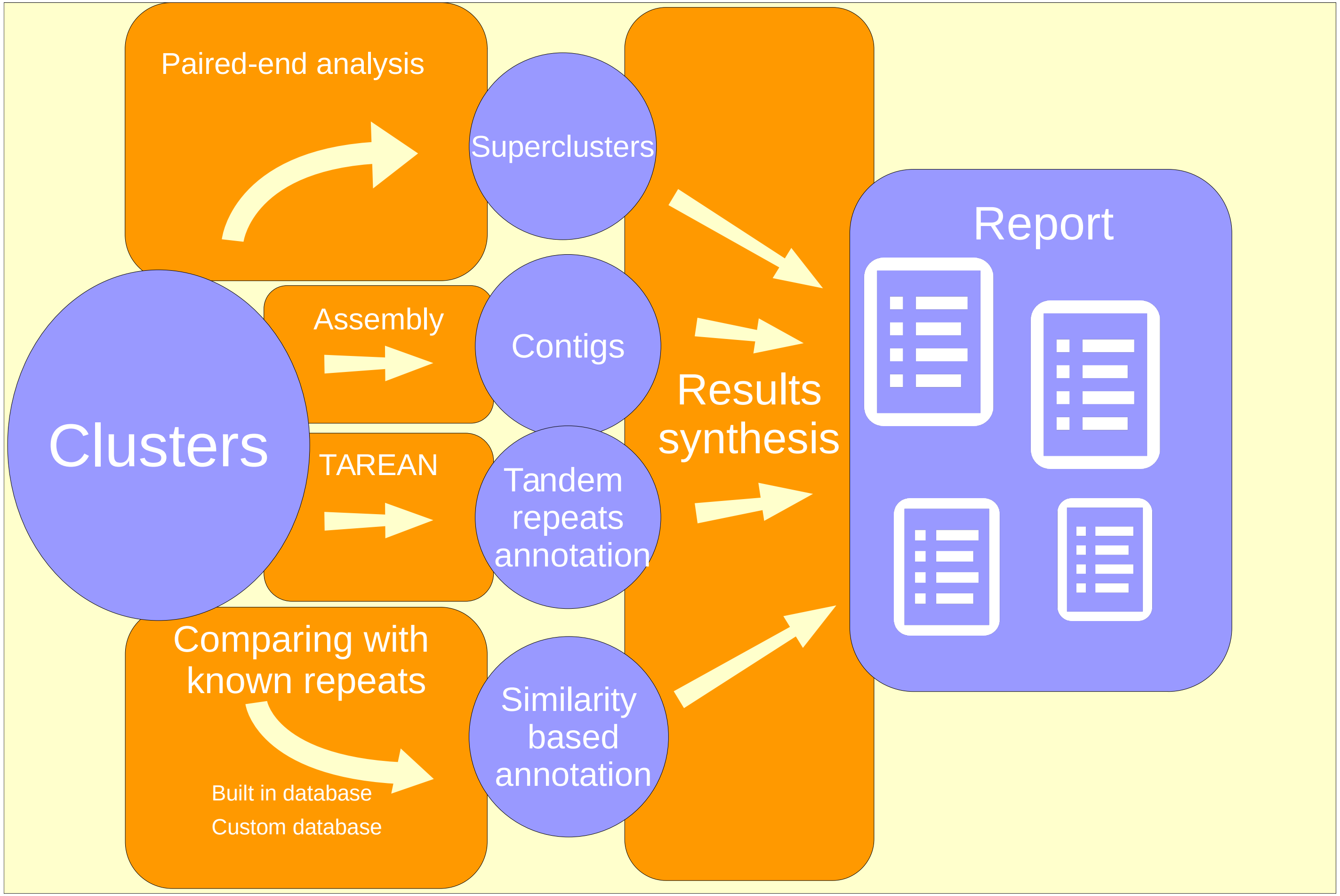
HSP filtering

Sequence Graph

Clustering

Clusters

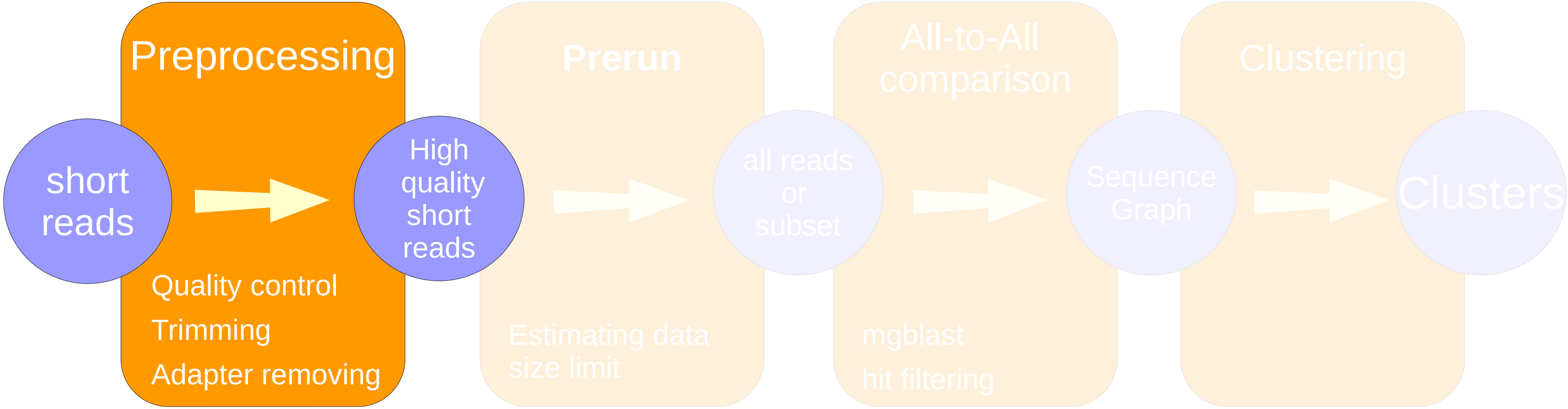
RepeatExplorer Pipeline



RepeatExplorer pipeline

Input data

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

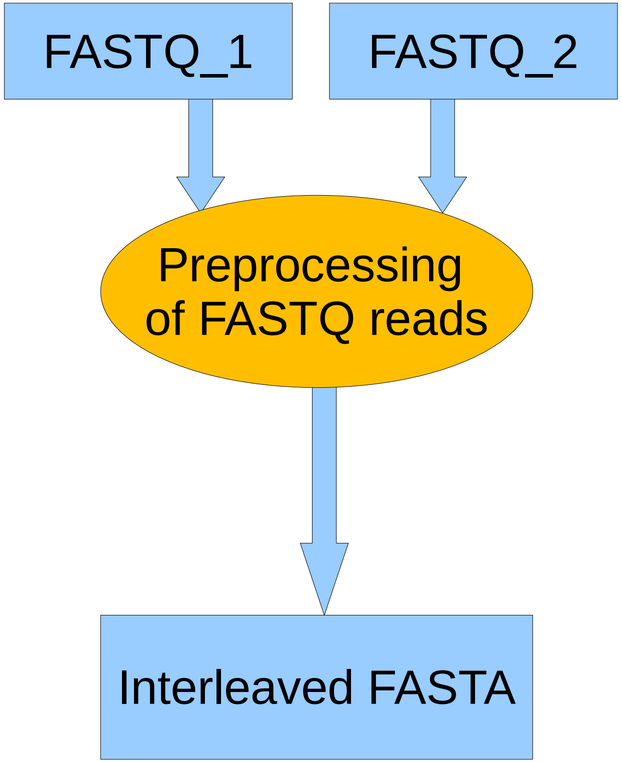


Preprocessing

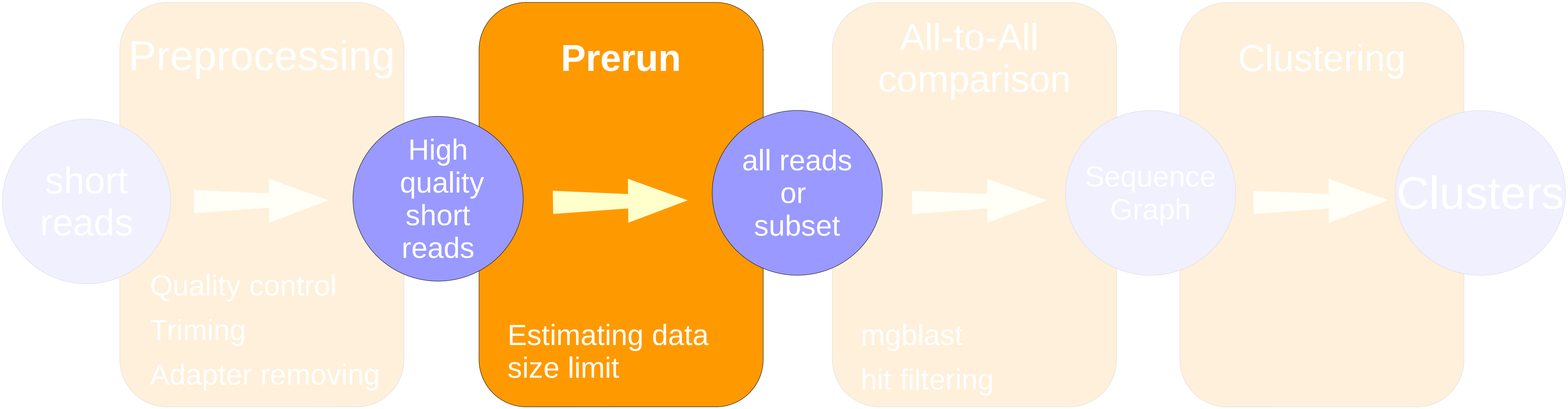
RepeatExplorer utilities:

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

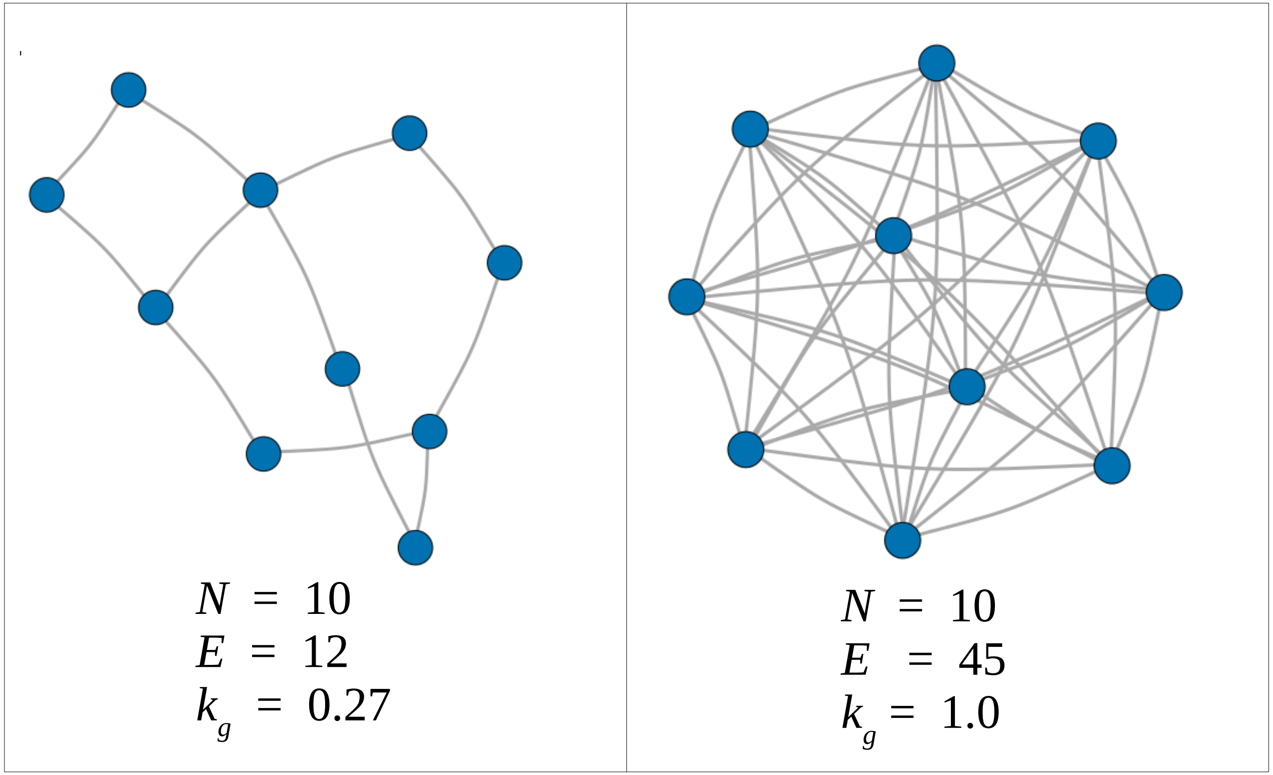


RepeatExplorer pipeline



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

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

$$N_{max} = \sqrt{m \frac{M}{k_g}}$$

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**)

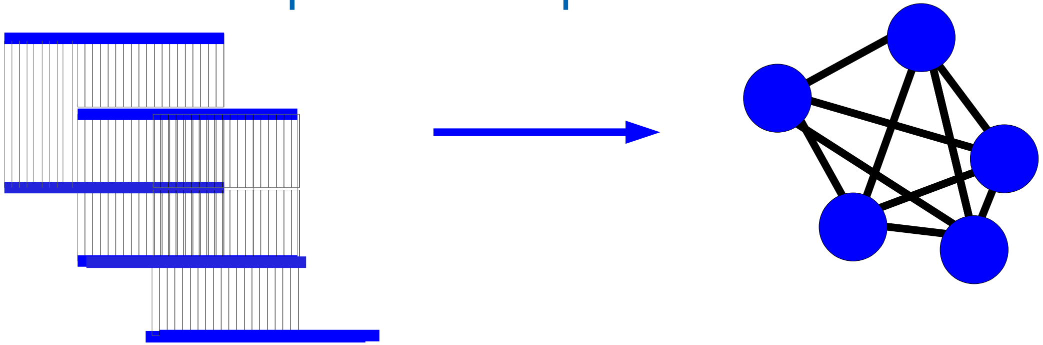
$$N_{max} = \sqrt{m \frac{M}{k_g}}$$

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

Number of reads which can be processed with 16GB RAM

Prerun

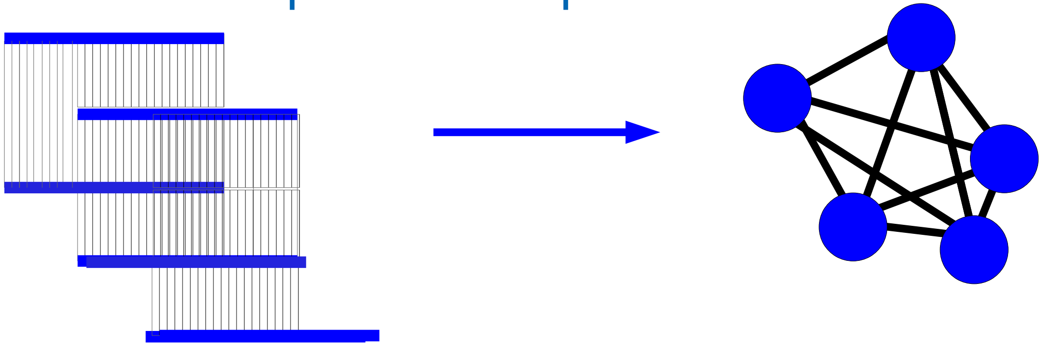
All-to-all sequence comparison on small sample of input data



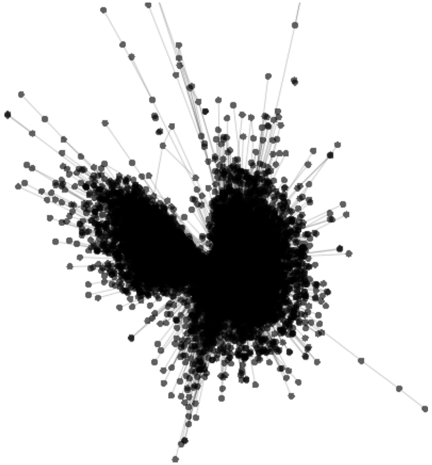
graph density = 1

Prerun

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



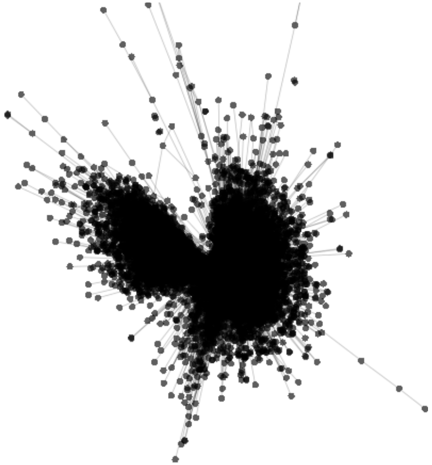
graph density = 1



graph density = 0.54

Prerun

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



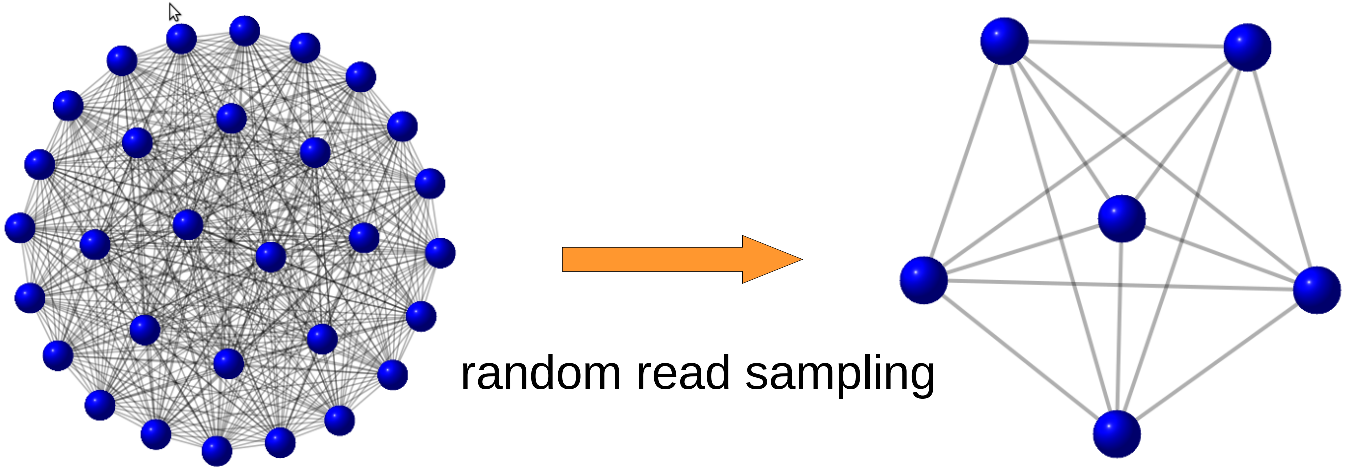
Number of reads (Vertices)	44,772
Number of similarity hits (Edges)	542,348,907

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

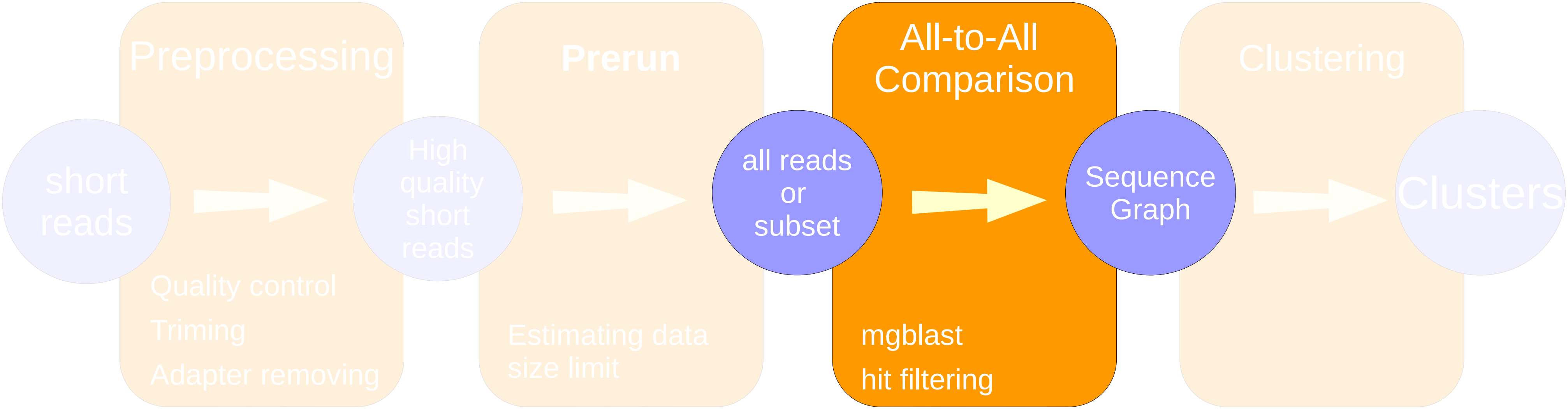
Satellite filtering (optional)



random read sampling

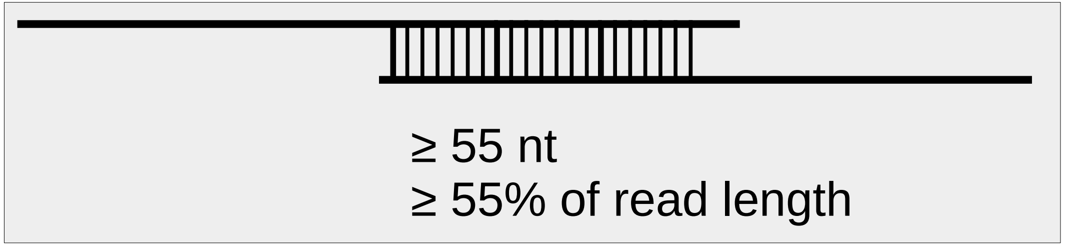
- Clusters composed from satellite reads can be scaled down without loosing information.
- Sample of 10% of reads of is kept in analysis to keep track of this satellite

RepeatExplorer pipeline



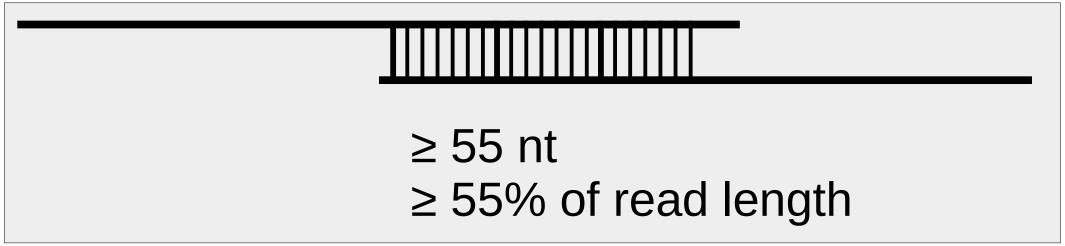
All-to-all comparison

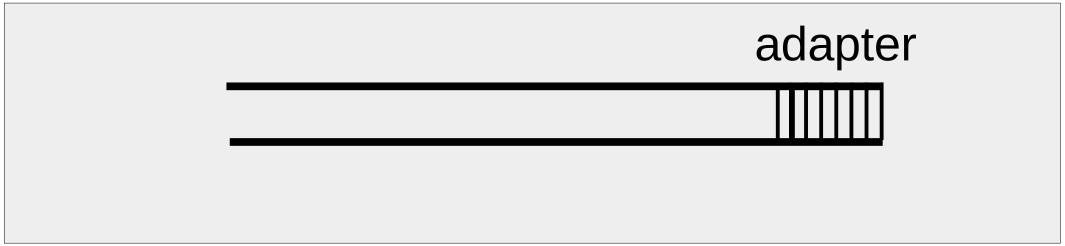
- Similarity search using **mgblast**
- Default threshold:
  - overlap : 55 nt and 55% of the length
  - minimal similarity 90%
- By default mgblast is using **DustMasker** (low complexity repeat filter)
  - simple repeats are underestimated or not detected (e.g. telomeric motifs, microsatellites)
 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



All-to-all comparison

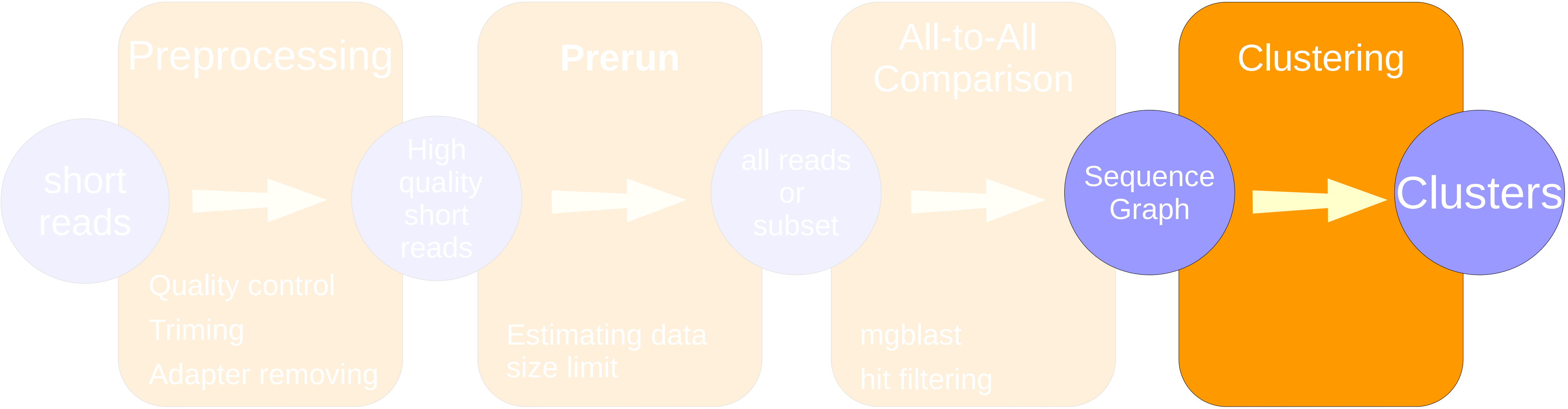
- Similarity search using **mgblast**
- Default threshold:
  - overlap : 55 nt and 55% of the length
  - minimal similarity 90%
- By default mgblast is using **DustMasker** (low complexity repeat filter)
  - simple repeats are underestimated or not detected (e.g. telomeric motifs, microsatellites)
 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



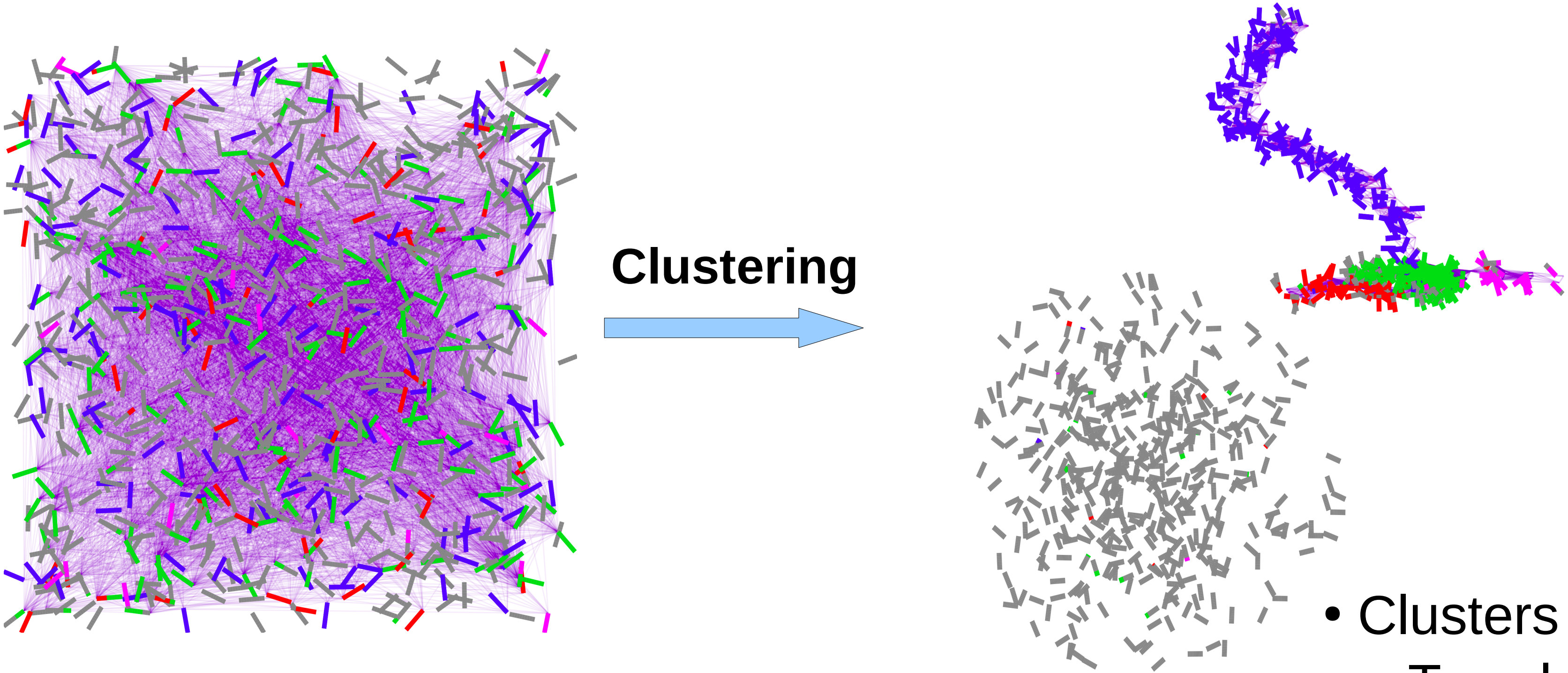


RepeatExplorer pipeline

- 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



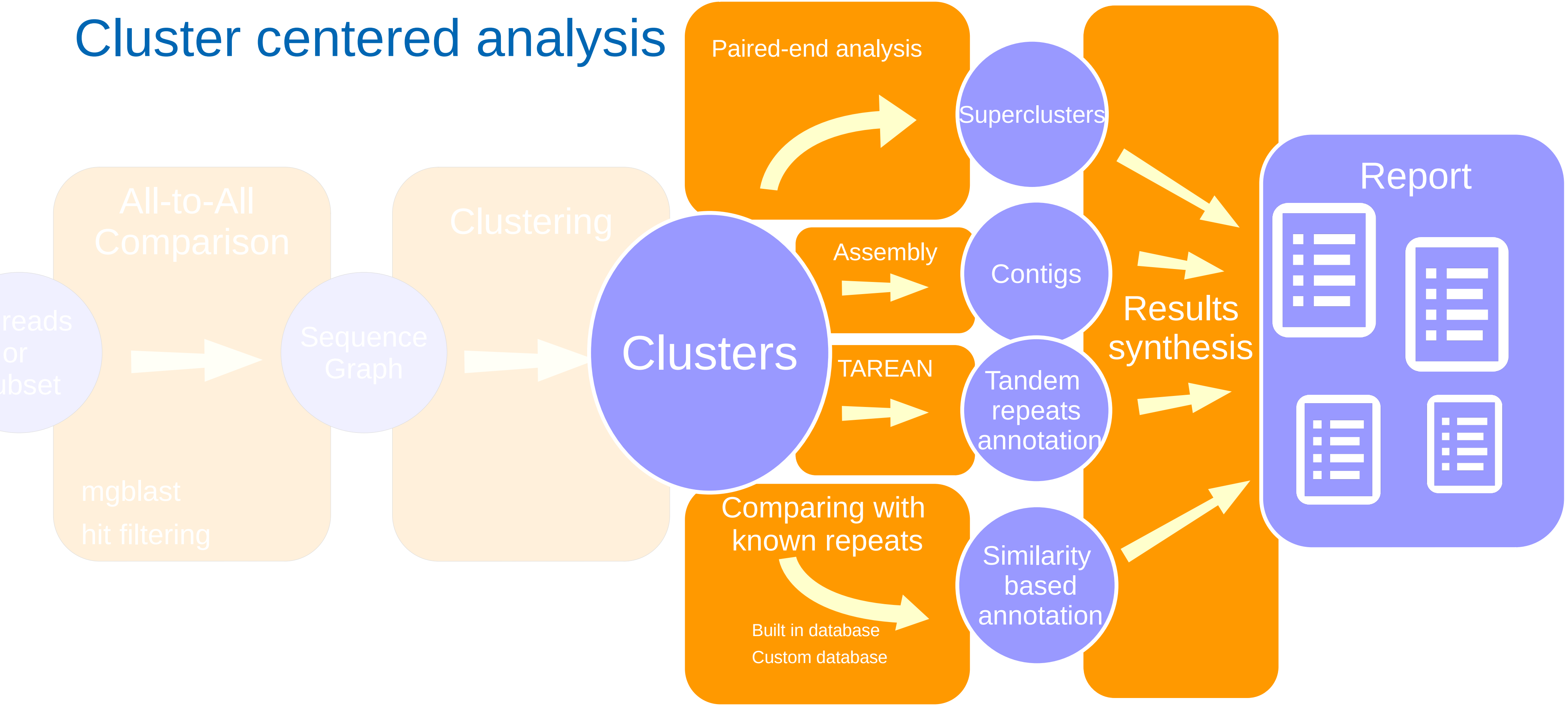
RepeatExplorer pipeline



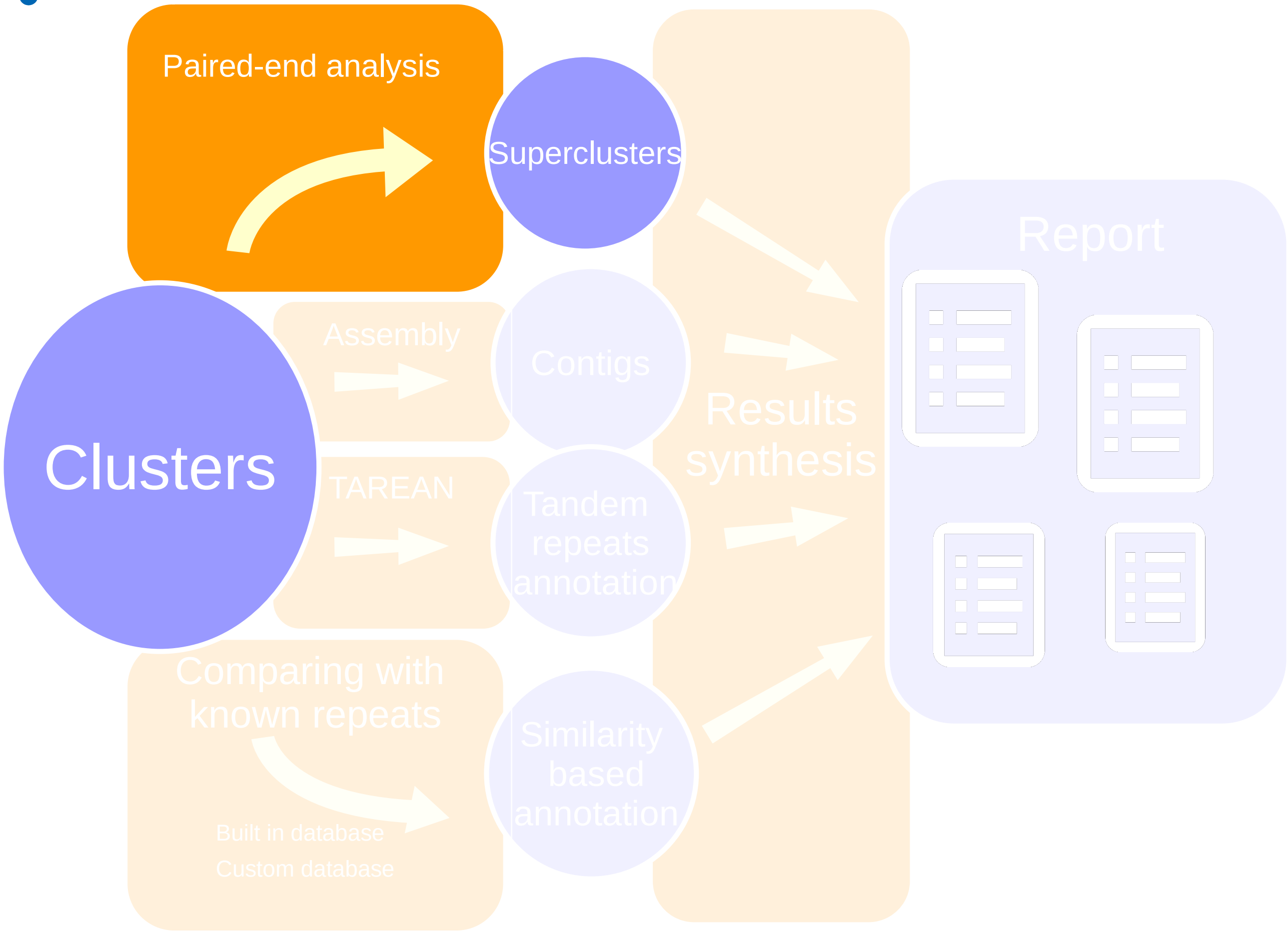
- Clusters
  - Top clusters
  - Small clusters
- Singlets

RepeatExplorer pipeline

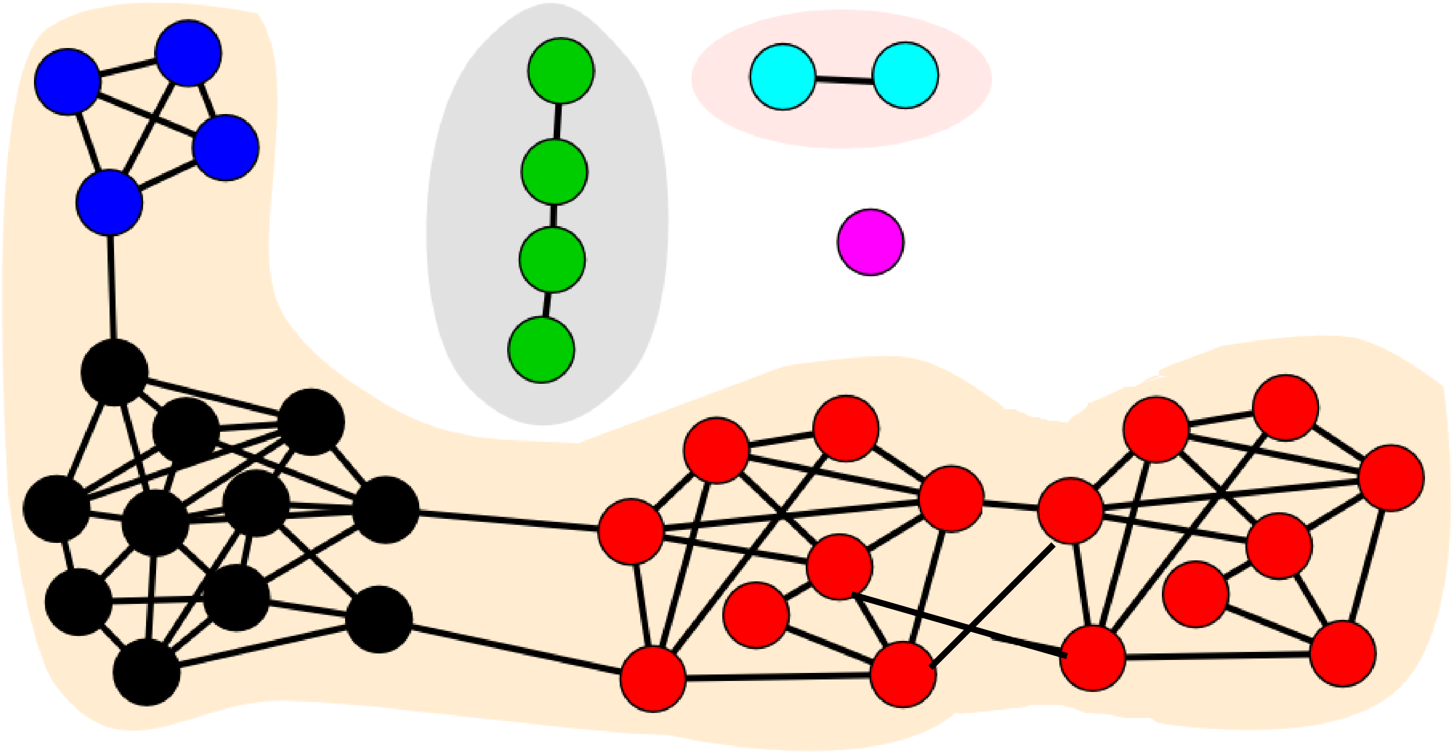
Cluster centered analysis



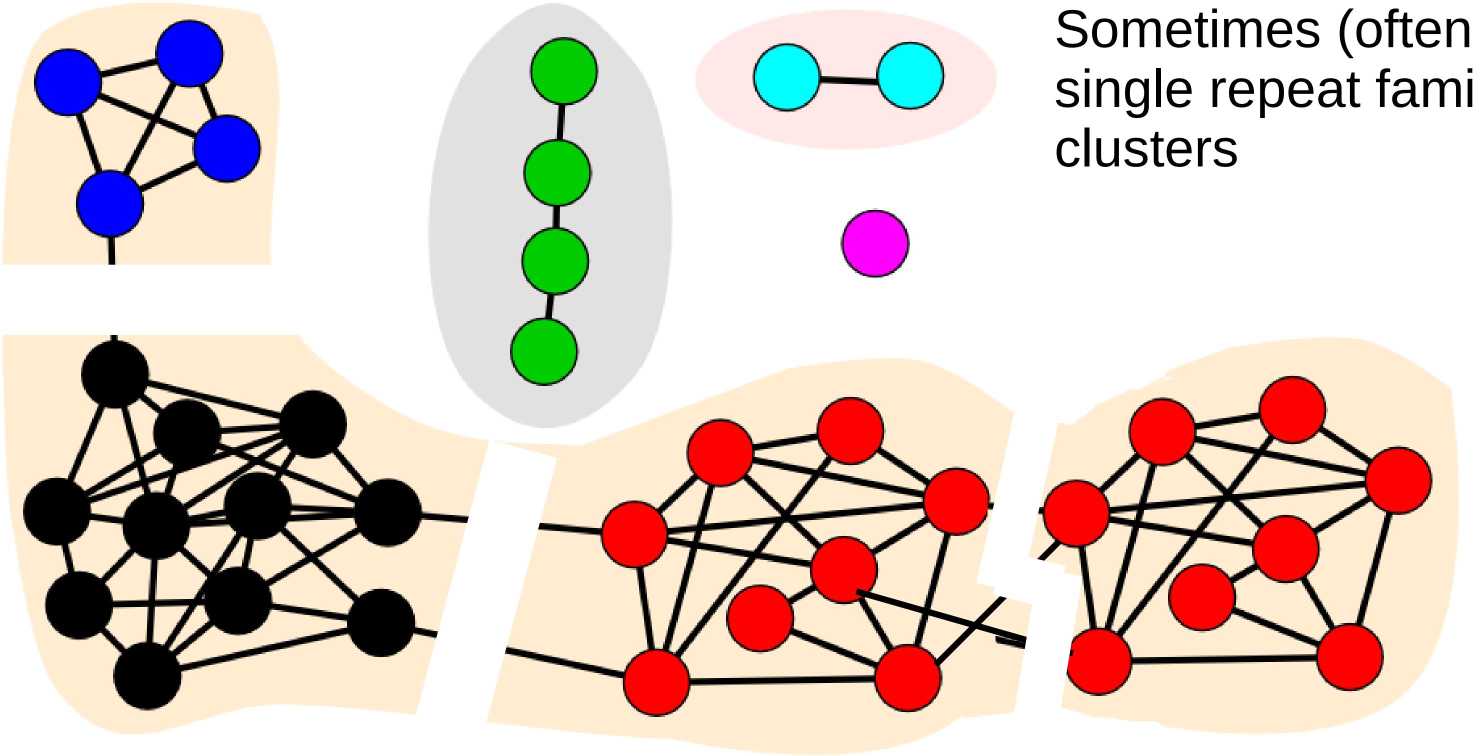
RepeatExplorer pipeline



Clusters and Superclusters



Clusters and Superclusters

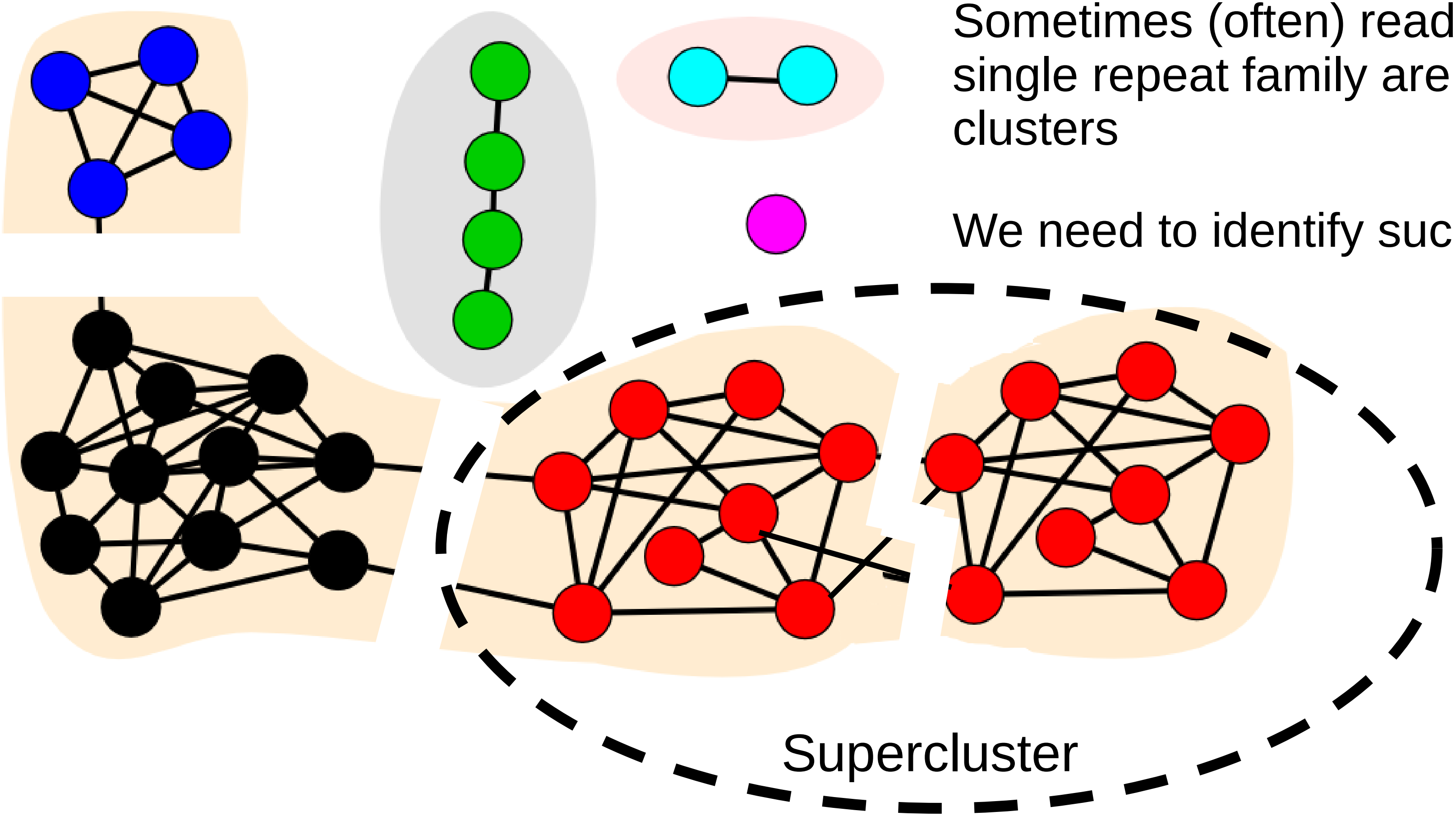


Sometimes (often) reads which belong to single repeat family are split into multiple clusters

Clusters and Superclusters

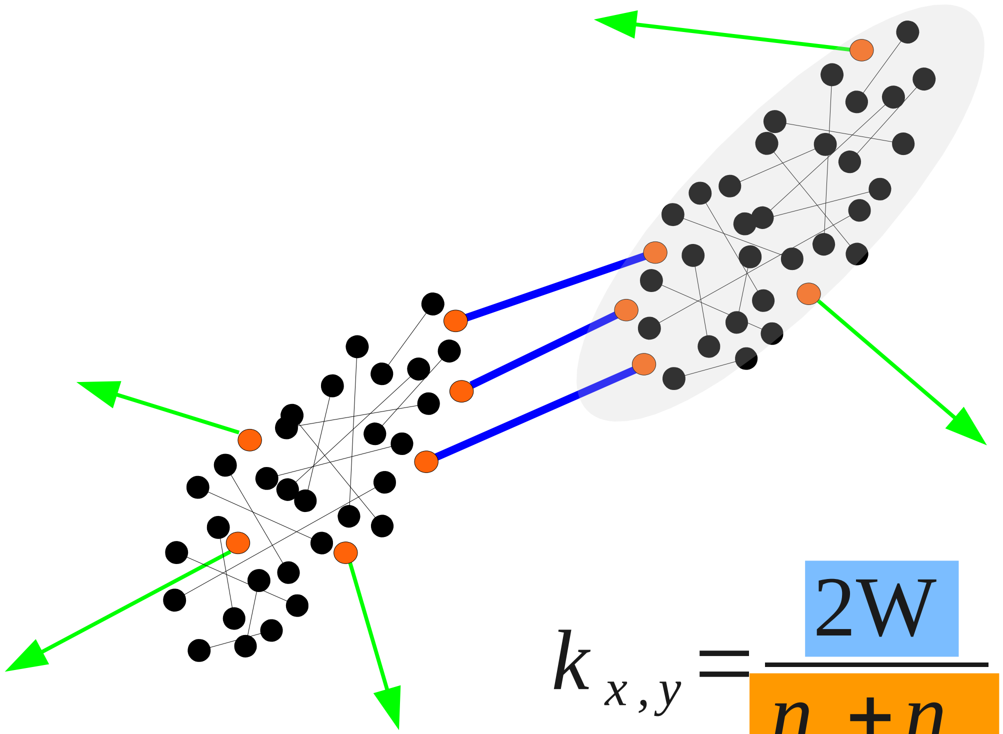
Sometimes (often) reads which belong to single repeat family are split into multiple clusters

We need to identify such false splits



Clusters and Superclusters

Identification of supercluster using paired-end reads



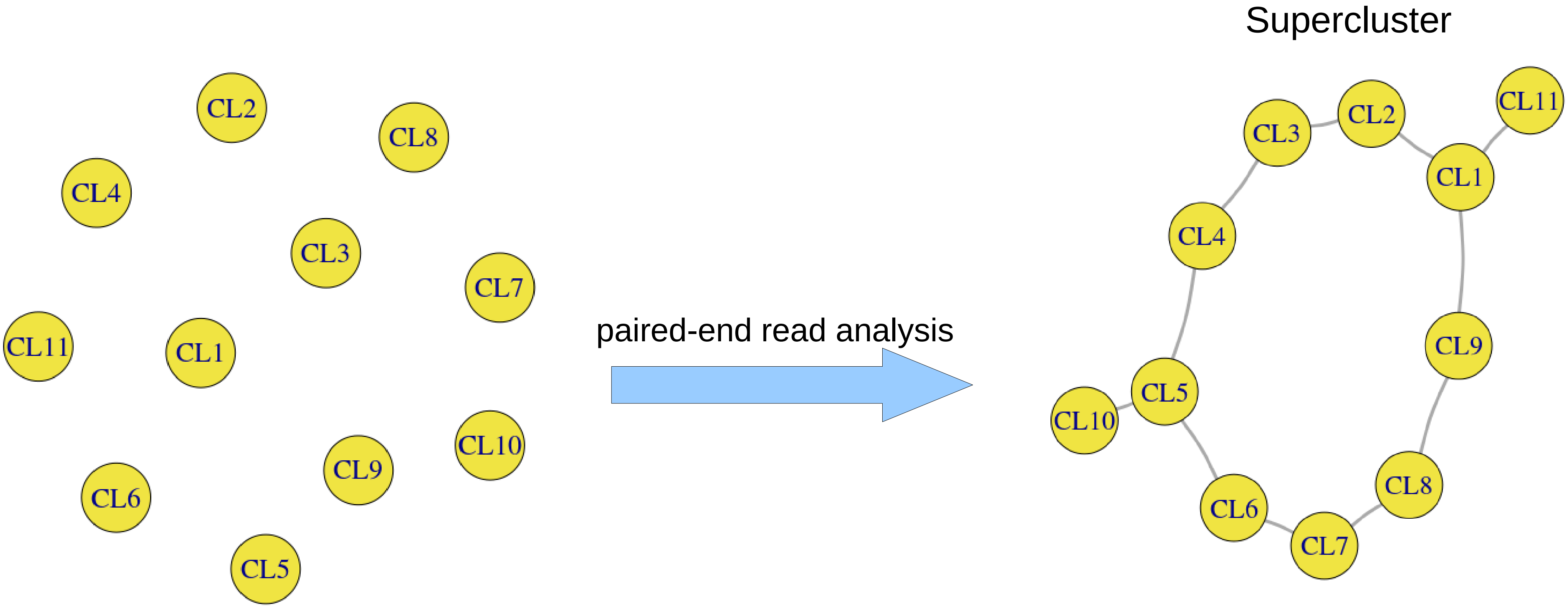
W number of reads pairs shared between clusters **x** and **y**
n*$_x$** and ***n*$_y$** is number of reads in cluster ***x and cluster ***y*** with absent read mate within the same cluster respectively

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

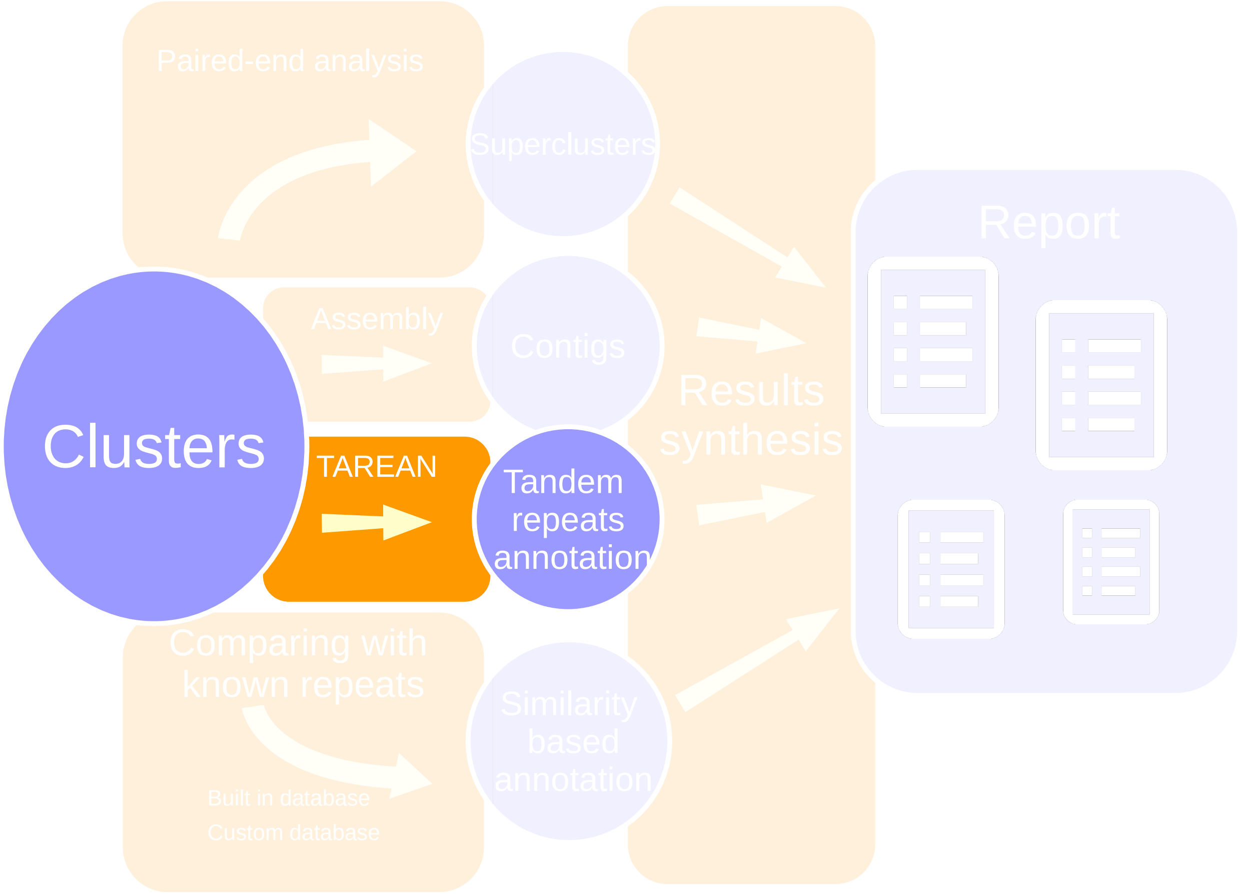
Suitable $\boldsymbol{k_{x,y}}$ cutoff 0.05 – 0.2
full connection: $\boldsymbol{k_{x,y}=1}$
no connection $\boldsymbol{k_{x,y}= 0}$

Clusters and Superclusters

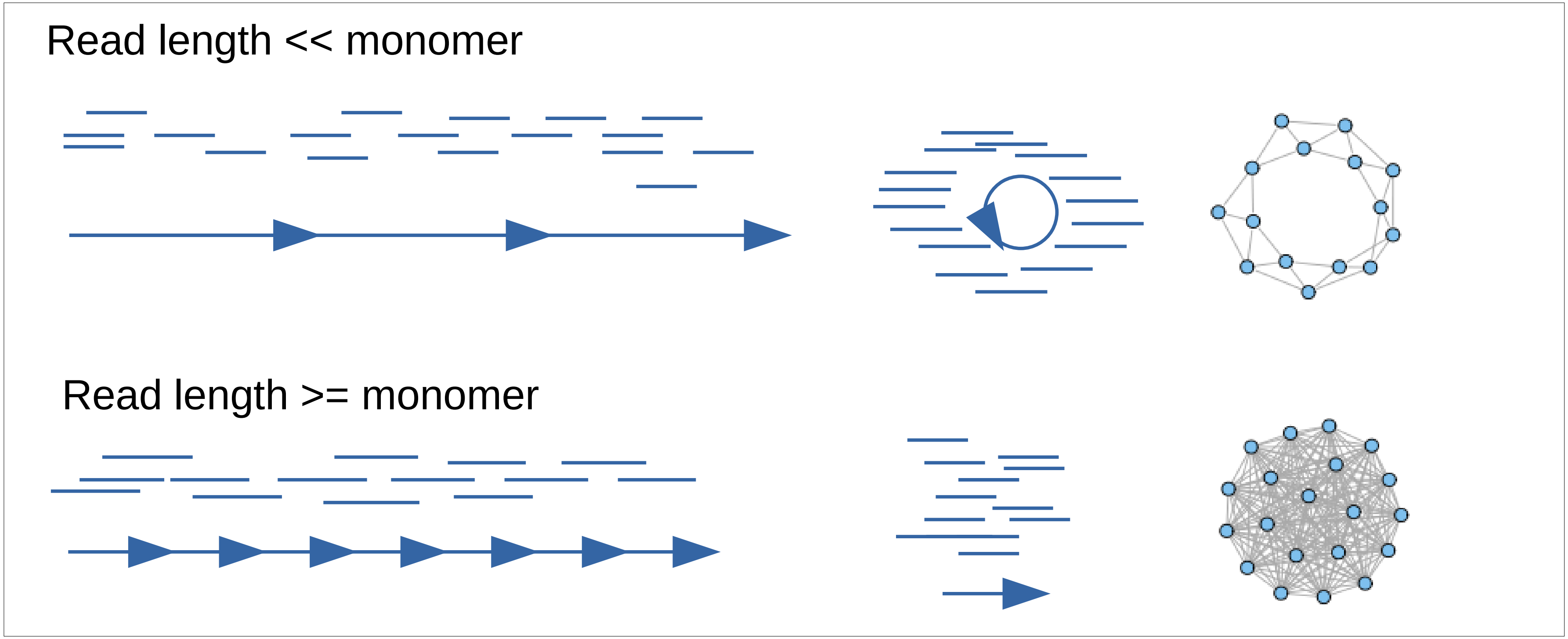
Identification of supercluster using paired-end reads



RepeatExplorer pipeline



Tandem Repeat Analyzer - TAREAN



Read length << monomer

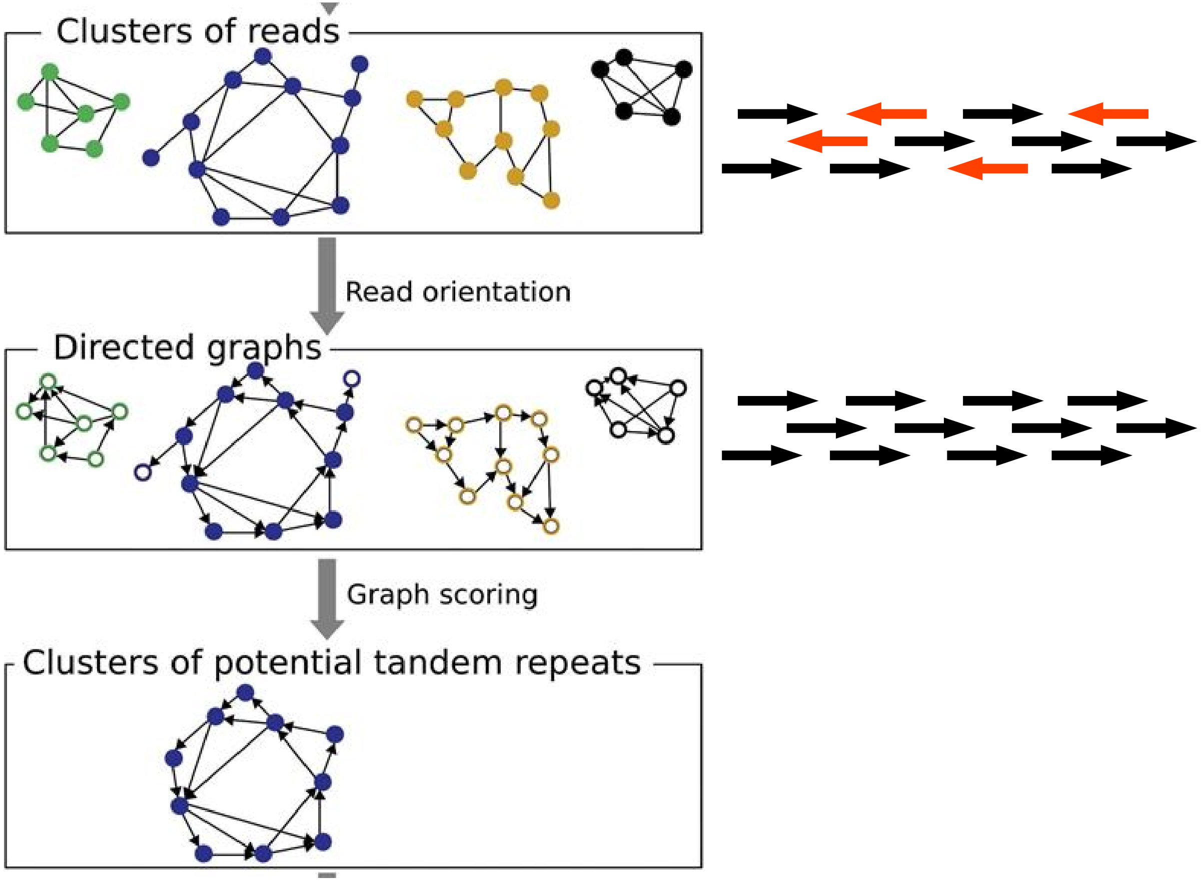
Read length >= monomer

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

Tandem Repeat Analyzer - TAREAN

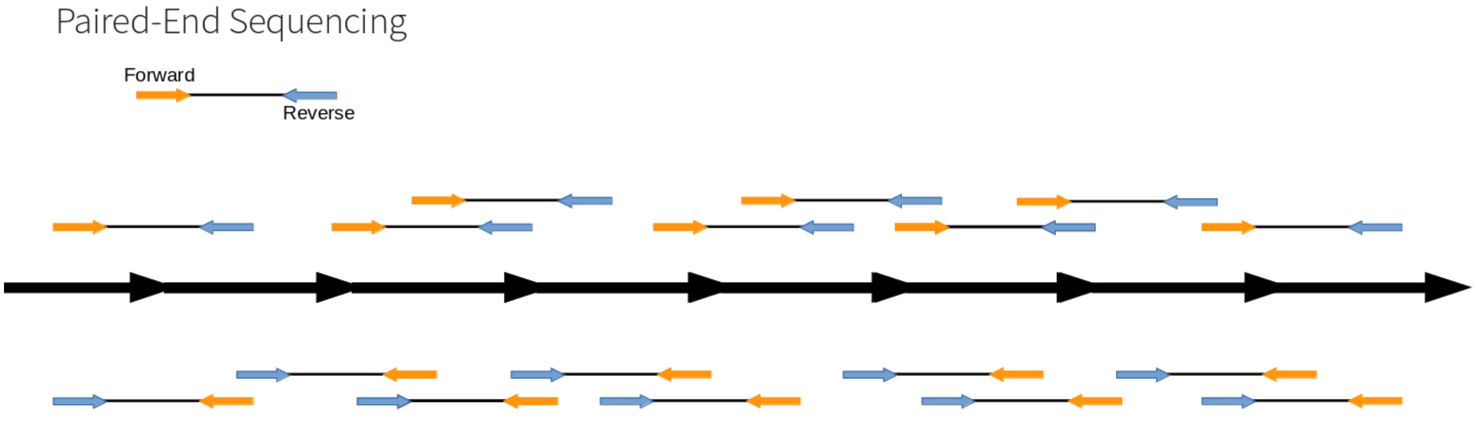
Principle

- All reads within cluster are set to have same 3’to5’ orientation with hypothetical tandem repeat monomer
- Directed graph is constucted
- Larges circular structures are detected (a.k.a strongly connected component)
- **Connected Component index**

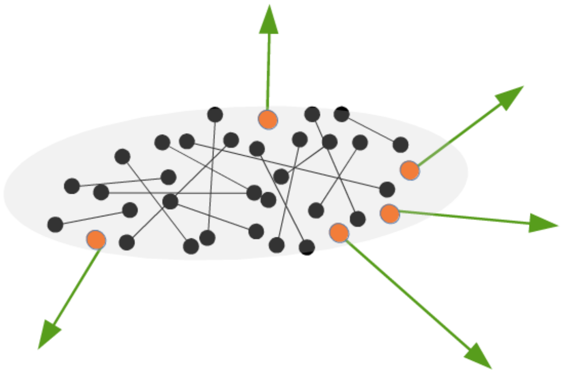


Tandem Repeat Analyzer - TAREAN

Principle



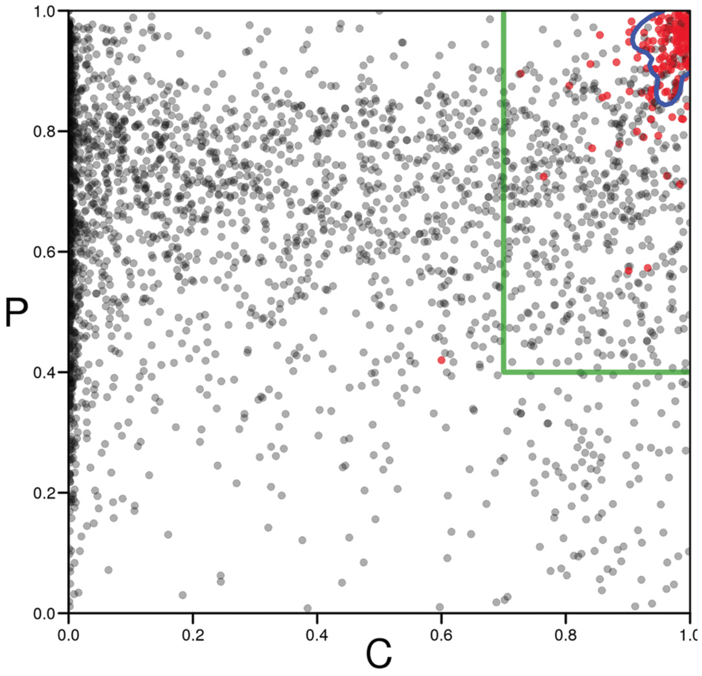
Pair completeness = fraction of complete pairs in cluster

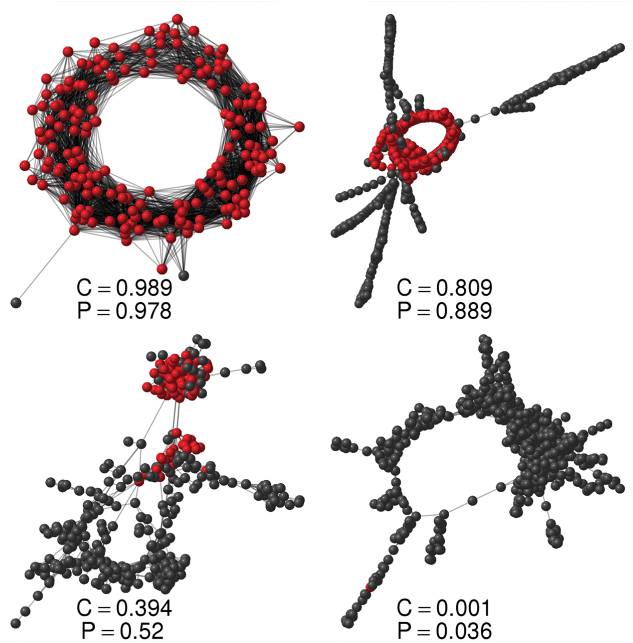


In clusters which are derived from tandem repeat, most of the paired-end reads should be complete

Tandem Repeat Analyzer - TAREAN

Principle





Five groups of clusters by TAREAN

- **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

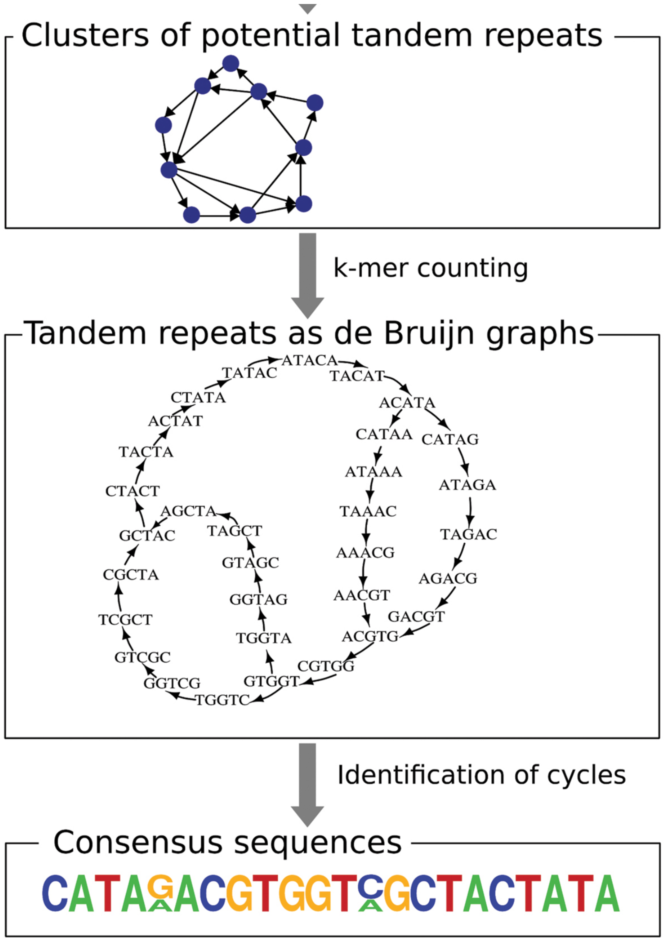
Principle

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



Clusters of potential tandem repeats

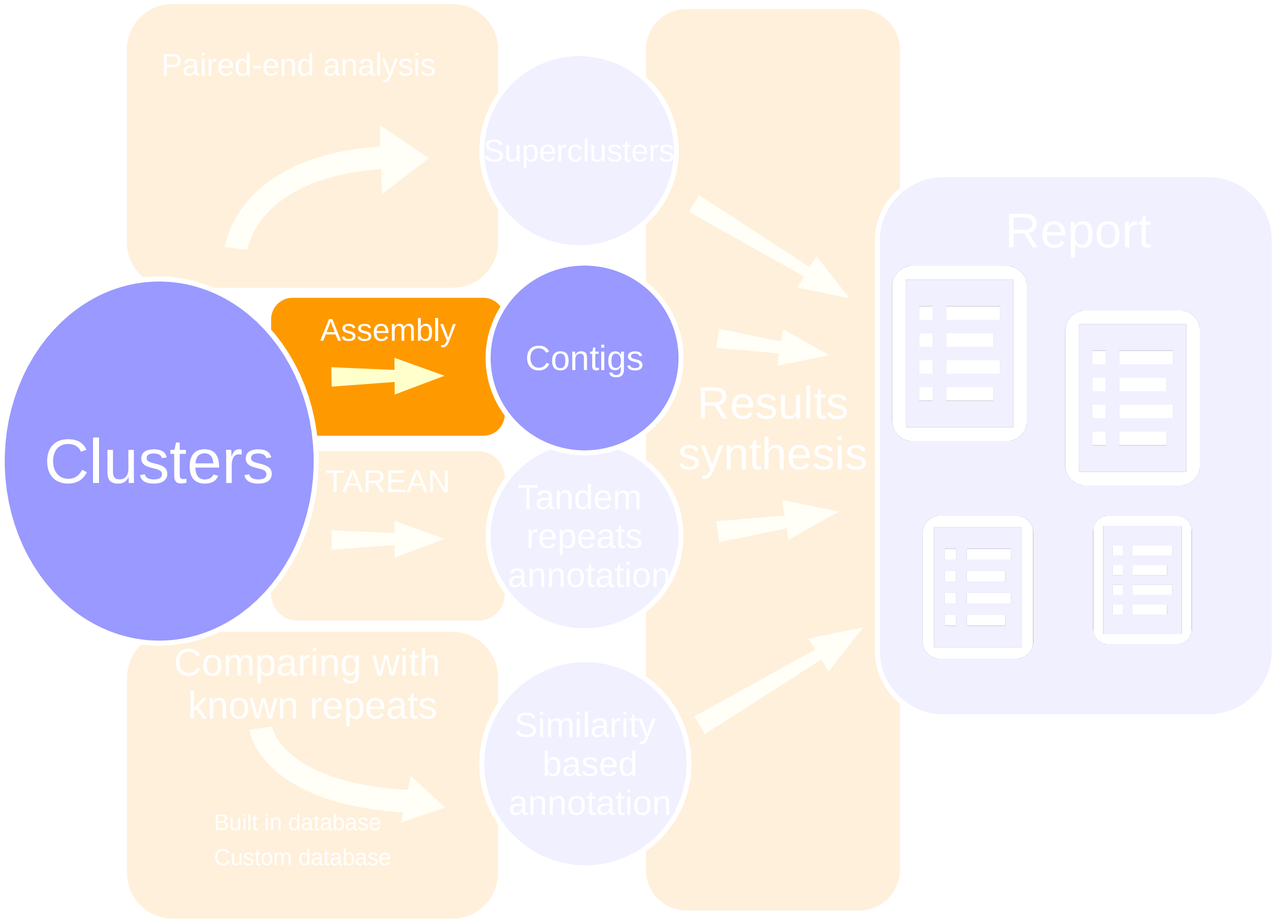
k-mer counting

Tandem repeats as de Bruijn graphs

Identification of cycles

Consensus sequences

RepeatExplorer pipeline

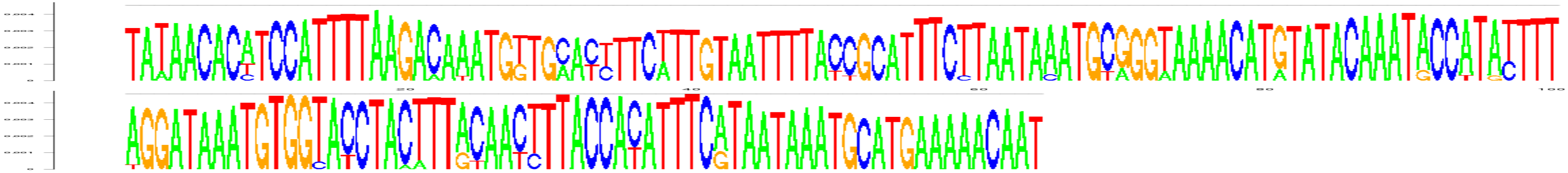


Contig assembly

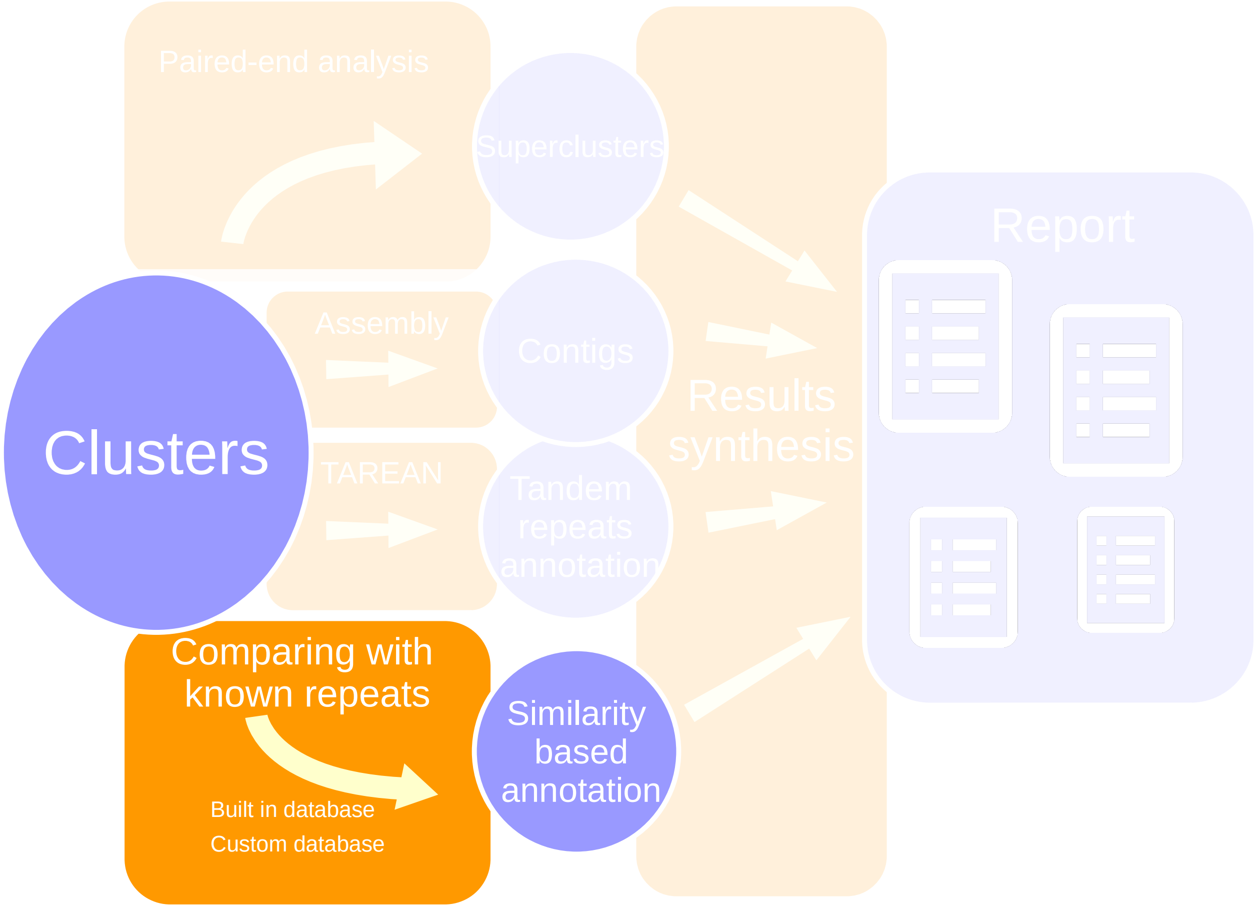
Reads are assembled by CAP3 program, each cluster separately:

```
ACTGTGTCGTCGTCGTCGTGTG
       CGTCGTCG-CGTGTGGT            Reads
              GTCGTGTG-TTGTCGTCTGA
ACTGTGTCGTCGTCGTCGTGTGGTTGTCGTCTGA  Contig
```

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



RepeatExplorer pipeline



All **reads** are compared with:

- Database of protein domains (REXdb)
- DNA database
  - rDNA, tRNA
  - Organele DNA
  - potential contaminants
- Custom database (optional)

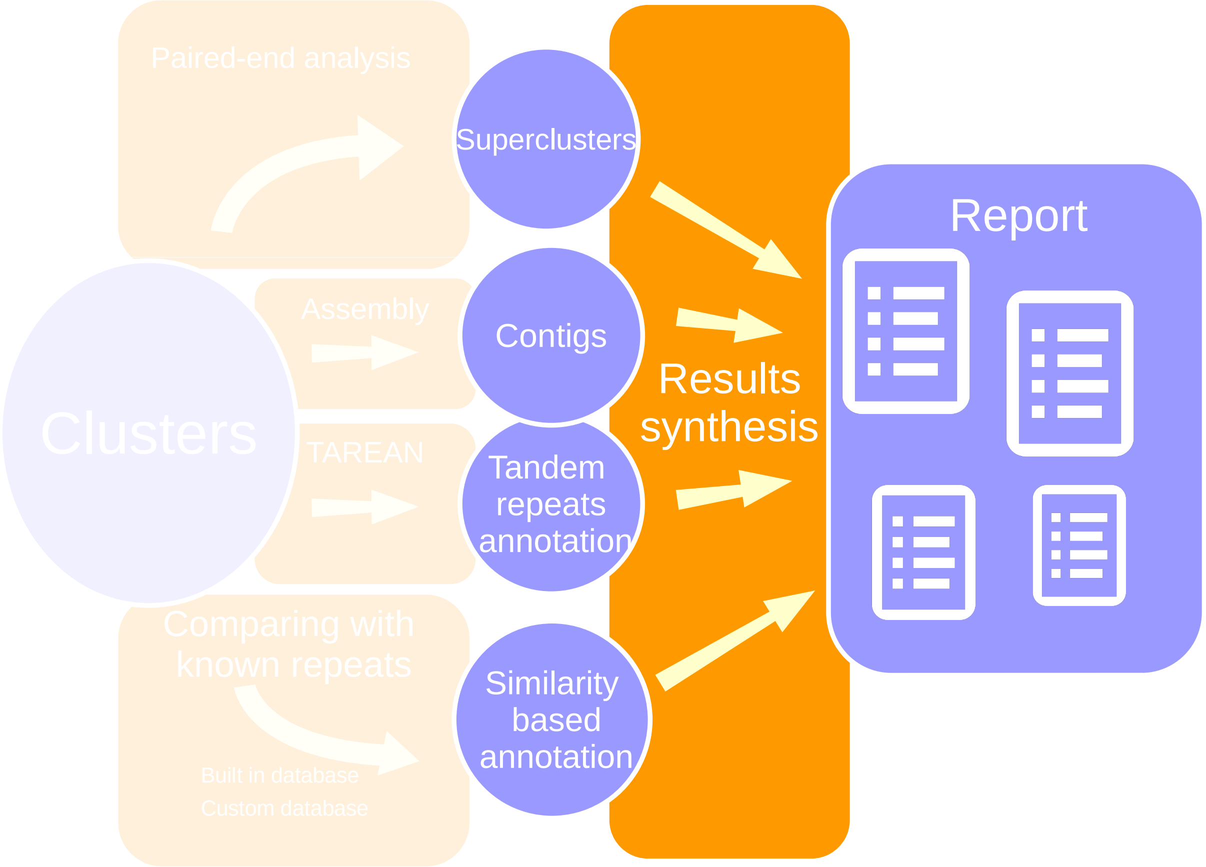
Similarity search

Database sequence classification	**Protein domain**	**Number of reads with similarity hit**	**Proportion** **No of reads / cluster size**
mitochondria		25	0.0023
Ogre Ty3-RH	Ty3-RH	2977	0.27402
Retand Ty3-RH	Ty3-RH	2	0.00018
Ogre Ty3-RT	Ty3-RT	3473	0.31968
Ogre Ty3-aRH	Ty3-aRH	1713	0.15768

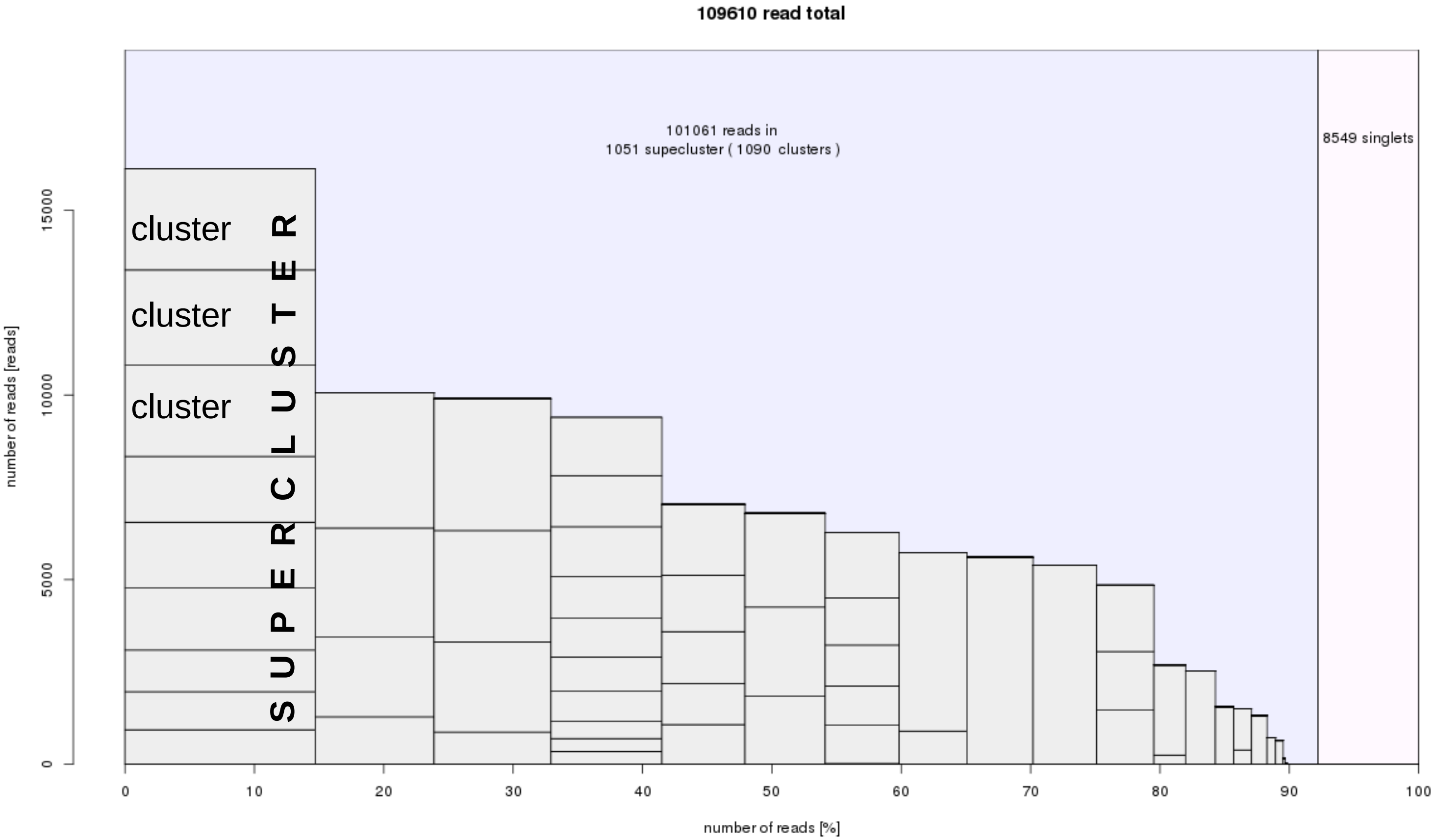
RepeatExplorer pipeline

Reporting:

- HTML reports
- Visualization
- Automatic classification

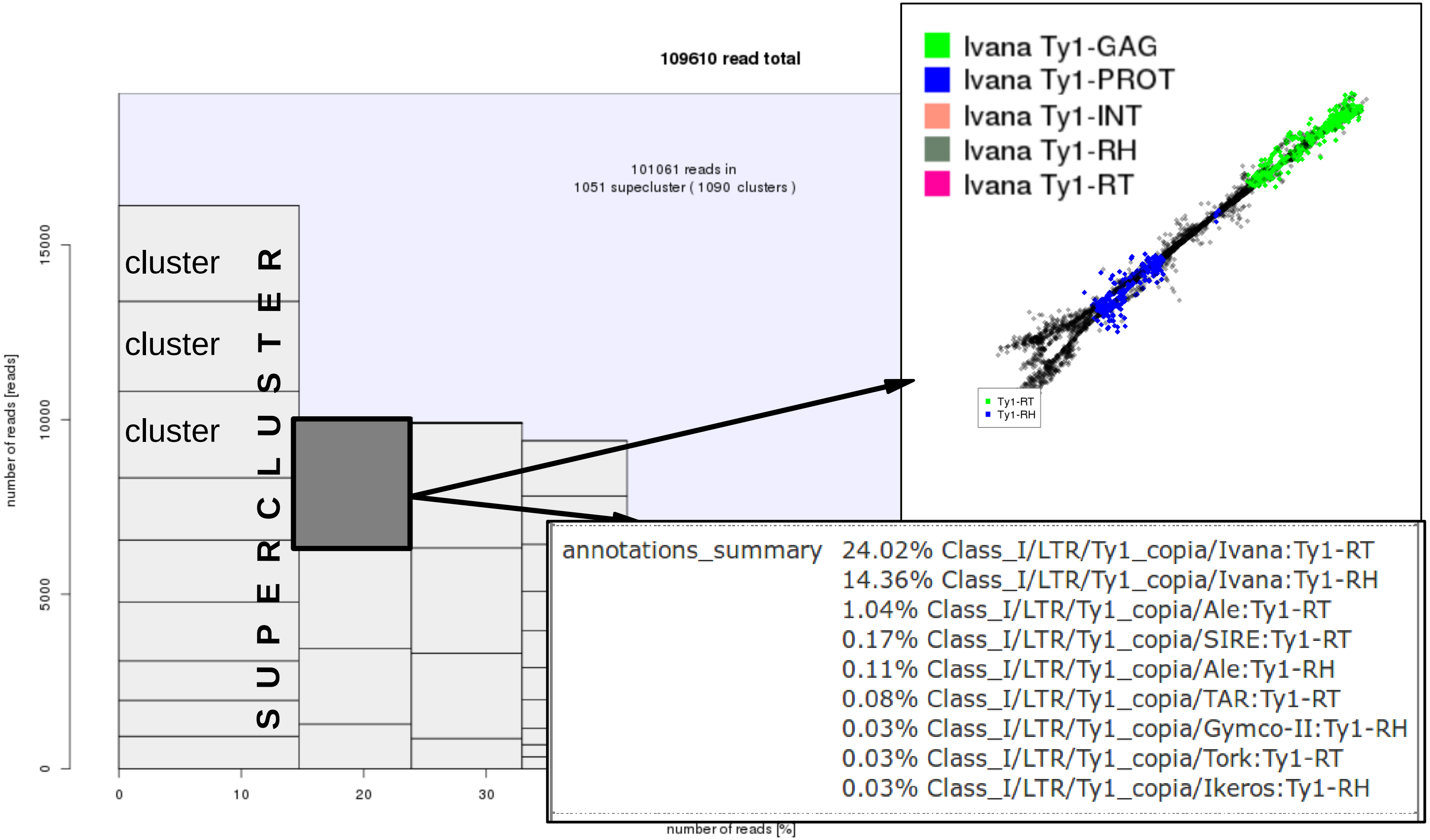


Report

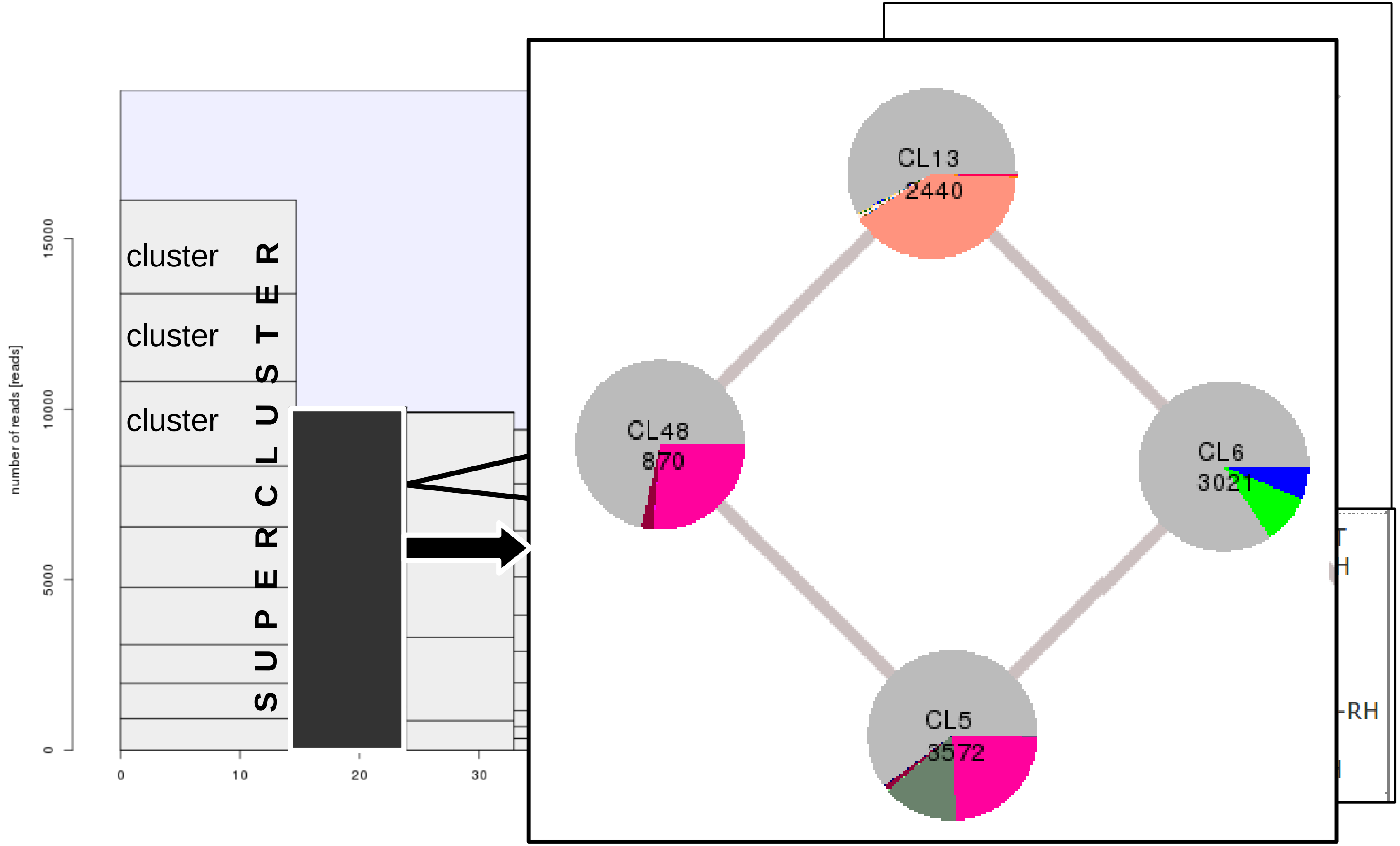


109610 read total
101061 reads in
1051 supecluster (1090 clusters)
8549 singlets
cluster
cluster
cluster
SUPERCLUSTER
number of reads [reads]
0
5000
10000
15000
0
10
20
30
40
50
60
70
80
90
100
number of reads [%]

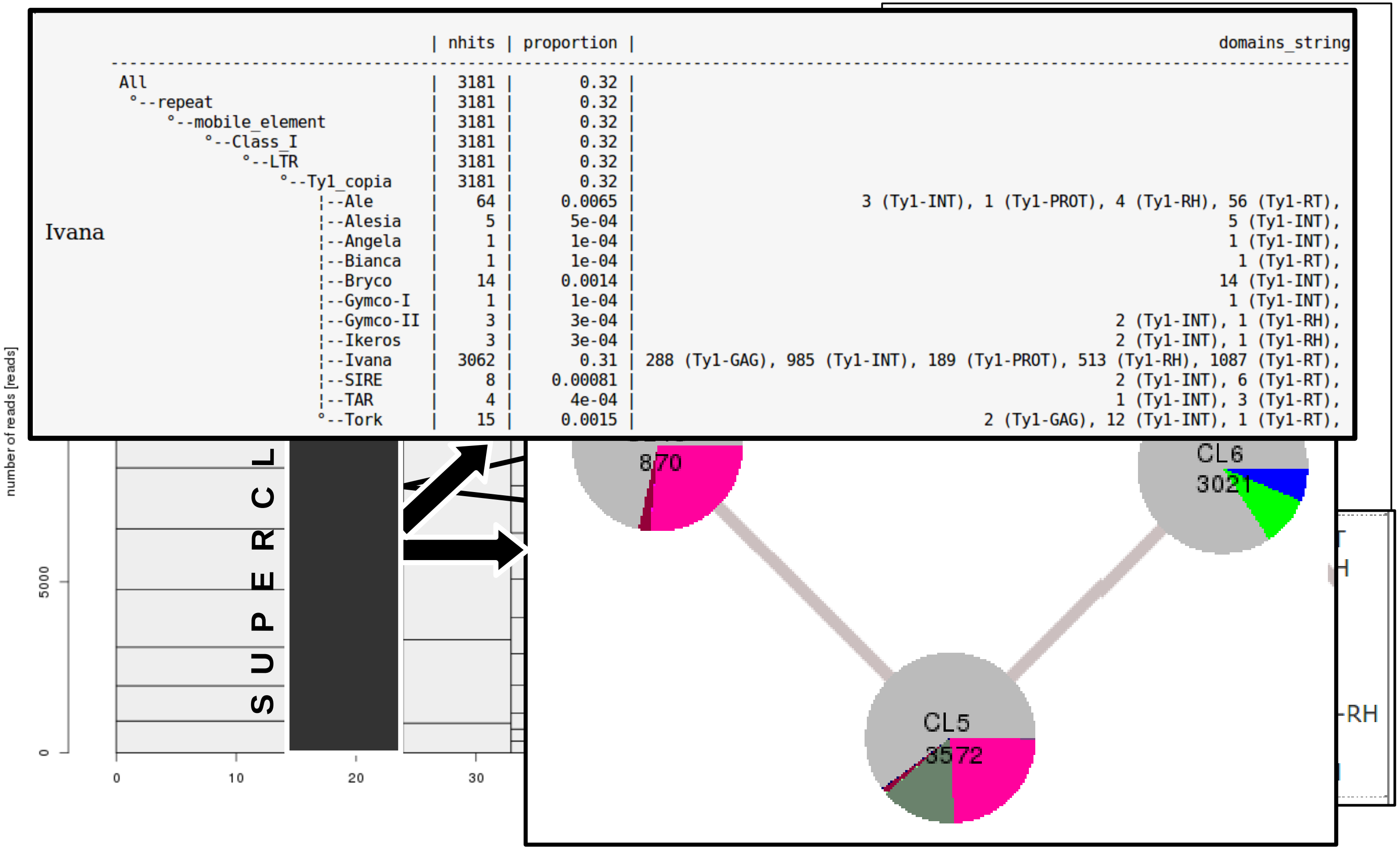
Report



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


number of reads [reads]
15000
10000
5000
0
0
10
20
30
cluster
cluster
cluster
SUPERCLUSTER
CL13
2440
CL48
870
CL6
3021
CL5
3572
RH

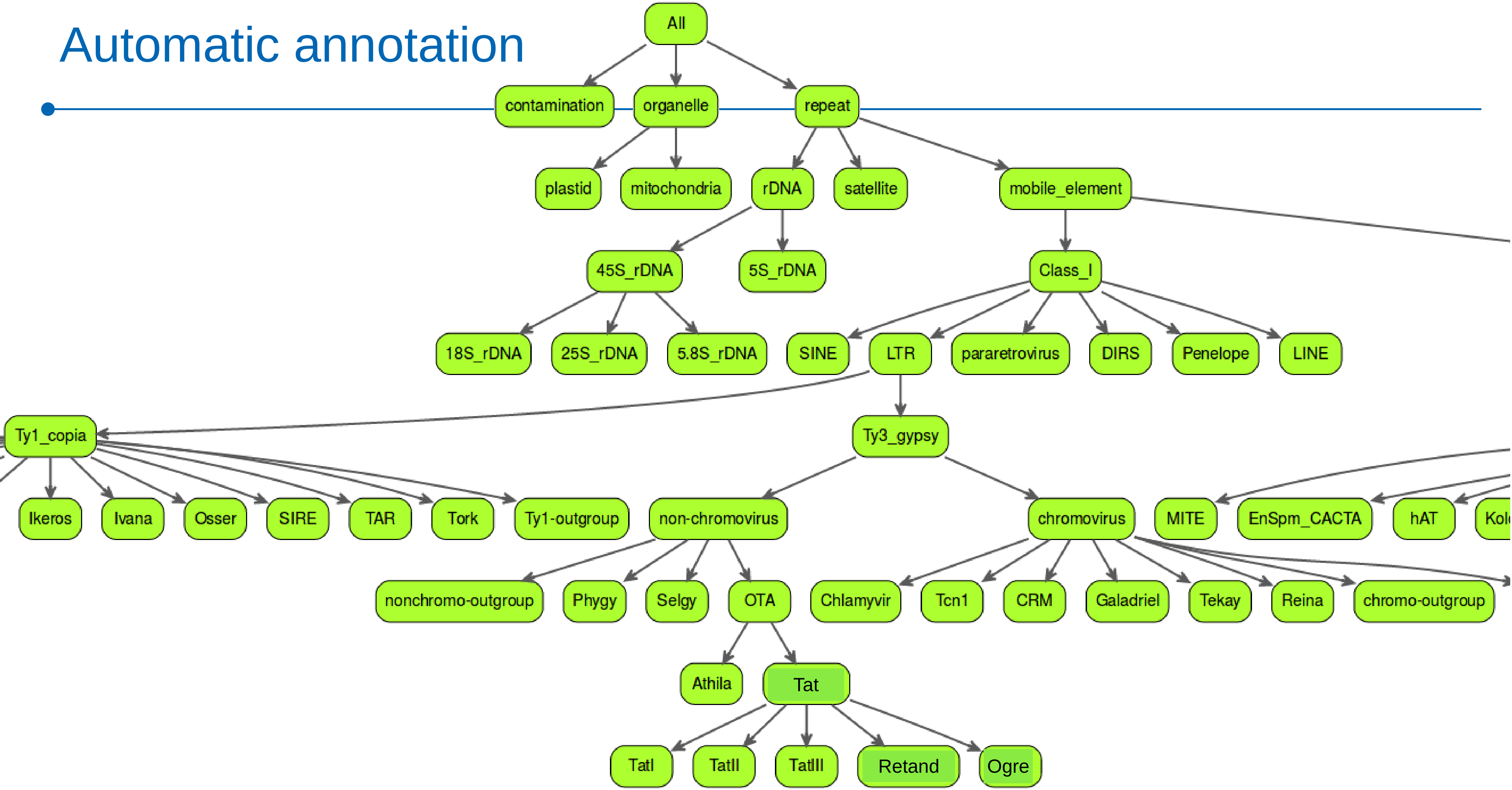
Reporting



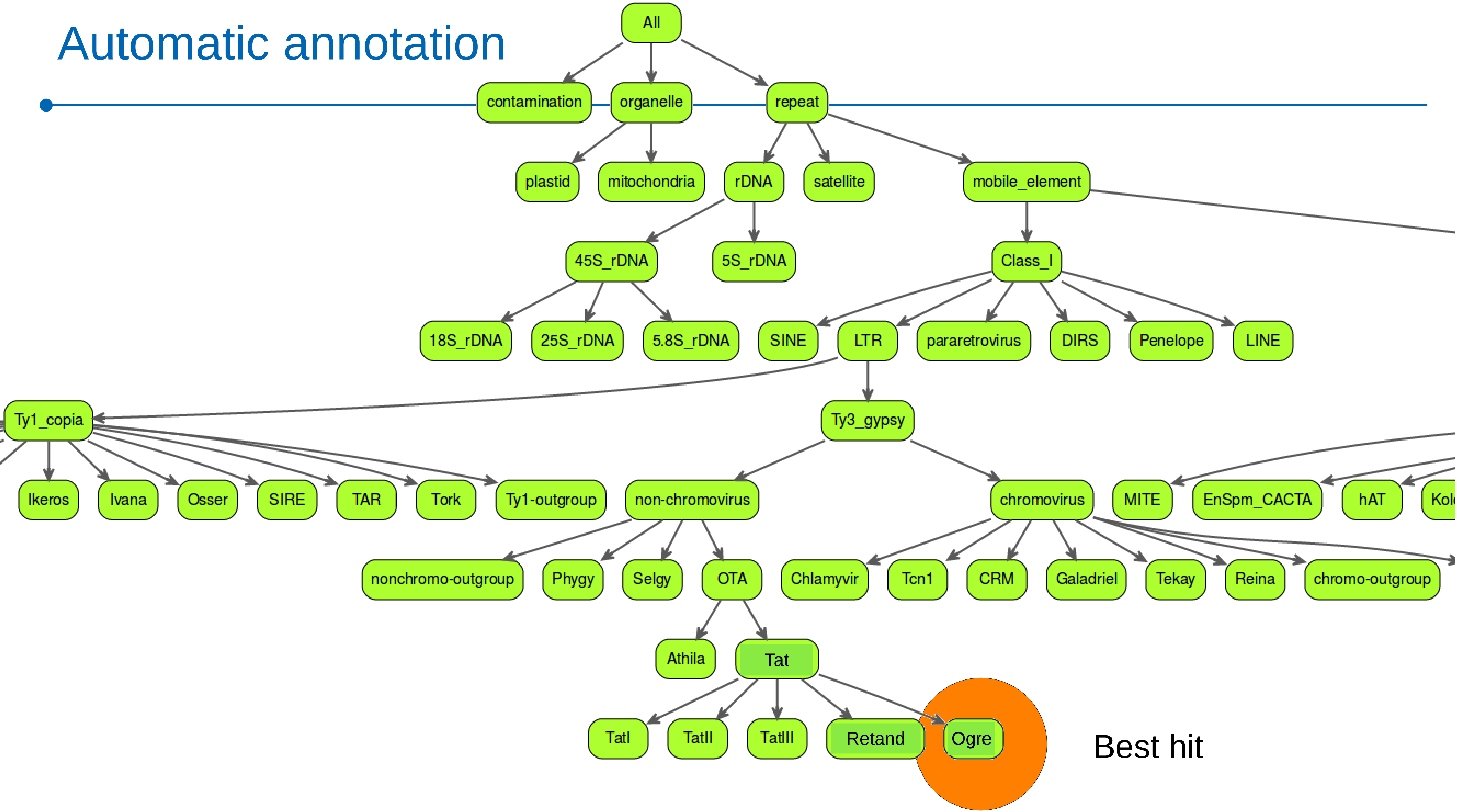
Ivana

```
                            | nhits | proportion |                                                         domains_string
  -------------------------------------------------------------------------------------------------------------------------
  All                       |  3181 |       0.32 |
   °--repeat                |  3181 |       0.32 |
       °--mobile_element    |  3181 |       0.32 |
           °--Class_I       |  3181 |       0.32 |
               °--LTR       |  3181 |       0.32 |
                   °--Ty1_copia  |  3181 |       0.32 |
                       ¦--Ale     |    64 |     0.0065 |                  3 (Ty1-INT), 1 (Ty1-PROT), 4 (Ty1-RH), 56 (Ty1-RT),
                       ¦--Alesia  |     5 |      5e-04 |                                                      5 (Ty1-INT),
                       ¦--Angela  |     1 |      1e-04 |                                                      1 (Ty1-INT),
                       ¦--Bianca  |     1 |      1e-04 |                                                       1 (Ty1-RT),
                       ¦--Bryco   |    14 |     0.0014 |                                                     14 (Ty1-INT),
                       ¦--Gymco-I |     1 |      1e-04 |                                                      1 (Ty1-INT),
                       ¦--Gymco-II |    3 |      3e-04 |                                         2 (Ty1-INT), 1 (Ty1-RH),
                       ¦--Ikeros  |     3 |      3e-04 |                                         2 (Ty1-INT), 1 (Ty1-RH),
                       ¦--Ivana   |  3062 |       0.31 | 288 (Ty1-GAG), 985 (Ty1-INT), 189 (Ty1-PROT), 513 (Ty1-RH), 1087 (Ty1-RT),
                       ¦--SIRE    |     8 |    0.00081 |                                         2 (Ty1-INT), 6 (Ty1-RT),
                       ¦--TAR     |     4 |      4e-04 |                                         1 (Ty1-INT), 3 (Ty1-RT),
                       °--Tork    |    15 |     0.0015 |                          2 (Ty1-GAG), 12 (Ty1-INT), 1 (Ty1-RT),
```

Automatic annotation

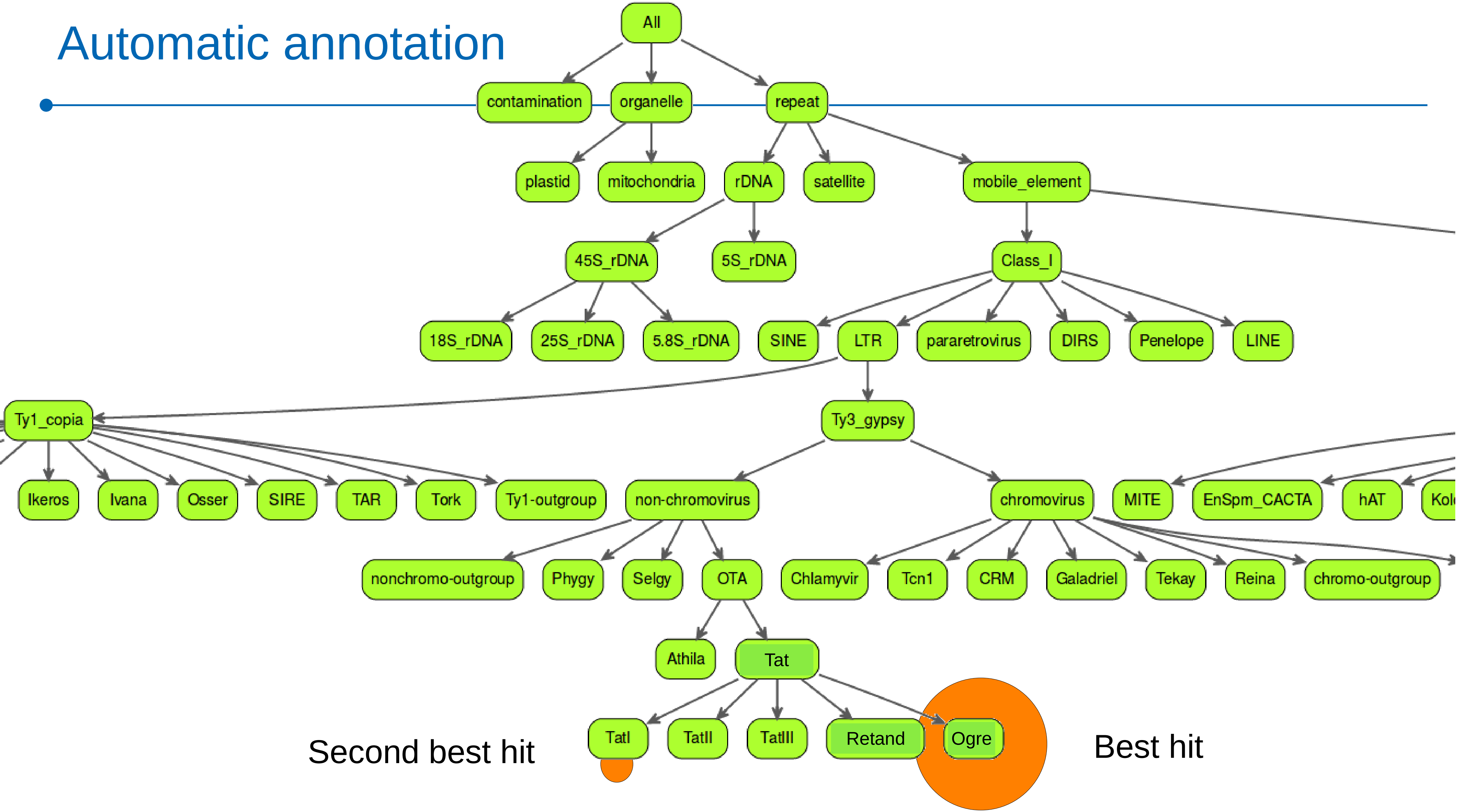


Automatic annotation



Best hit

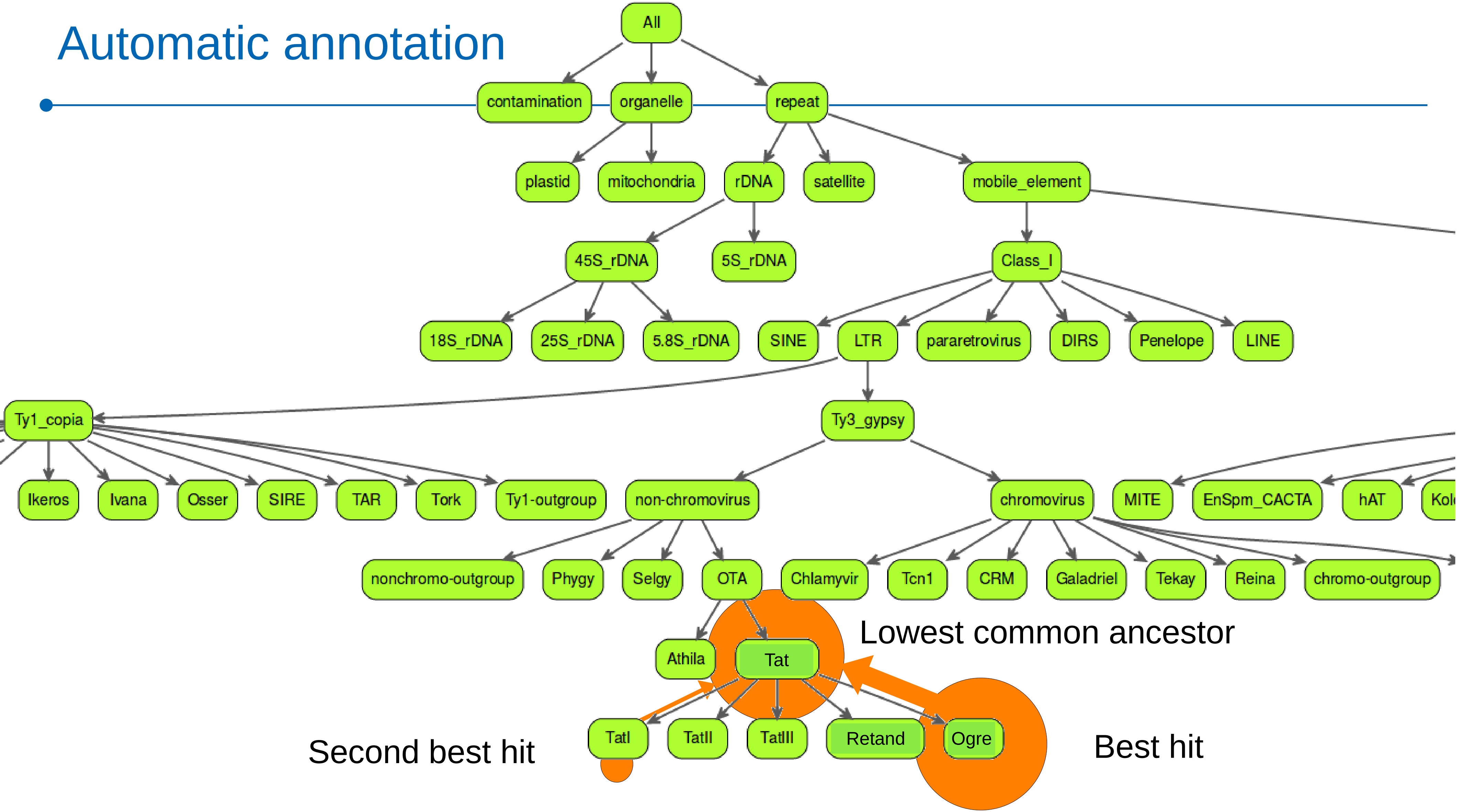
Automatic annotation



Second best hit

Best hit

Automatic annotation

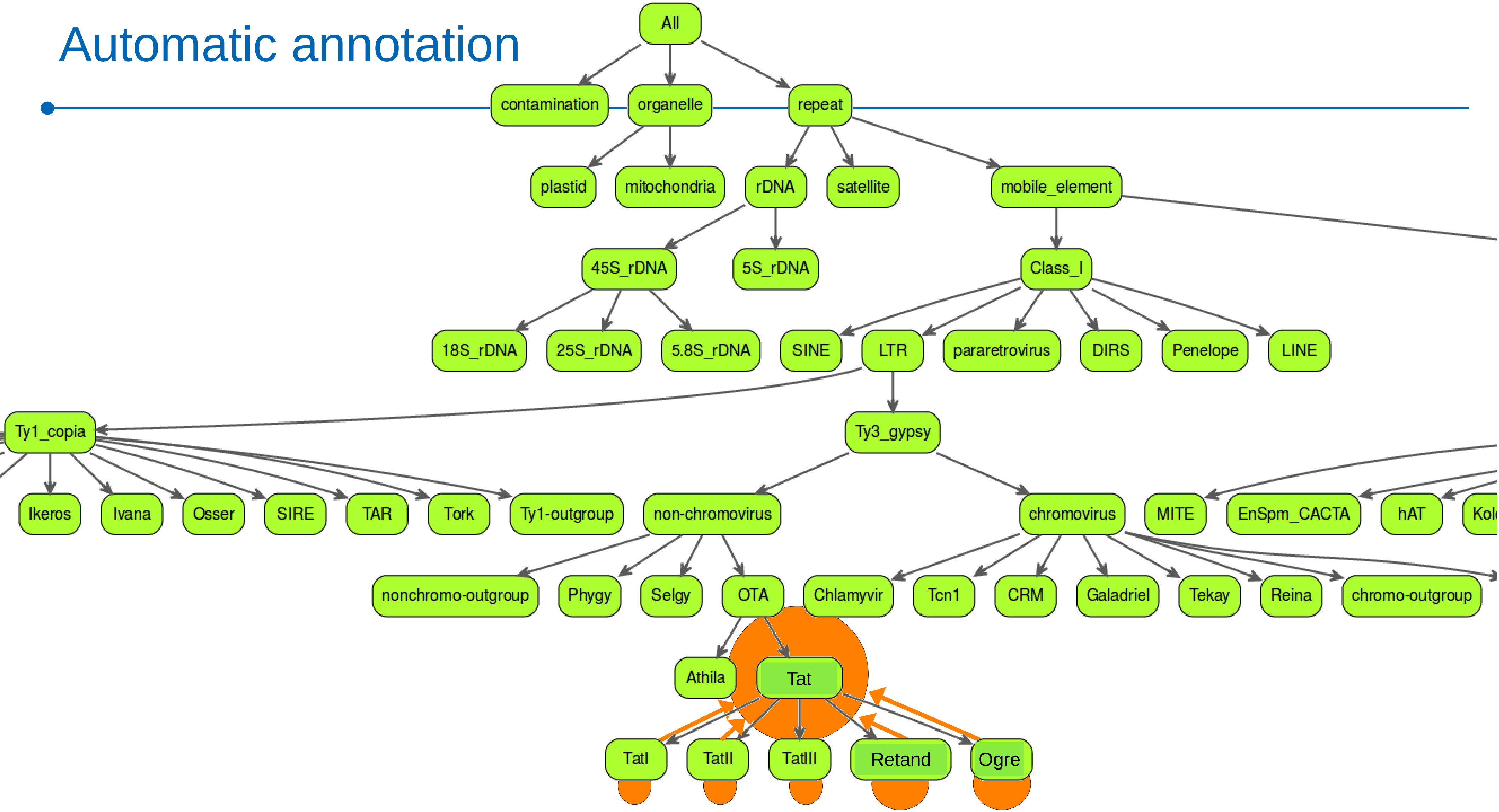


Lowest common ancestor

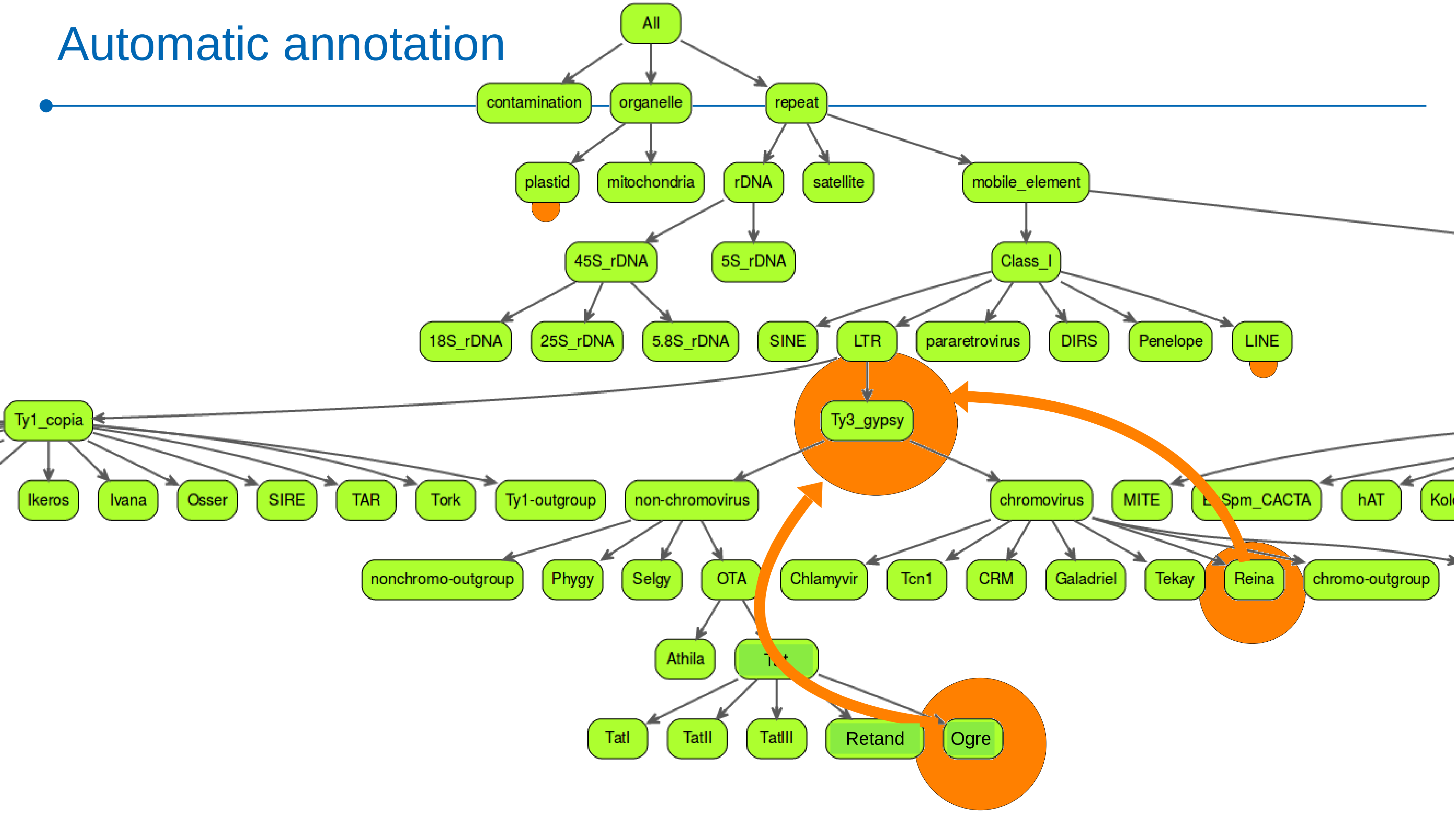
Second best hit

Best hit

Automatic annotation

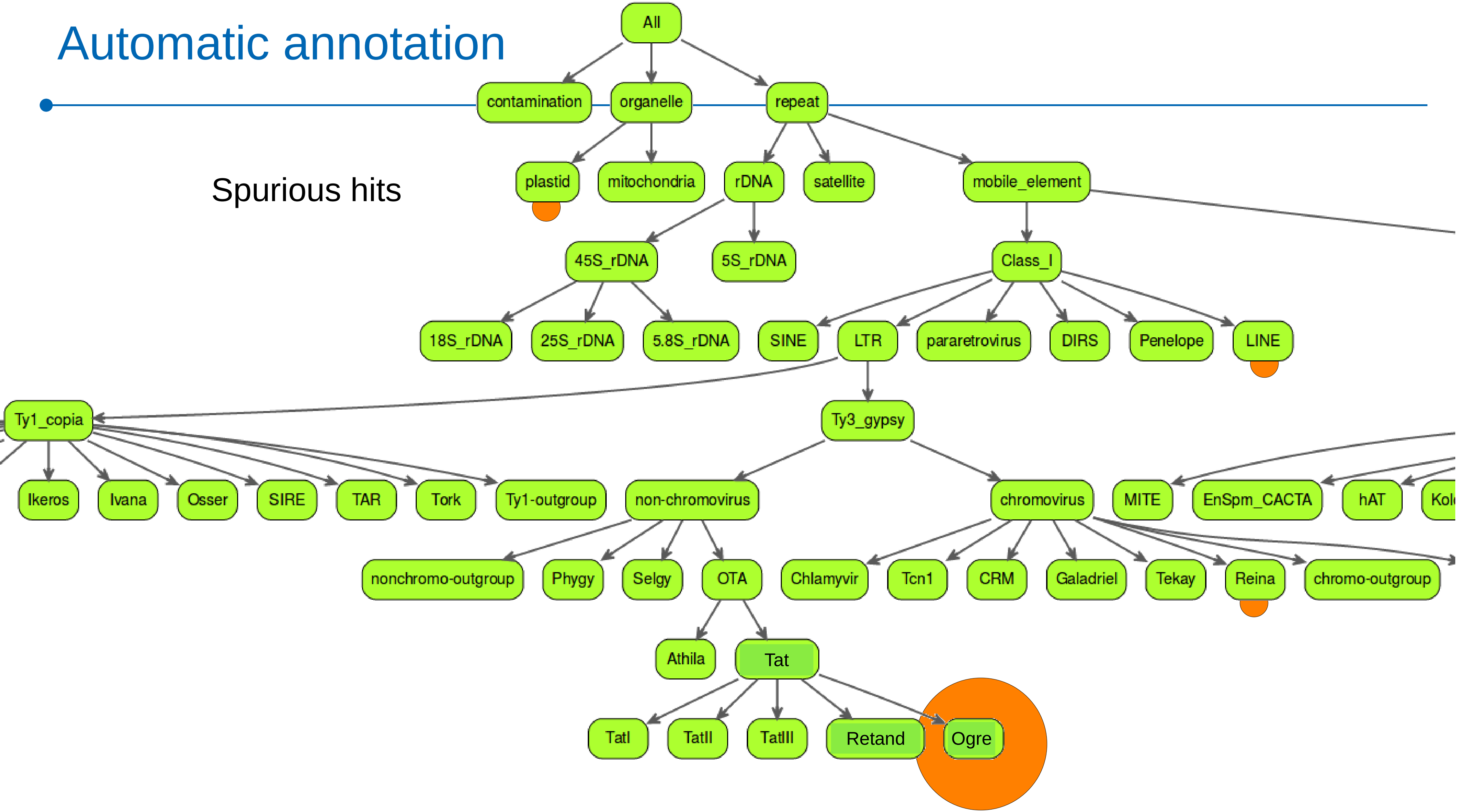


Automatic annotation

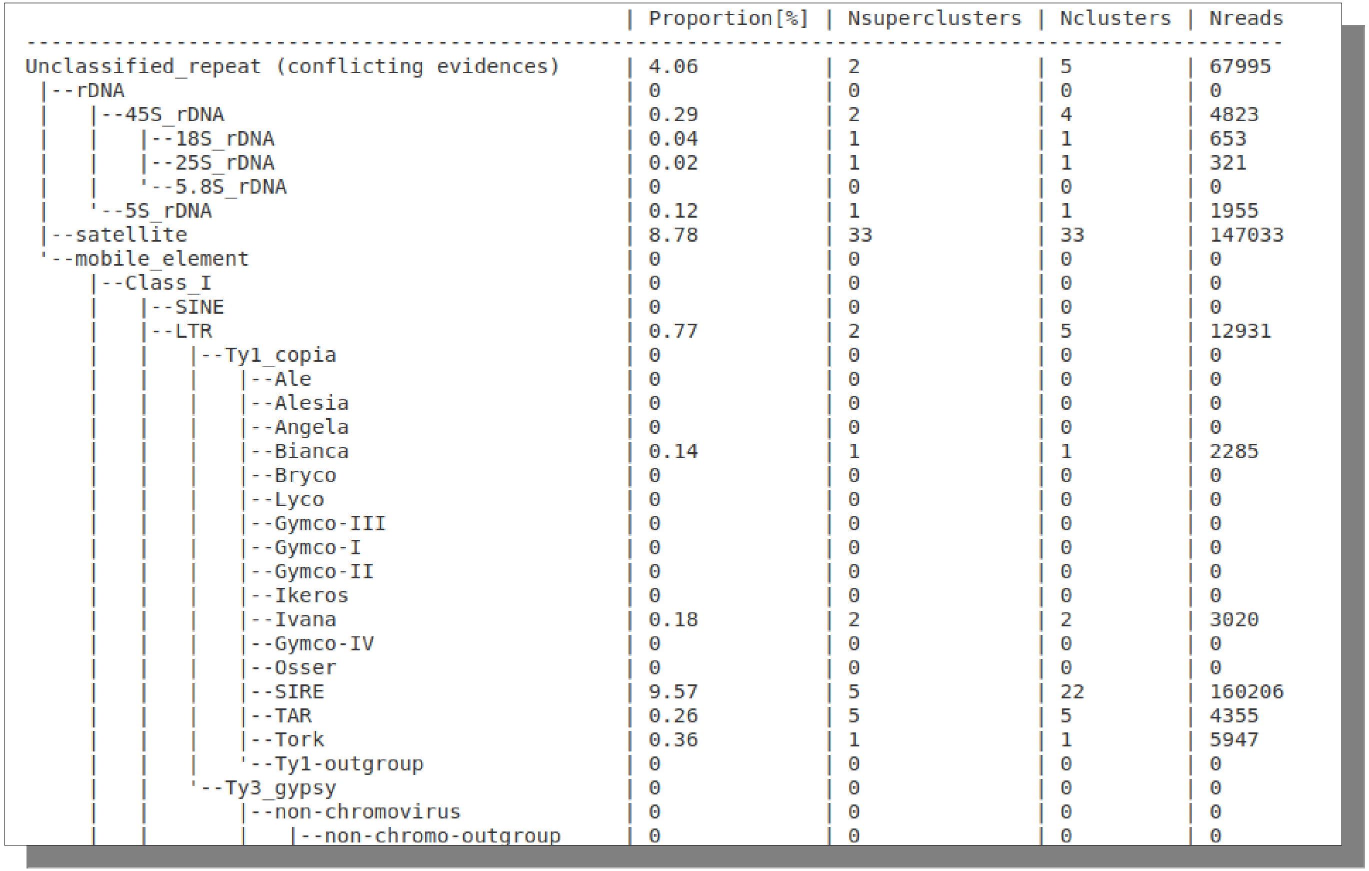


Automatic annotation

Spurious hits

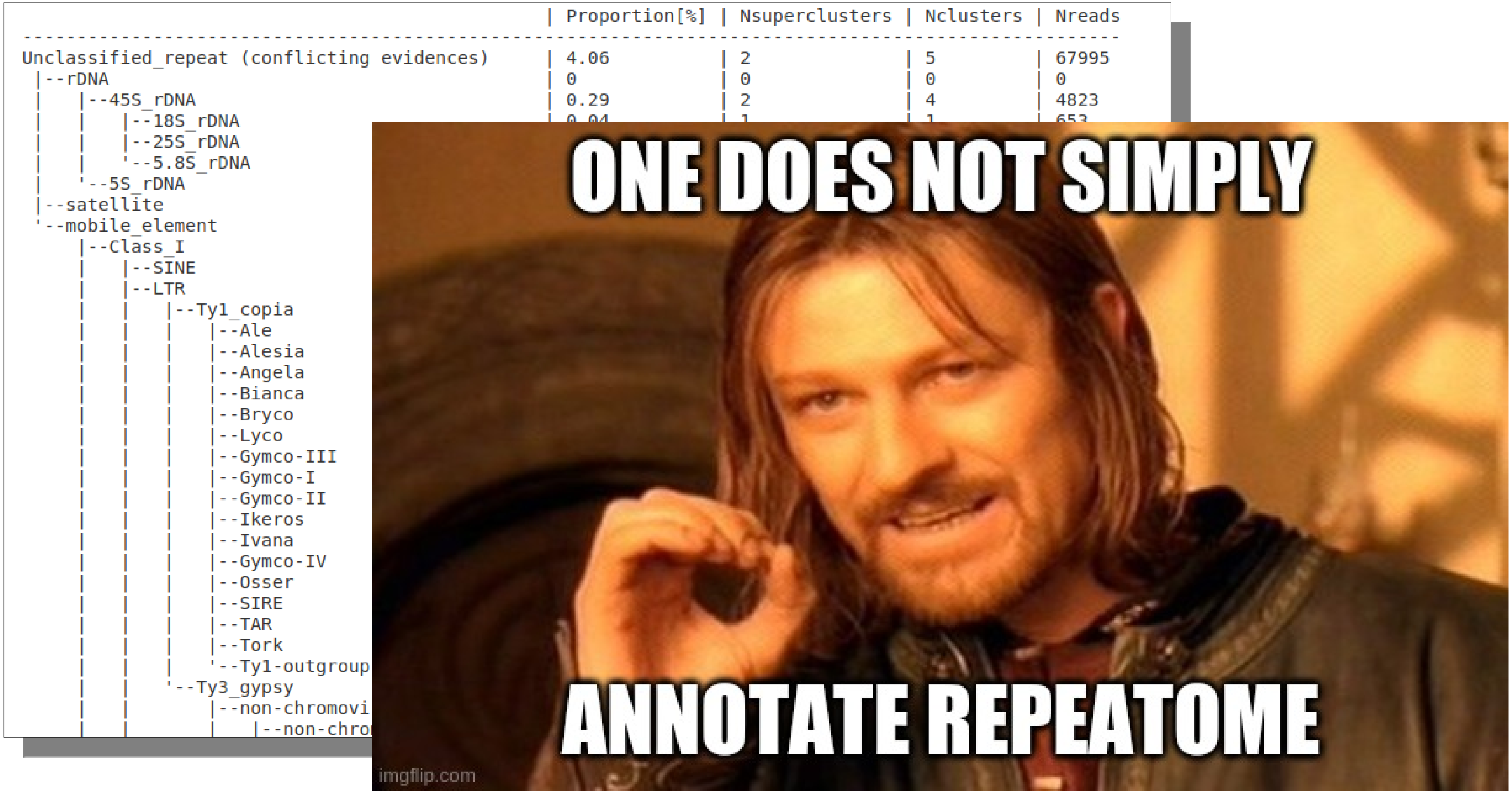


Automatic annotation



```
                                                | Proportion[%] | Nsuperclusters | Nclusters | Nreads
-----------------------------------------------------------------------------------------------------
Unclassified_repeat (conflicting evidences)     | 4.06          | 2              | 5         | 67995
 |--rDNA                                        | 0             | 0              | 0         | 0
 |   |--45S_rDNA                                | 0.29          | 2              | 4         | 4823
 |   |   |--18S_rDNA                            | 0.04          | 1              | 1         | 653
 |   |   |--25S_rDNA                            | 0.02          | 1              | 1         | 321
 |   |   '--5.8S_rDNA                           | 0             | 0              | 0         | 0
 |   '--5S_rDNA                                 | 0.12          | 1              | 1         | 1955
 |--satellite                                   | 8.78          | 33             | 33        | 147033
 '--mobile_element                              | 0             | 0              | 0         | 0
     |--Class_I                                 | 0             | 0              | 0         | 0
     |   |--SINE                                | 0             | 0              | 0         | 0
     |   |--LTR                                 | 0.77          | 2              | 5         | 12931
     |   |   |--Ty1_copia                       | 0             | 0              | 0         | 0
     |   |   |   |--Ale                         | 0             | 0              | 0         | 0
     |   |   |   |--Alesia                      | 0             | 0              | 0         | 0
     |   |   |   |--Angela                      | 0             | 0              | 0         | 0
     |   |   |   |--Bianca                      | 0.14          | 1              | 1         | 2285
     |   |   |   |--Bryco                       | 0             | 0              | 0         | 0
     |   |   |   |--Lyco                        | 0             | 0              | 0         | 0
     |   |   |   |--Gymco-III                   | 0             | 0              | 0         | 0
     |   |   |   |--Gymco-I                     | 0             | 0              | 0         | 0
     |   |   |   |--Gymco-II                    | 0             | 0              | 0         | 0
     |   |   |   |--Ikeros                      | 0             | 0              | 0         | 0
     |   |   |   |--Ivana                       | 0.18          | 2              | 2         | 3020
     |   |   |   |--Gymco-IV                    | 0             | 0              | 0         | 0
     |   |   |   |--Osser                       | 0             | 0              | 0         | 0
     |   |   |   |--SIRE                        | 9.57          | 5              | 22        | 160206
     |   |   |   |--TAR                         | 0.26          | 5              | 5         | 4355
     |   |   |   |--Tork                        | 0.36          | 1              | 1         | 5947
     |   |   |   '--Ty1-outgroup                | 0             | 0              | 0         | 0
     |   |   '--Ty3_gypsy                       | 0             | 0              | 0         | 0
     |   |       |--non-chromovirus             | 0             | 0              | 0         | 0
     |   |       |   |--non-chromo-outgroup     | 0             | 0              | 0         | 0
```

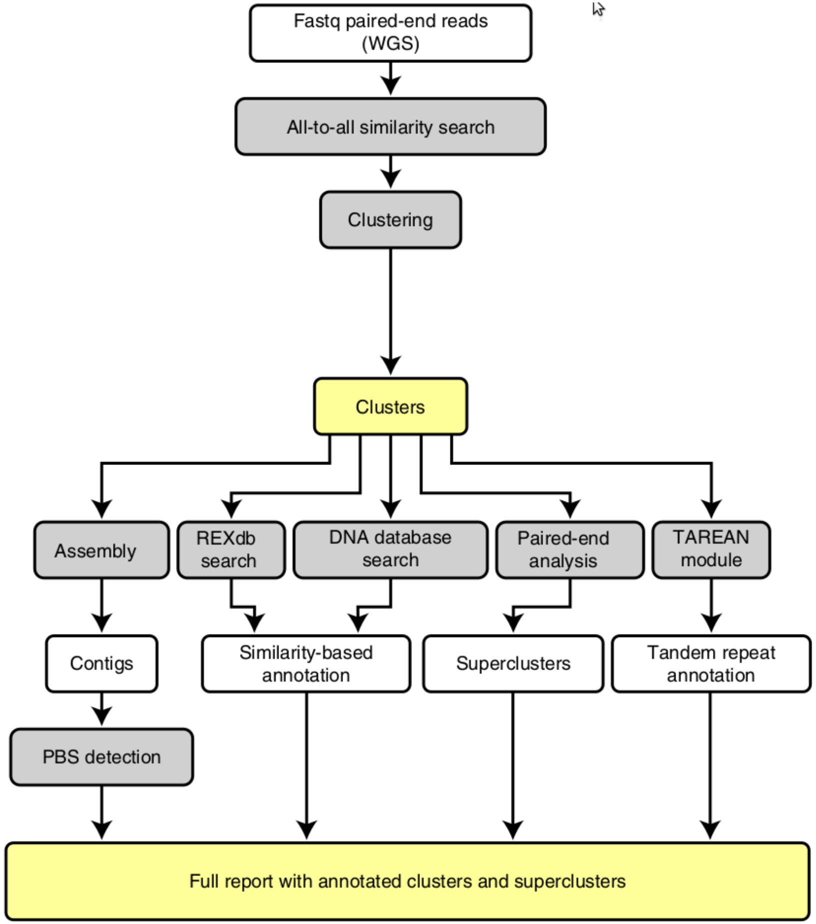
Automatic annotation

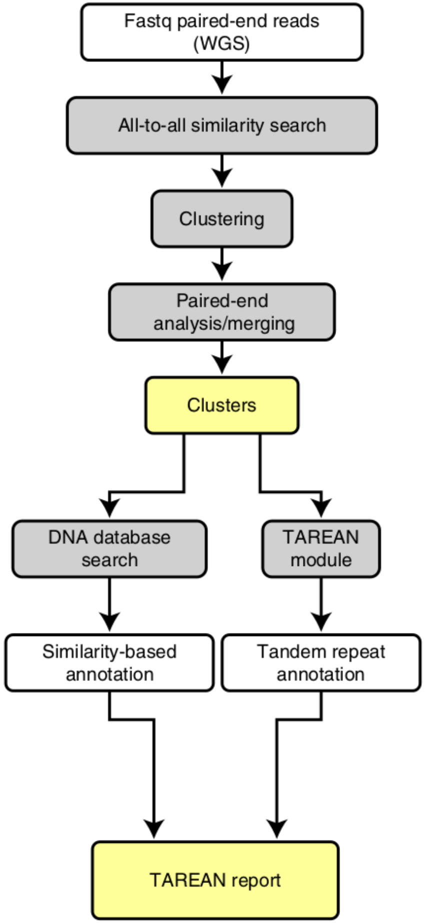


```
                                                | Proportion[%] | Nsuperclusters | Nclusters | Nreads
-----------------------------------------------------------------------------------------------------
Unclassified_repeat (conflicting evidences)     | 4.06          | 2              | 5         | 67995
 |--rDNA                                        | 0             | 0              | 0         | 0
 |   |--45S_rDNA                                | 0.29          | 2              | 4         | 4823
 |   |   |--18S_rDNA                            | 0.04          | 1              | 1         | 653
 |   |   |--25S_rDNA
 |   |   '--5.8S_rDNA
 |   '--5S_rDNA
 |--satellite
 '--mobile_element
     |--Class_I
     |   |--SINE
     |   |--LTR
     |   |   |--Ty1_copia
     |   |   |   |--Ale
     |   |   |   |--Alesia
     |   |   |   |--Angela
     |   |   |   |--Bianca
     |   |   |   |--Bryco
     |   |   |   |--Lyco
     |   |   |   |--Gymco-III
     |   |   |   |--Gymco-I
     |   |   |   |--Gymco-II
     |   |   |   |--Ikeros
     |   |   |   |--Ivana
     |   |   |   |--Gymco-IV
     |   |   |   |--Osser
     |   |   |   |--SIRE
     |   |   |   |--TAR
     |   |   |   |--Tork
     |   |   |   '--Ty1-outgroup
     |   |   '--Ty3_gypsy
     |   |       |--non-chromovi
     |   |       |   |--non-chro
```

RepeaExplorer

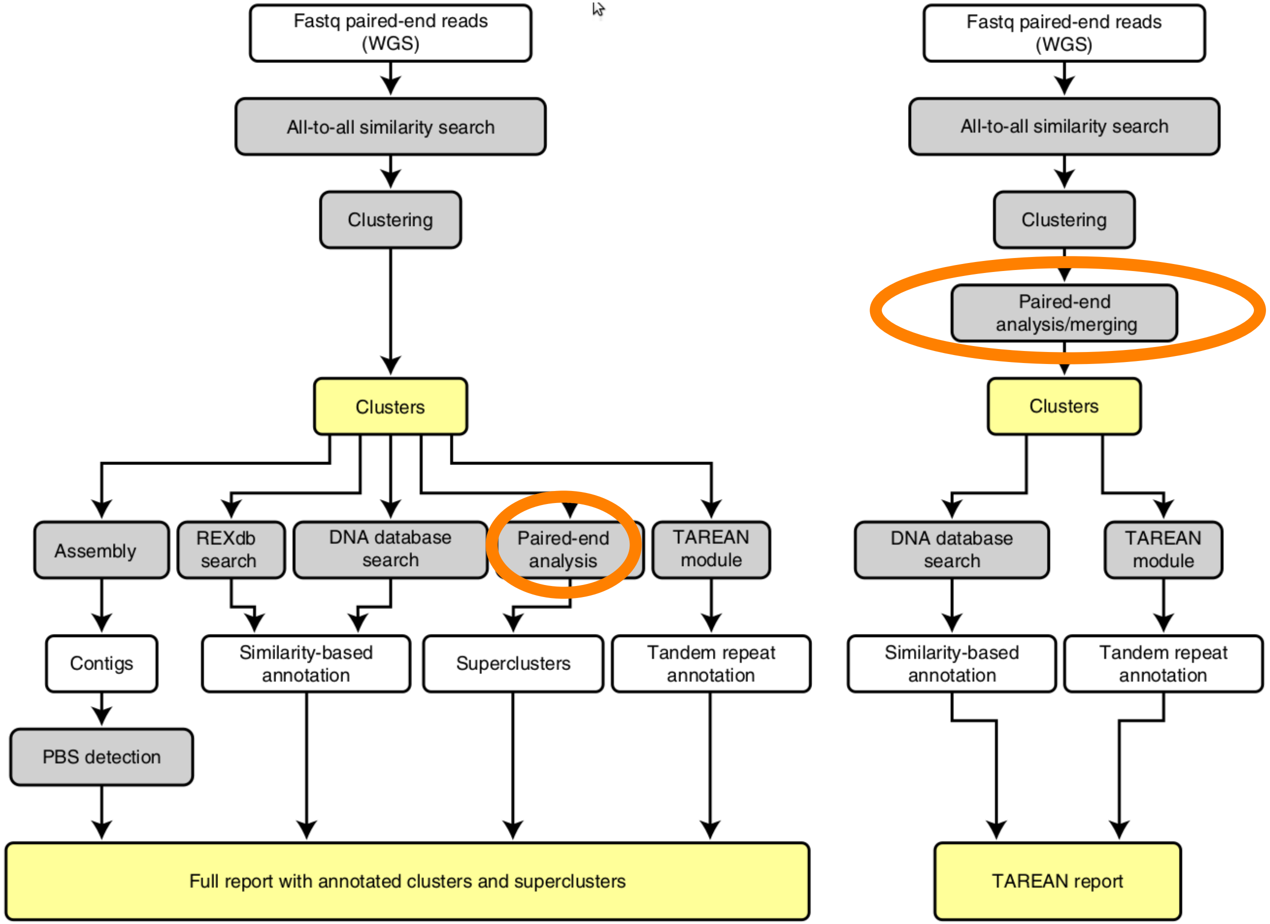
TAREAN





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

RepeaExplorer TAREAN



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
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

RepeatExplorer Related Tools

- DANTE – **D**omain based **AN**notation of **T**ransposable **E**lements
  - assembly annotation using REXdb
  - same TE classification system as RepeatExplorer based on REXdb
- Profrep
  - assembly annotaion based on RE results
- ChIP-Seq Mapper
  - Inentification of repeats associated with CENH3 or with a epiginetic marks

Availability

RepeatExplorer Galaxy Server

https://repeatexplorer-elixir.cerit-sc.cz/

regalaxy@rt.cesnet.cz

Support:

Martina Macháč
Zdeněk Salvet
Miroslav Ruda
Ivana Křenková

SC cerit scientific cloud

cesnet

elixir CZECH REPUBLIC

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

Thank you!

Questions?