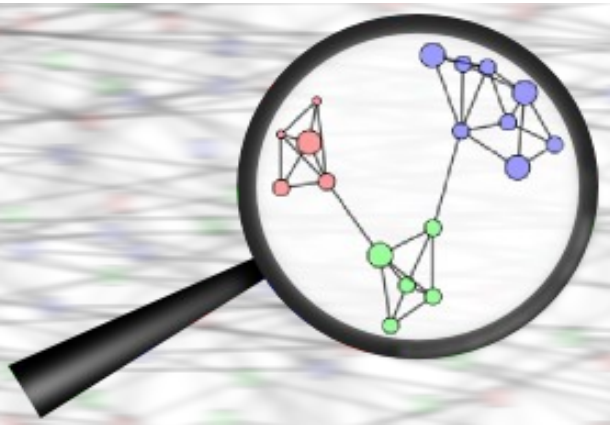
RepeatExplorer 2.0

Discover repeats in your next generation sequencing data



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

elixir CZECH REPUBLIC

BIOLOGY CENTRE CAS

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

Available Tools:

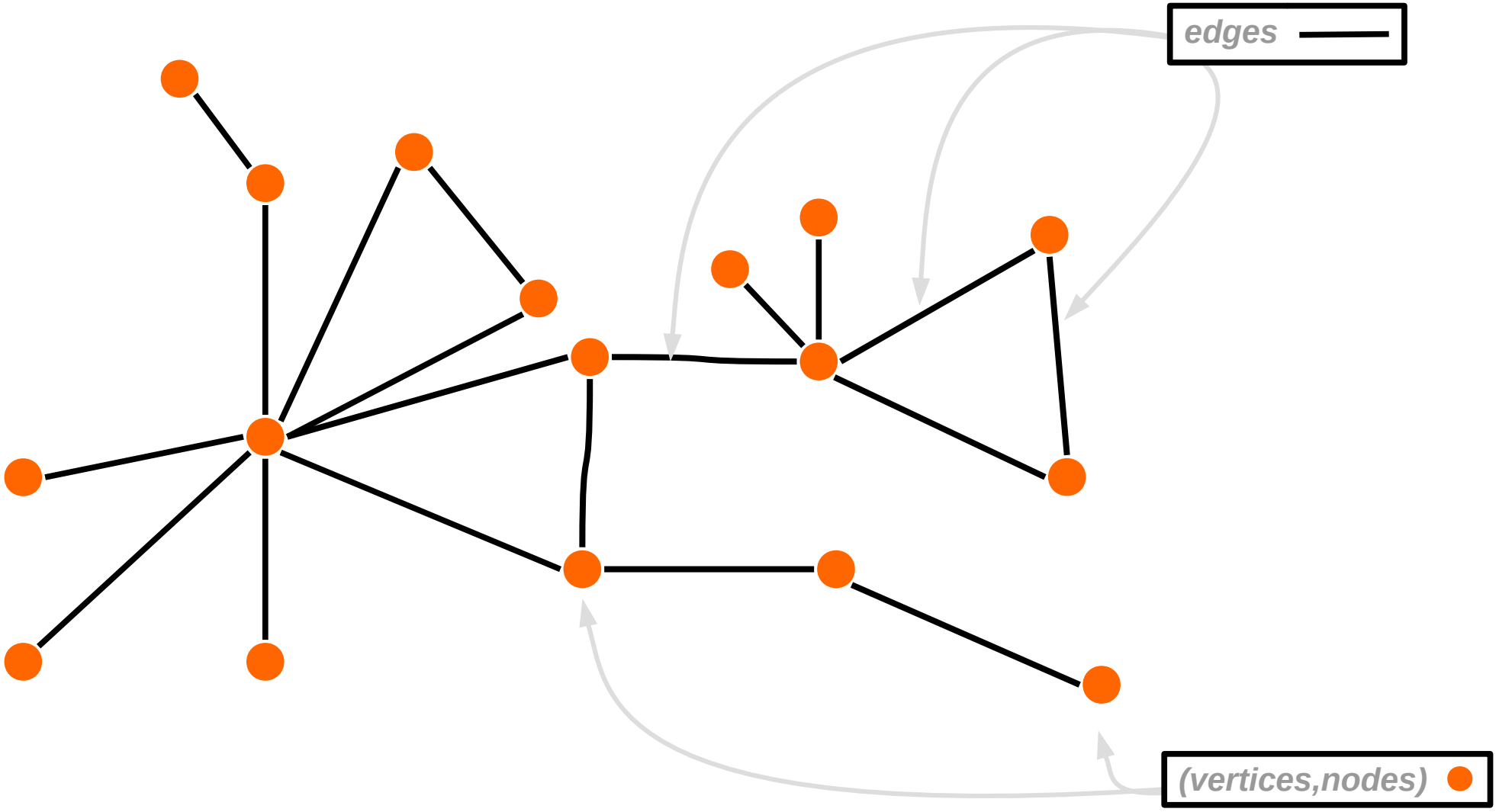
- NGS data preprocessing
- Graph-based clustering
  - Characterization of repeats:
    - Identification, Annotation, Quantification
- Satellite identification
- Chip-Seq analysis
- Domain based ANnotation of Transposable Element – DANTE
- Profrep

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?



Graph Based Representation of Sequence Reads

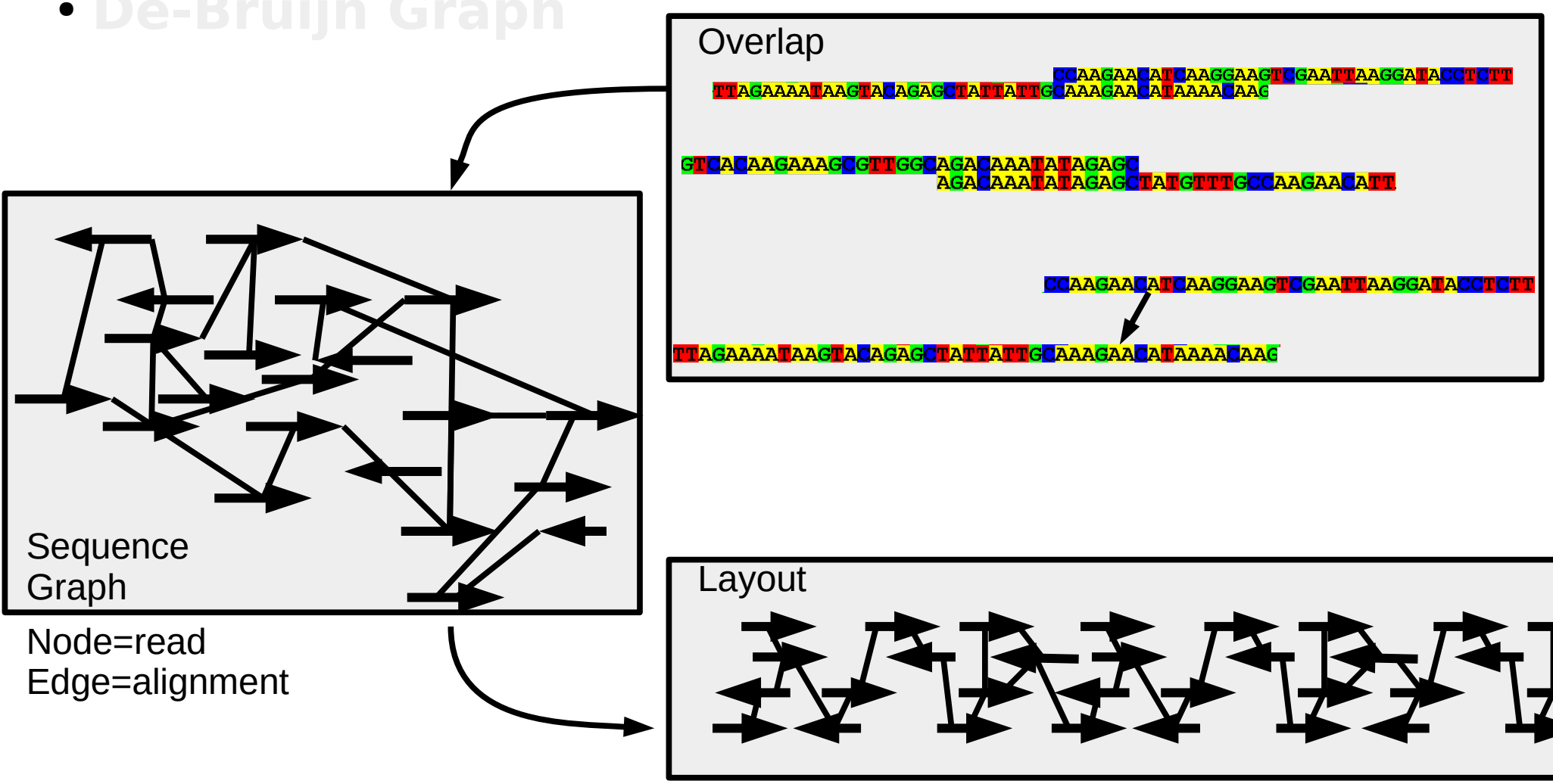
The are two approaches how to use a graph to describe and analyze sequences reads:

- **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

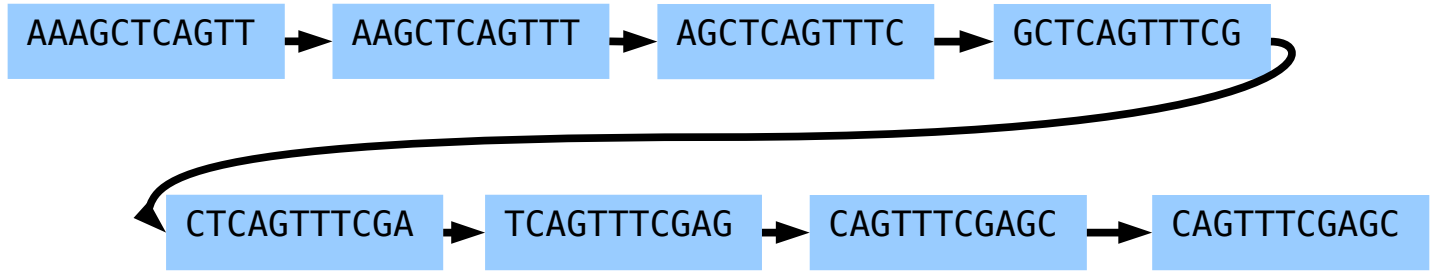
The are two approaches how to use a graph to describe and analyze sequences reads:

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

Sequence read:

```
AAAGCTCAGTTTCGAGCCAGAGACCAGAAAGTGTGGGAGCTTACAGCGCAACTTCAGCAAGAGCGGAG

AAAGCTCAGTT
 AAGCTCAGTTT
  AGCTCAGTTTC
   GCTCAGTTTCG
    CTCAGTTTCGA
     TCAGTTTCGAG
      CAGTTTCGAGC
       AGTTTCGAGCC
         .......
```



Graph Based Representation of Sequence Reads

Why to use graph representations :

The are number of available algorithms for graph analysis

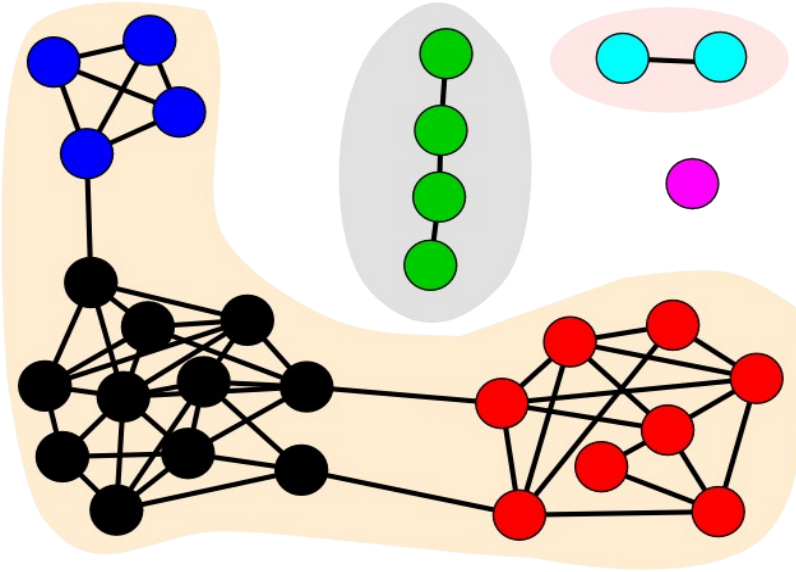
- Robust partitioning/classification of reads based on mutual similarities
- Informative graphical representation (layouts)
- Path in graph can be converted to contigs

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

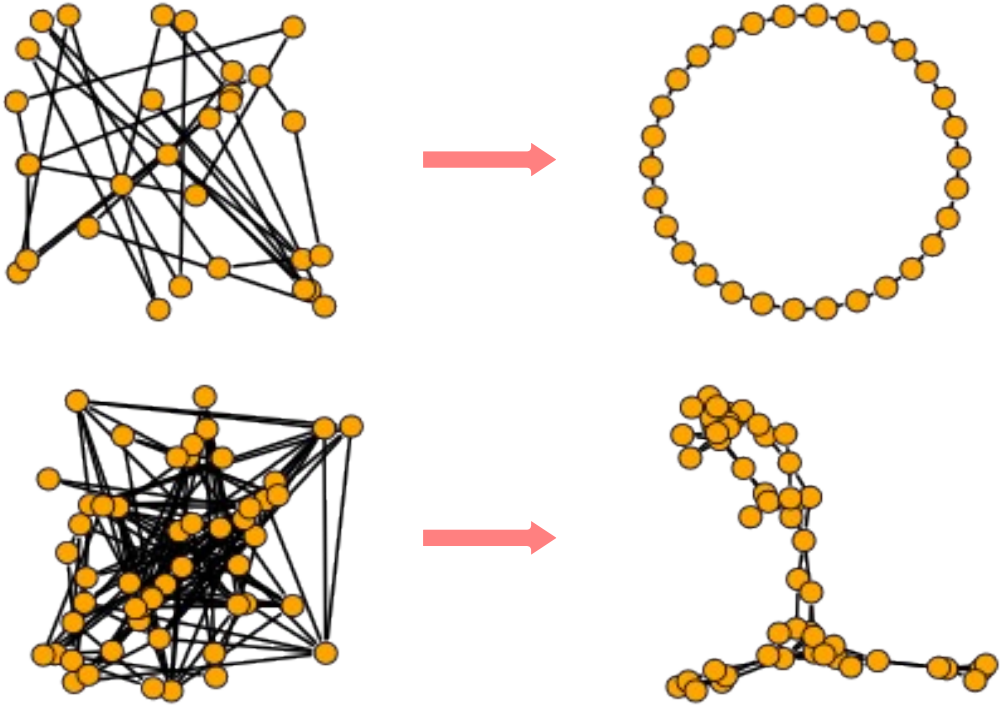


Graph Based Representation of Sequence Reads

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

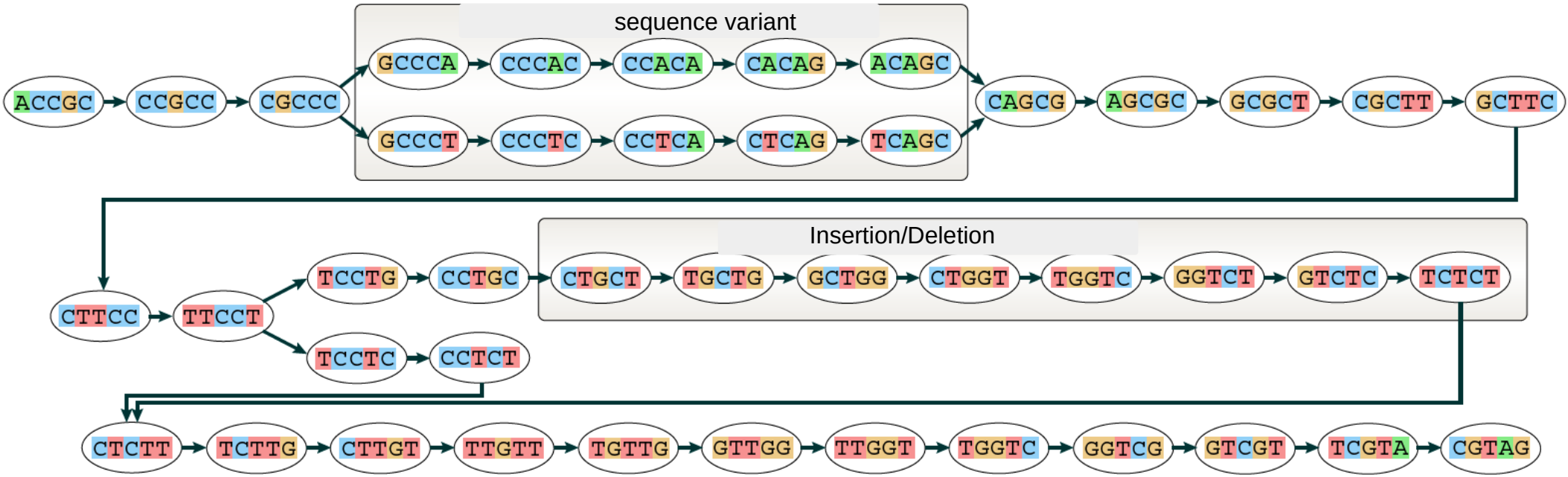


Graph Based Representation of Sequence Reads

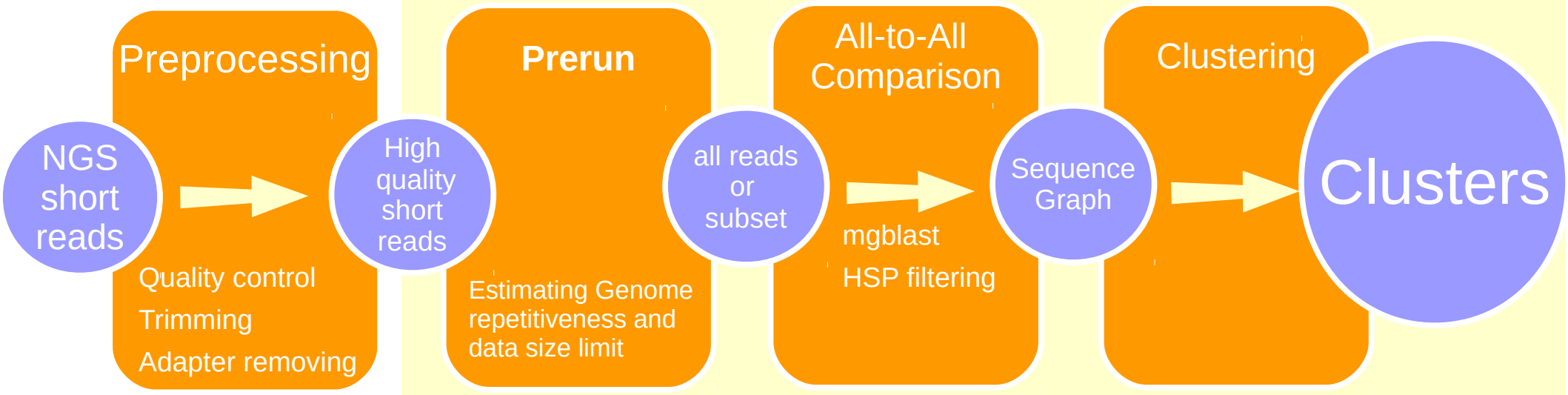
Why to use graph representations :

The are number of available algorithms for graph analysis

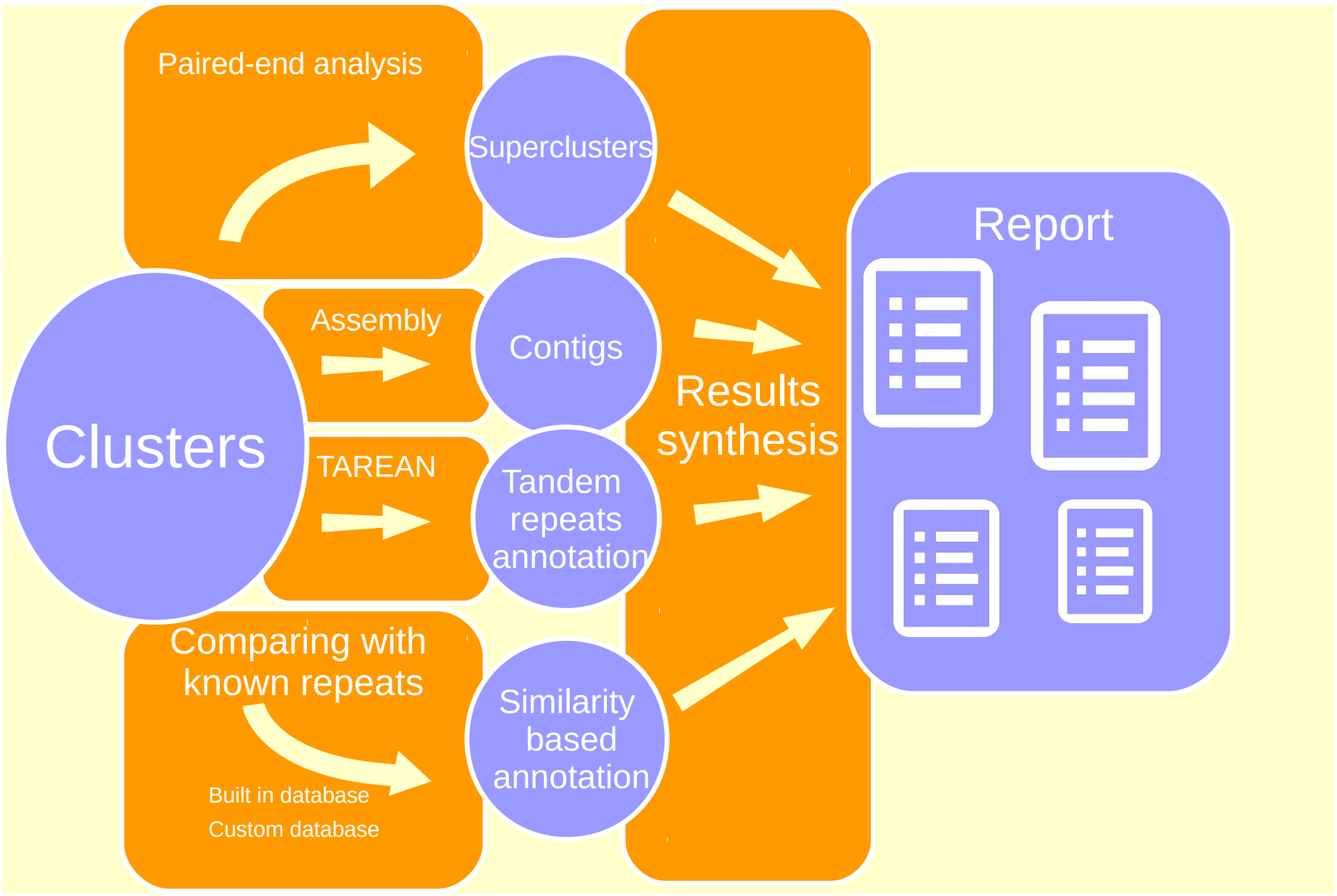
- Robust partitioning/classification of reads based on mutual similarities
- Informative graphical representation (layouts)
- Path in graph can be converted to contigs



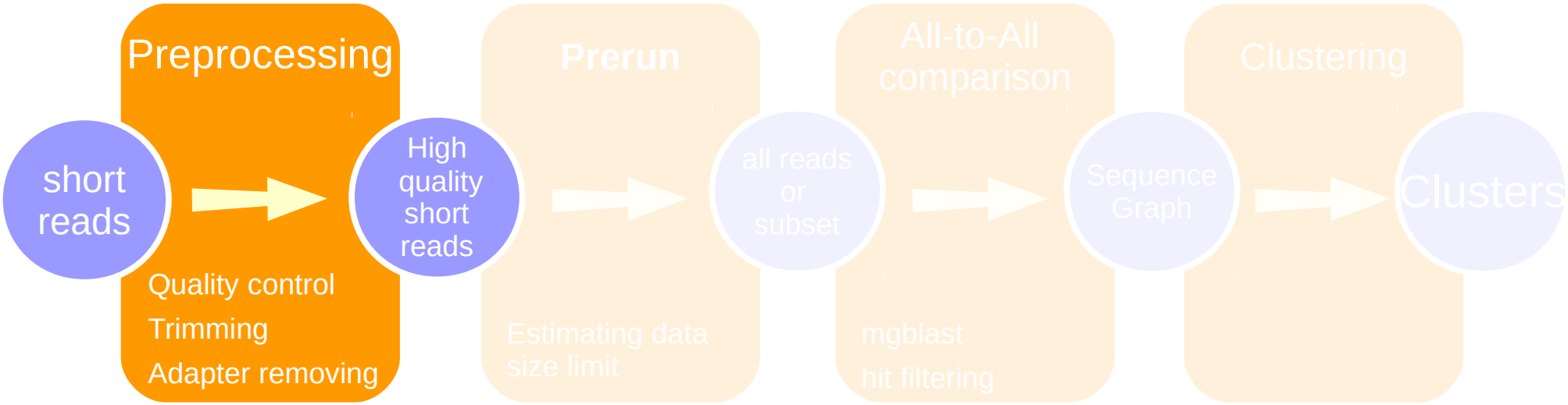
RepeatExplorer workflow



RepeatExplorer workflow



Preprocessing



Preprocessing

RepeatExplorer operates under GIGO principle:

Garbage In

RepeatExplorer

Garbage Out

Preprocessing

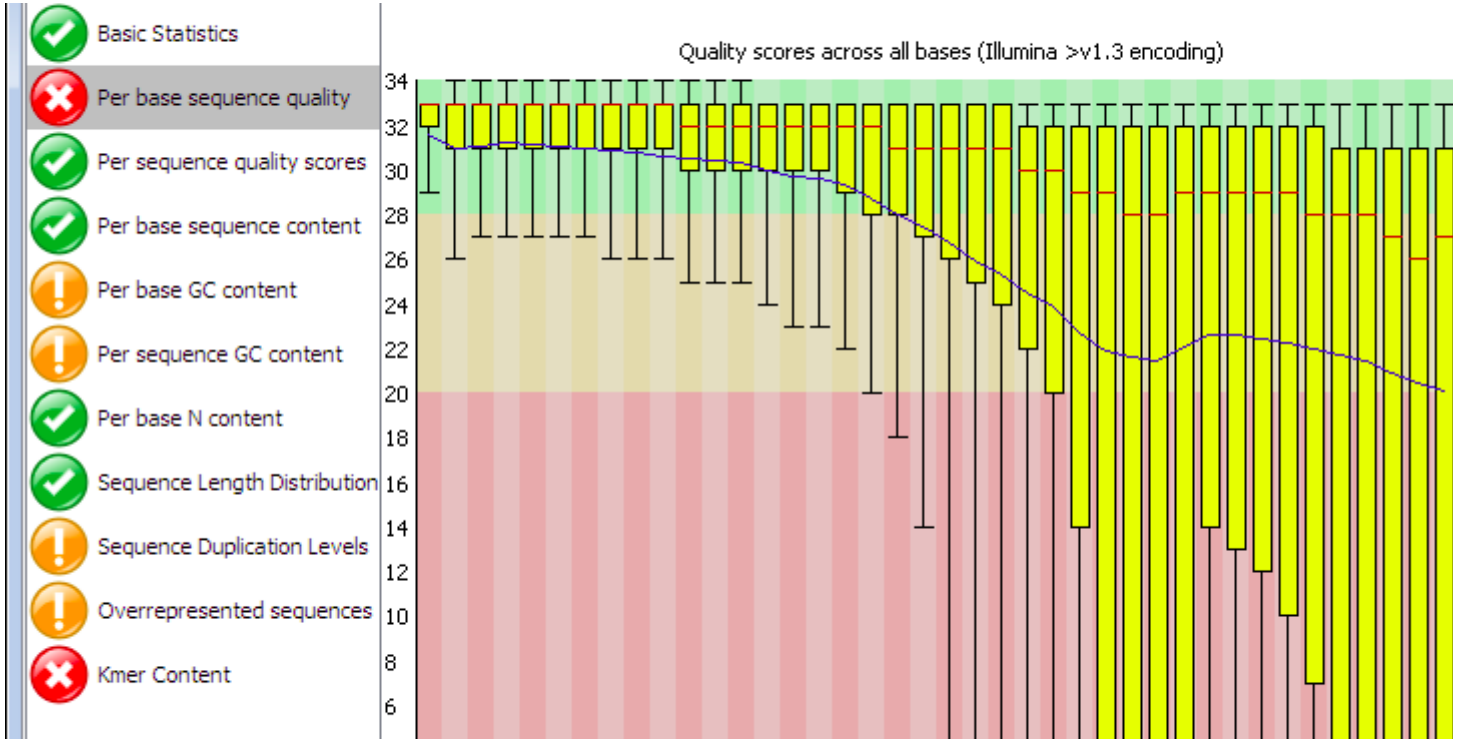
- **Quality control**
- **Trimming, filtering, adapter removing**
- **Convert fastq to fasta**
- **Interlacing, sampling**
- **Modification of sequence names**

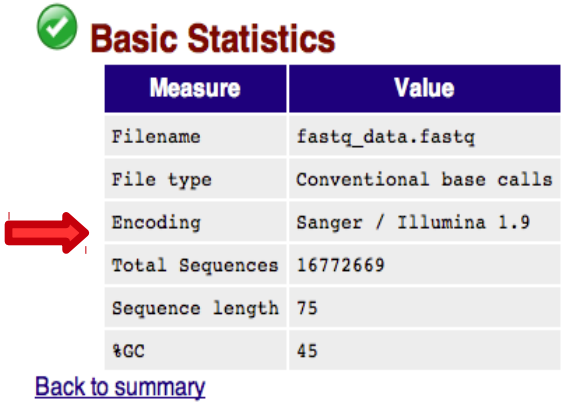
Preprocessing

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

FastQC program

- Galaxy
- GUI based
- Command line





Basic Statistics

Measure	Value
Filename	fastq_data.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	16772669
Sequence length	75
%GC	45

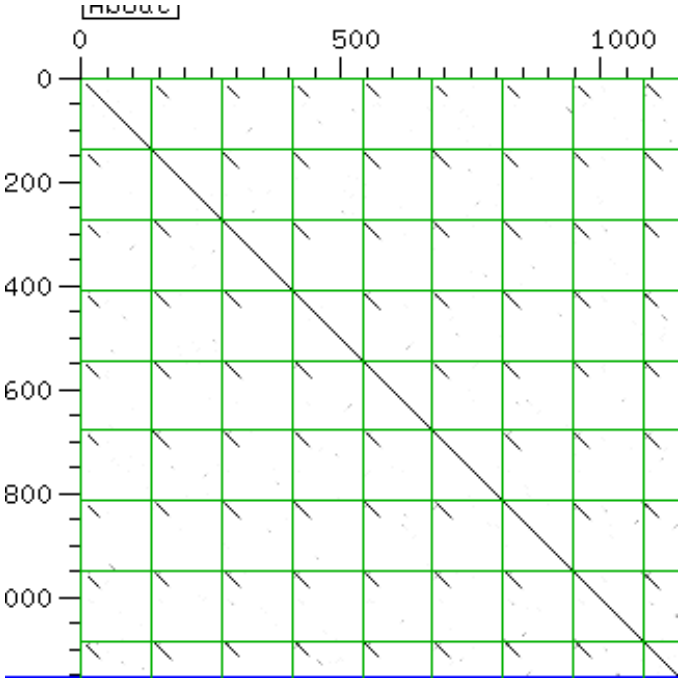
Back to summary

Preprocessing

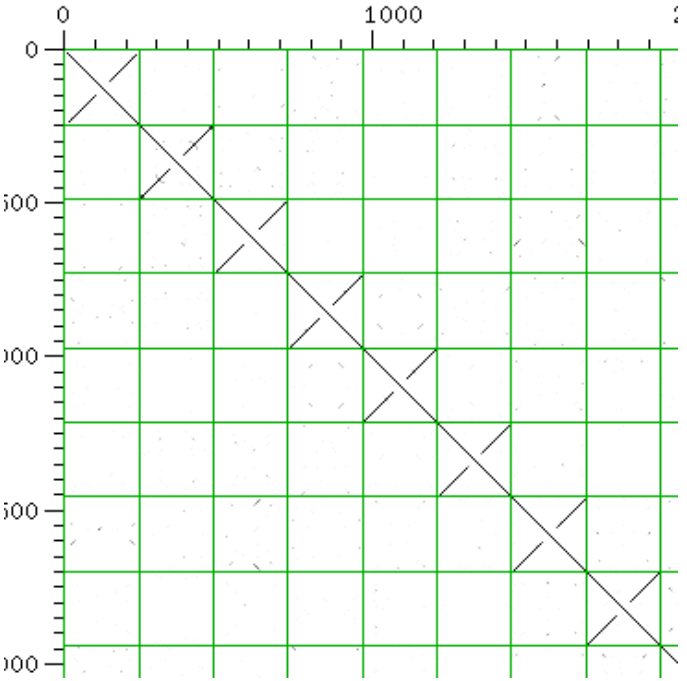
- **Quality control**
- **Trimming, filtering, adapter removing**
- **Convert fastq to fasta**
- **Interlacing, sampling**
- **Modification of sequence names**

Dotter - graphical dotplot program for detailed comparison of two sequences

Simple addapter detection



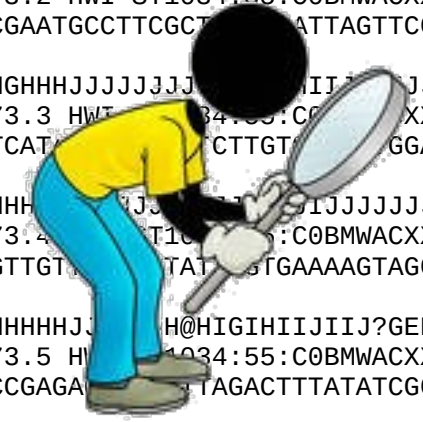
Concatenated reads!



Preprocessing

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

Visual inspection!



```
@DRR013373.1 HWI-ST1034:55:C0BMWACXX:1:1101:1149:2090/1
ATGAGATAACGCCTAATGAATCCGACGAAACTGCGGATCTCCGGCAAAGTCTTCGAATCTCCGGTGAAGAAATCAGATCCGGTGAGCGACGGAGTCGTCGA
+
CCCFFFFFHHHHHJJJJJJJJJJJIJJIJJIIIJIIFHJJJJJJHHFFFEEEEEDDDDDDDDDD;BBCDDDDDDDACCDDDD>B?B9>3>>@#########
@DRR013373.2 HWI-ST1034:55:C0BMWACXX:1:1101:1205:2126/1
AACTTCATCGAATGCCTTCGC[illegible]ATTAGTTCGGAGAATCCAATCACCGATGCCATCTTTACAGCCGATAACACCACCACCTTCTTGTCTTCTTACGT
+
CCCFFFFFHGHHHJJJJJJJJ[illegible]JJJJJIIJJJIGIJJJJJIGHHFFFFFFECEEEDDDDDDDDCDBBDDDDBDCCDDDDDDDDDCCCDB
@DRR013373.3 HWI-ST1034:55:C0BMWACXX:1:1101:1300:2065/1
ATACTCAGTCAT[illegible]CTTGT[illegible]GGACTCCAAAAACCCTACACAAGCTTATTATGGCTTCCCCAAGTTTTCTAAGTGAGGACGAAGACCCT
+
CCCFFFFFHHH[illegible]IJJJJJJJJJJJJJJJJJJIJJJJIJIHHGIJJGIIJIJIJJCHHGGF@DFFFDEDDECECCAABDDDDDDDDBC
@DRR013373.4 [illegible]:C0BMWACXX:1:1101:1479:2108/1
TGCAACTCGTTGT[illegible]GTGAAAAGTAGGAGACTTGAAGATGCACGGAAGGTGTATGACGAAATGTGTGAGAGAGGTGGCTGTGTTGATAATTA
+
@C@FFFFFHHHHHJ[illegible]H@HIGIHIIJIIJ?GEHEHGHJAHIJIIJJJJJJHDHI@G@=FGGGGDHGBE?B;;;ACC?AB5;?((85>5:<353:3@:
@DRR013373.5 HW[illegible]1034:55:C0BMWACXX:1:1101:1352:2147/1
ATCCGTTGCCGAGA[illegible]TAGACTTTATATCGCAGCACAGCACCCGCGCACACACCGTCTCCGGGGAGGCGAATGCTAGCCGCTCGTTTGCTTCTTCC
+
CCCFFFFFHHGHFGHGHHFHGIGCGIJJJEHCHIJJJJJFIGGIGHGIGIFFCC>B=BBB<@BCDDDDD9@B?B<>9AC<ACDCBD###############
@DRR013373.6 HWI-ST1034:55:C0BMWACXX:1:1101:1374:2154/1
ATGCATTTAAACACTTAGCTAGCTAAAGAAACAGAGGATTCCACTAACAACCTATCCCATCCTTAAGAACTTTGCAGAACTACTCAAGAACACTGAAGAAC
+
@@CFADDFHHFFHIEHJJJJ<EEHHIIHEGGGEHGIFBHGHG@GIJHIJJIJJJJJIAHGHGHJJGIGIHDGIH=?AHHHHFFFFFCCEC@CCDDDDDDCD
@DRR013373.7 HWI-ST1034:55:C0BMWACXX:1:1101:1328:2186/1
GTATATGTGGAAGGTGTGTCACGTGTAAAAAGGCTAAGTCCAAAGTGCAACCGAACGGTTTGTATACTCCTTTACCCATTCCTTCTCATCCTTGGACTGAC
+
BBBFFFFFHHHHHIFHHIHHHIHEHFGIJIJIIDHIIJEHEIJJIDGGIGFFGGI4@B;@B=;);;;BC>>@C:6.;(;@@CAC:>5(5:>>A5:,>A?>A
@DRR013373.8 HWI-ST1034:55:C0BMWACXX:1:1101:1430:2218/1
TGGCACCTCGATGTCGGCTCTTCGCCACCTGGGGCTGTAGTATGTTCCAAGGGTTGGGCTGTTCGCCCATTAAAGCGGTACGTGGGCTGGGTTCAGAACGC
+
CCCFFFFFHHHHHHIIJIJIHGGGGIJIJIIFJJIGI2BGBFGGFGCGDCFFI.=@HH@;B;>3?6;?6?A@C@CA?;5059<3(0(+5@B##########
@DRR013373.9 HWI-ST1034:55:C0BMWACXX:1:1101:1469:2227/1
AATCAATATAGATTGATCCGAAATATGATTCAAATCCAATTCCAATATAGTCCCTATGGGTACATAAGAAATGTATTGAATCGATTCTTTTTAATGAAGAG
```

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

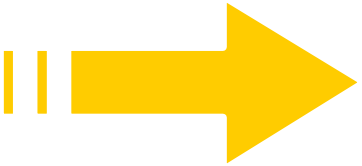
Preprocessing

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

Tool: **affixer**

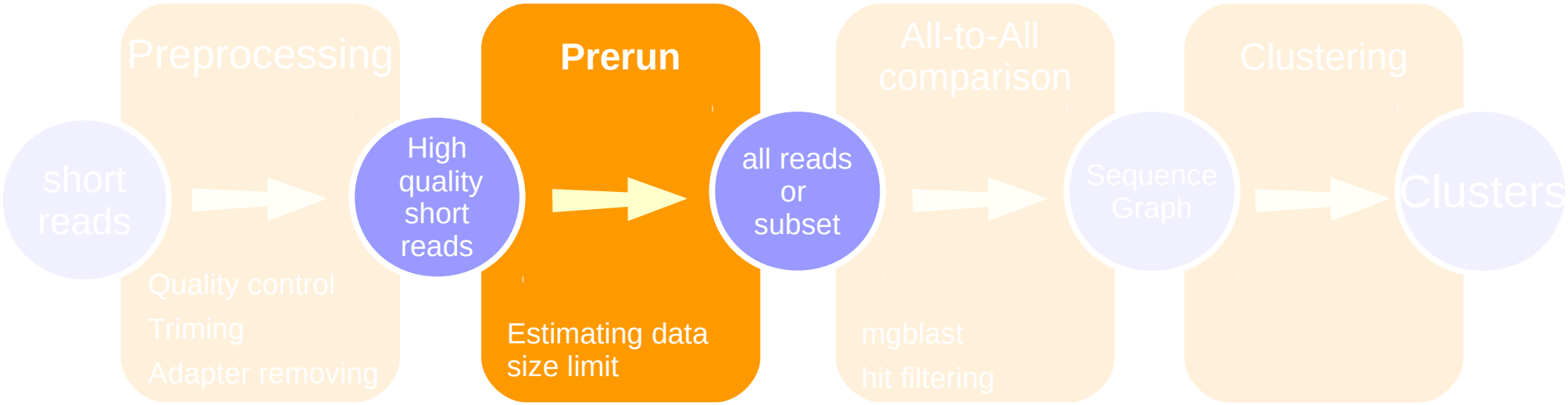
```
Genome AB
>FUXEH4V01EAG68....
acgacagctgactaatgc
>FUXEH4V01BKPDK
cttcgaggctacacgagct
>FUXEH4V01AJAJV
actatcgacactgccggcgcg
...
```

```
Genome XY
>F0X5OLU02GZ8YF....
gccccgtcgccgtccgtgtcg
>F0X5OLU02I1AMY
tgtgtgcccgtctgcgcgccccc
>F0X5OLU02HYN8U
atatgctatgcgcgc
...
```



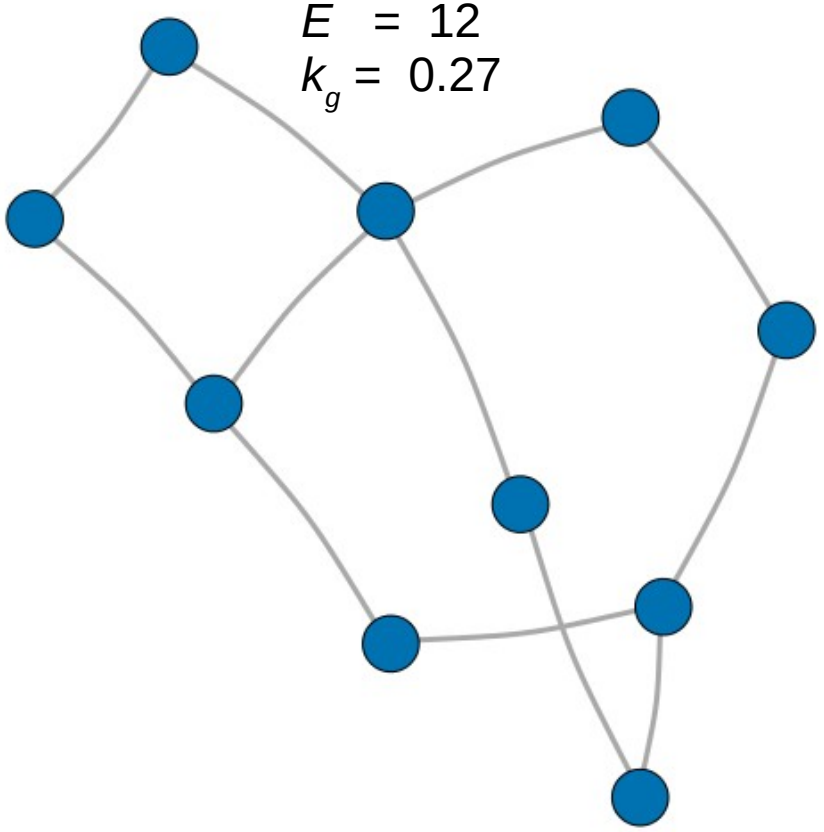
```
comparative analysis:
>AB1
acgacagctgactaatgc
>AB2
cttcgaggctacacgagct
>AB3
Actatcgacactgccggcgcg
...
>XY1
gccccgtcgccgtccgtgtcg
>XY2
tgtgtgcccgtctgcgcgccccc
>XY3
atatgctatgcgcgc
```

Prerun

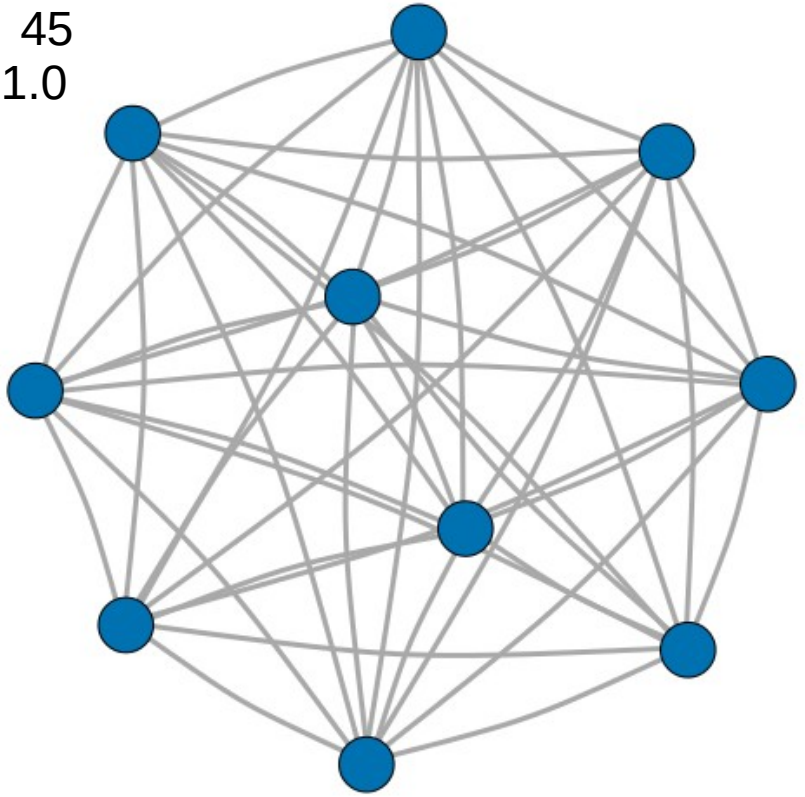

short reads
Preprocessing
Quality control
Triming
Adapter removing
High quality short reads
Prerun
Estimating data size limit
all reads or subset
All-to-All comparison
mgblast
hit filtering
Sequence Graph
Clustering
Clusters

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



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

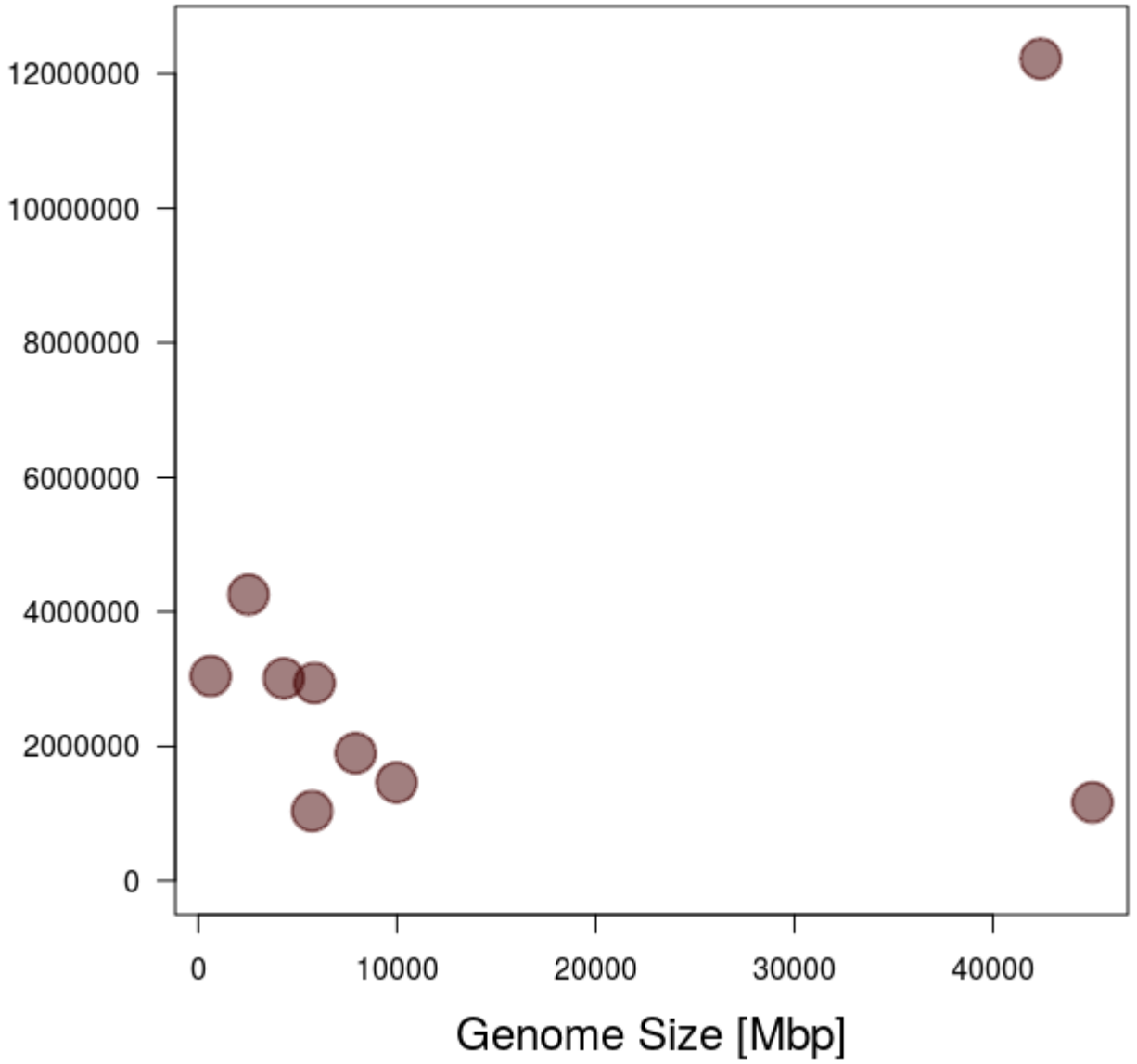
k_g genome specific coefficient - **graph density** depends on repetitive content and genome size

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

k $_g$ is used to estimate maximum number of 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

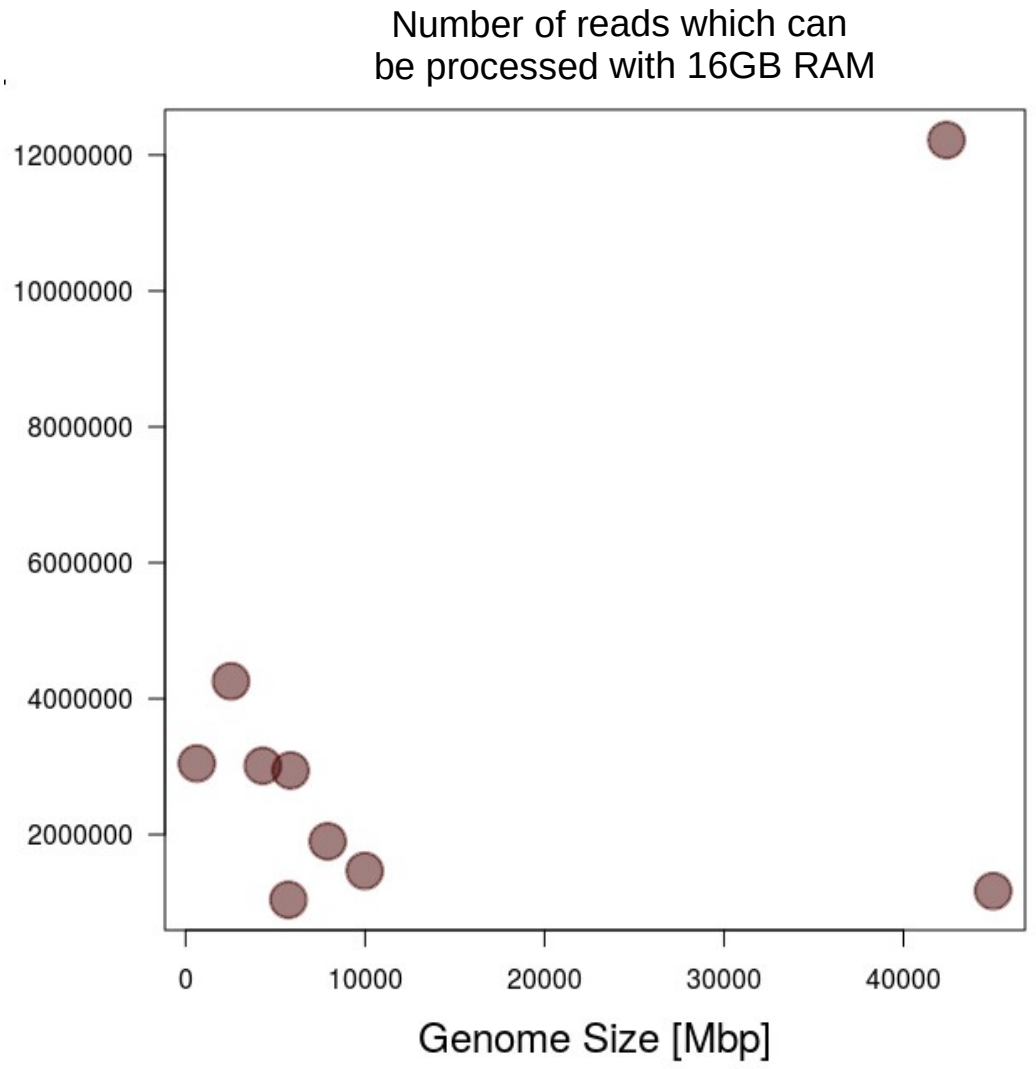
$$k_g = \frac{2E}{N(N-1)}$$

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

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

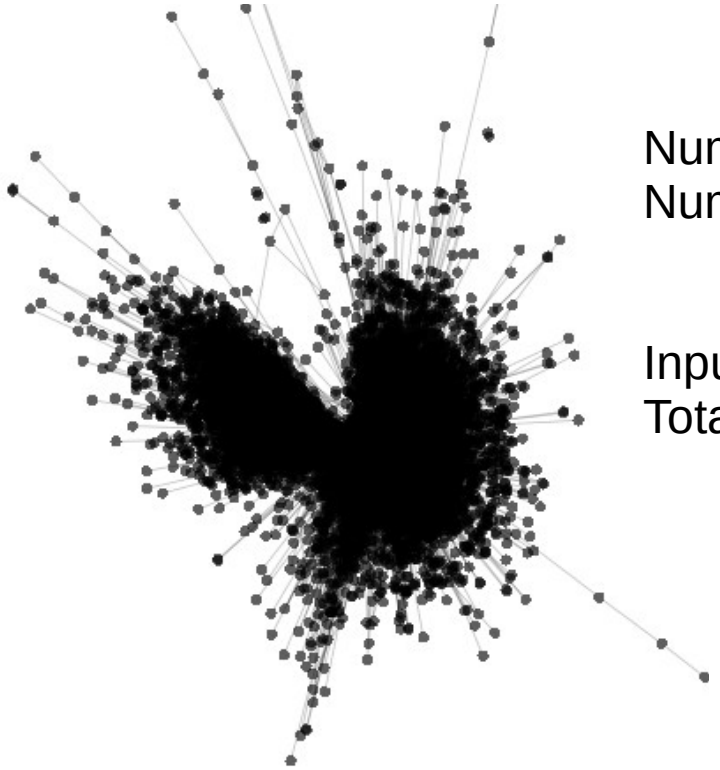
Density corresponds to probability that two randomly taken sequences from genome will be similar

k_g is used to estimate maximum number of reads providing that **we can process ~ *340 ·10⁶* 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	Density
Number of similarity hits (Edges)	542,348,907	0.54

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 approx 2% of genome

Prerun – optional filtering of abundant satellite sequences

Example of dense satellite cluster:

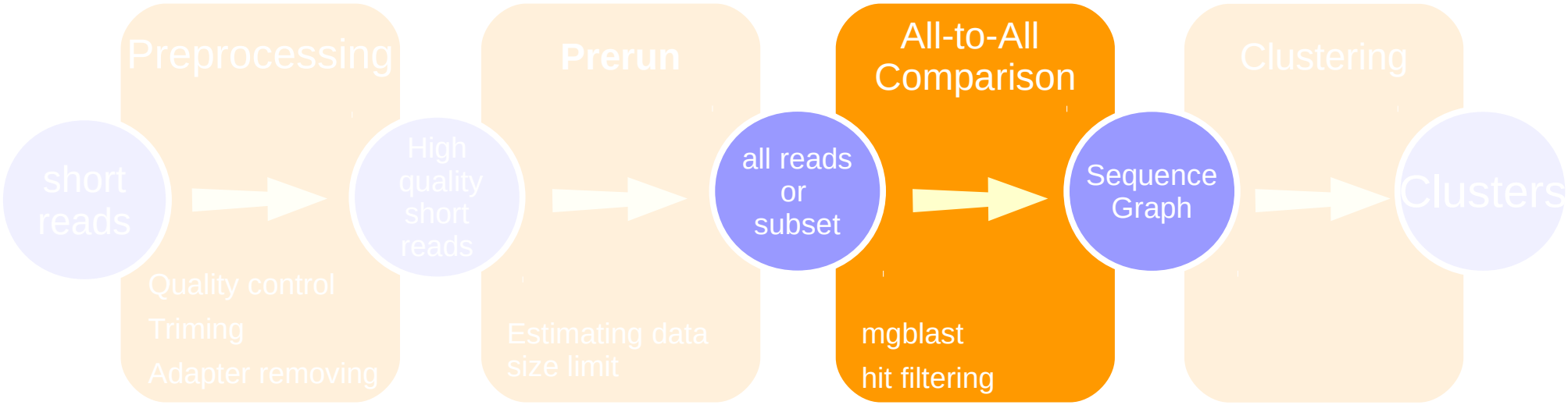
Such clusters can be filtered out from the clustering

Filtering critera:

- 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



All-to-All Comparison

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 then 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 then 100 nt

Presence of unfiltered **adapter** sequence
Does not pass similarity threshold, but

All-to-all comparison becomes extremely slow!

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 then 100 nt

Low complexity repeat – **DustMasker**

Simple repeats are underestimated or not detected at all:
e.g.

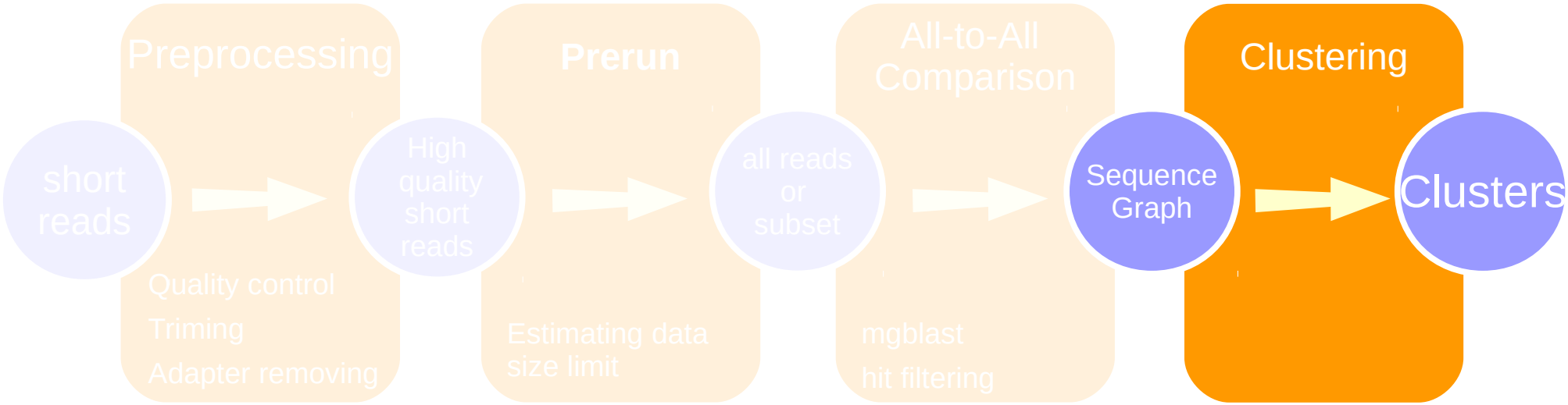
- Telomeric motifs
- microsatellites

…

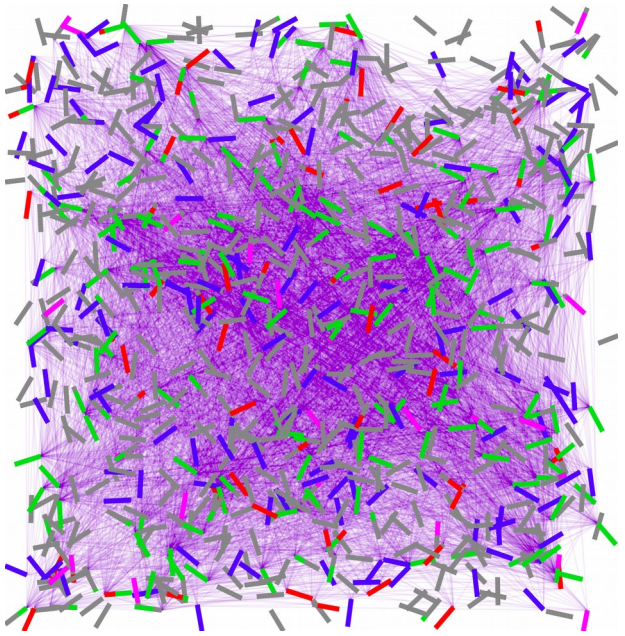
By default, DustMasker is on but it can be disabled to increase sensitivity of simple repeats detection.

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

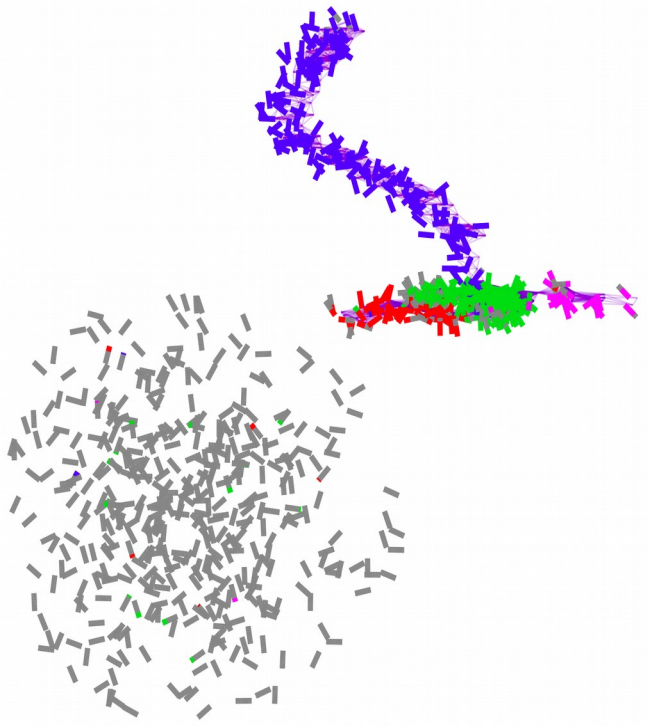
Clustering


short reads
Preprocessing
Quality control
Triming
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

Clustering

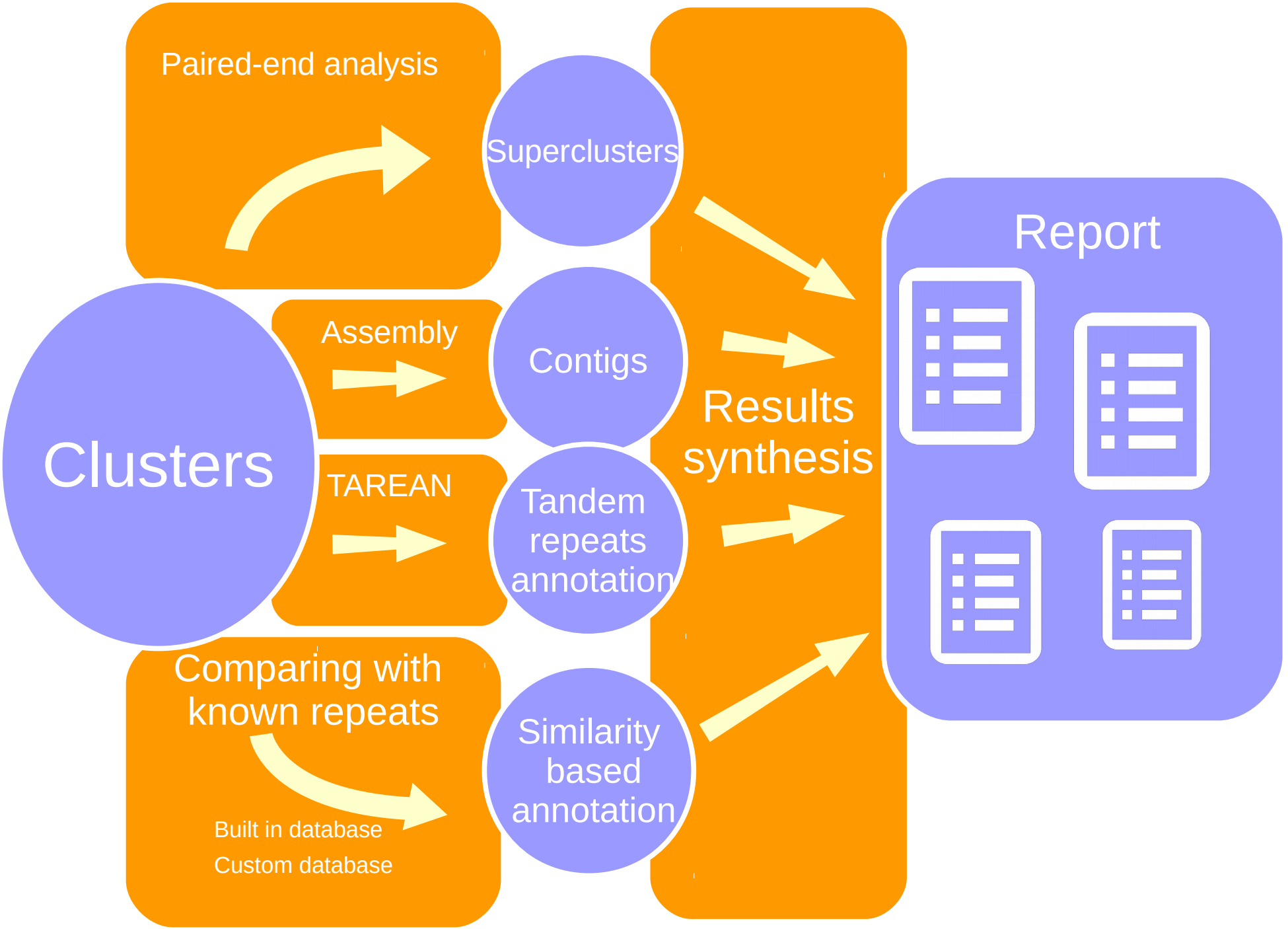


Clustering

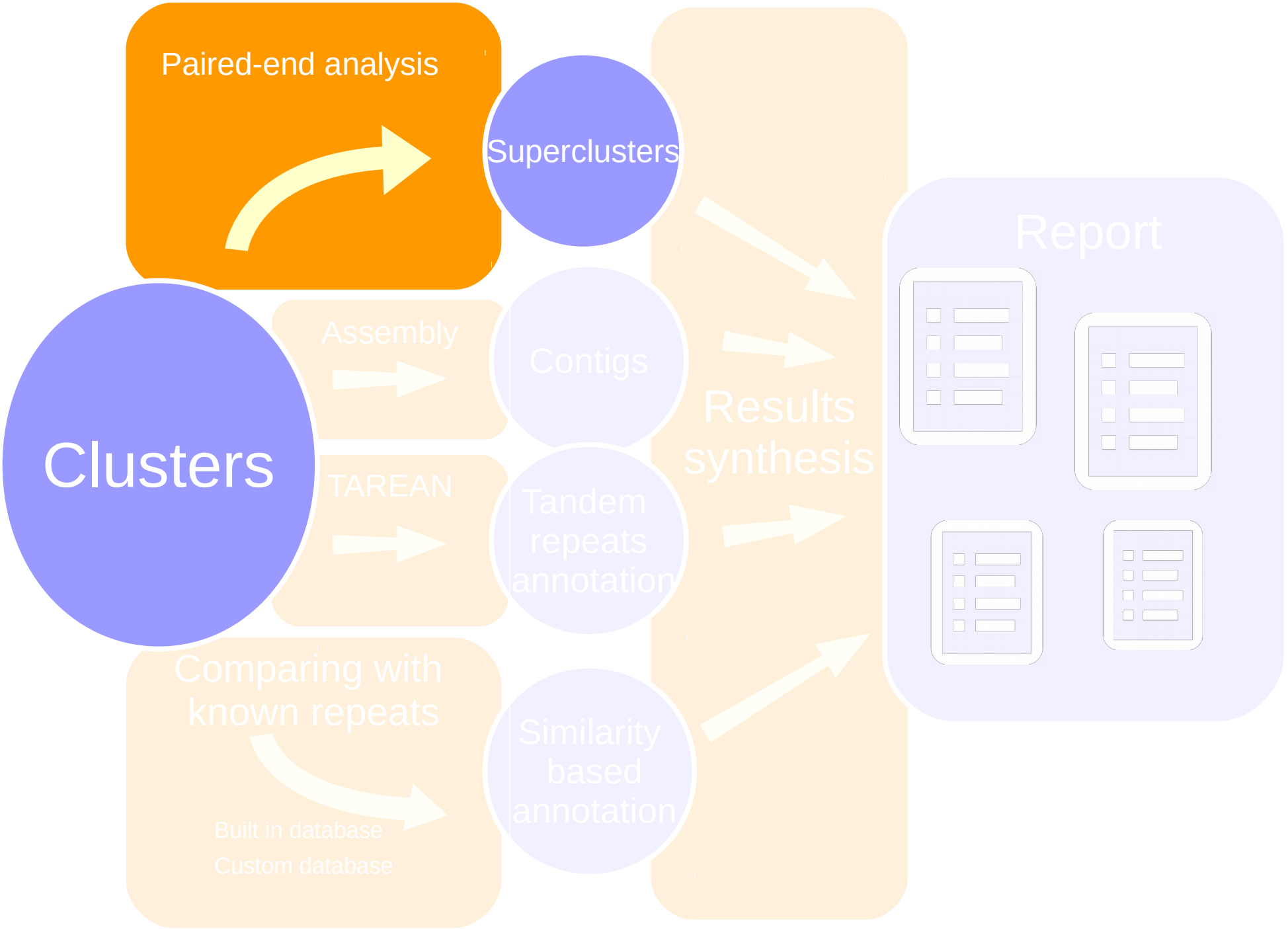


- Graph is divided into subgraphs (clusters/communities)
- Quality of division is measured using **modularity**
- Modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random
- Clusters have dense connections between the nodes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

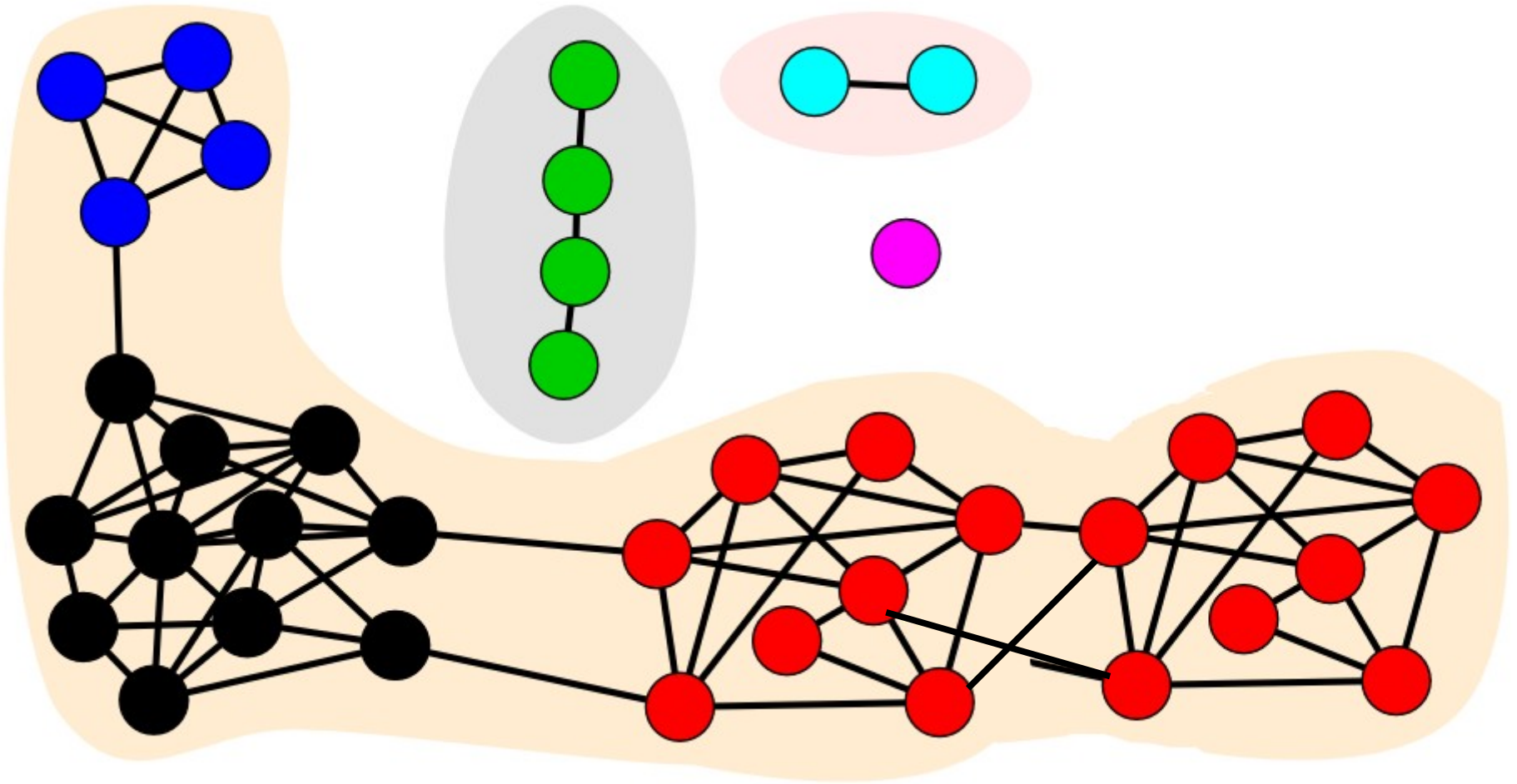
Cluster centered analysis



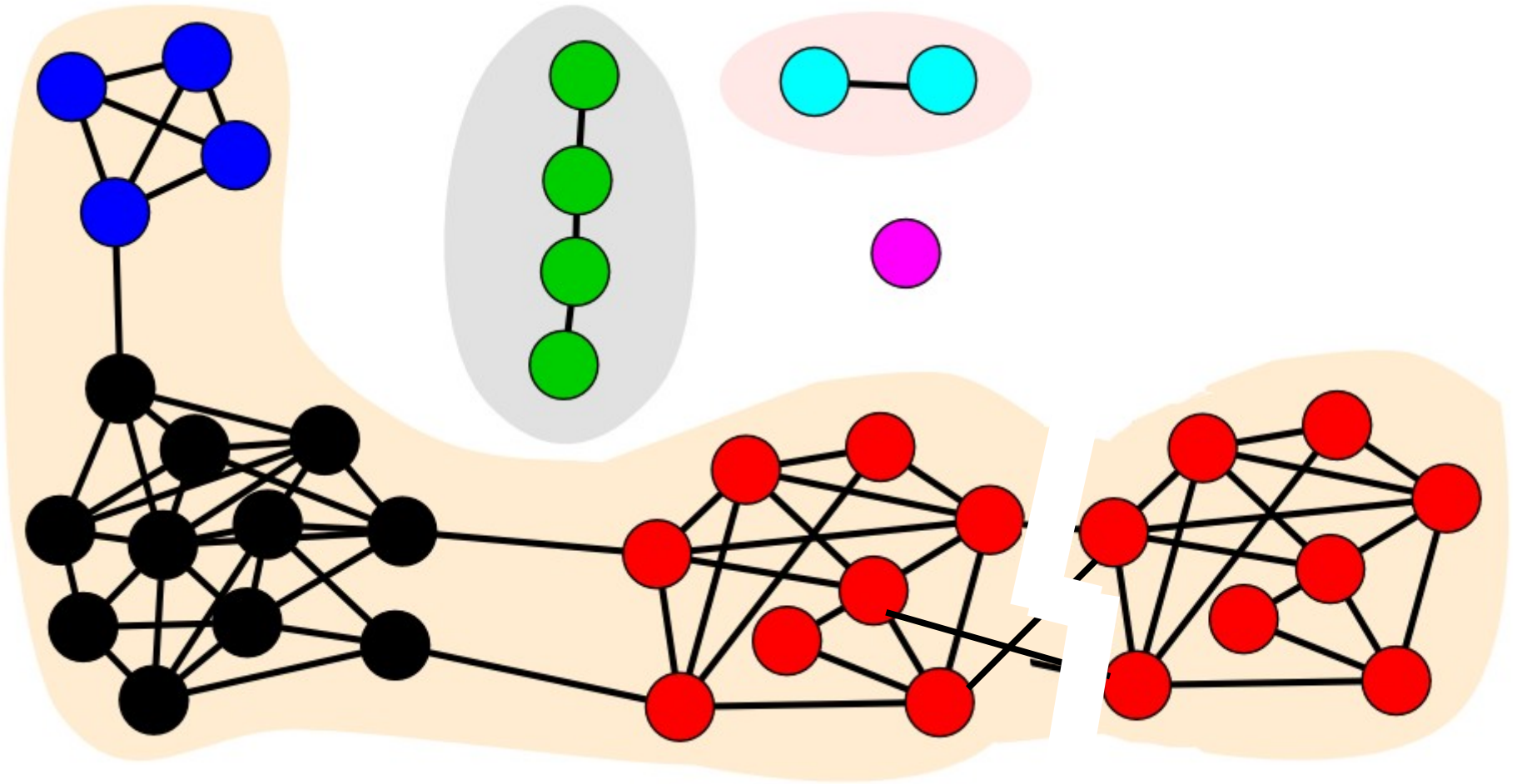
Cluster centered analysis



Supercluster Identification



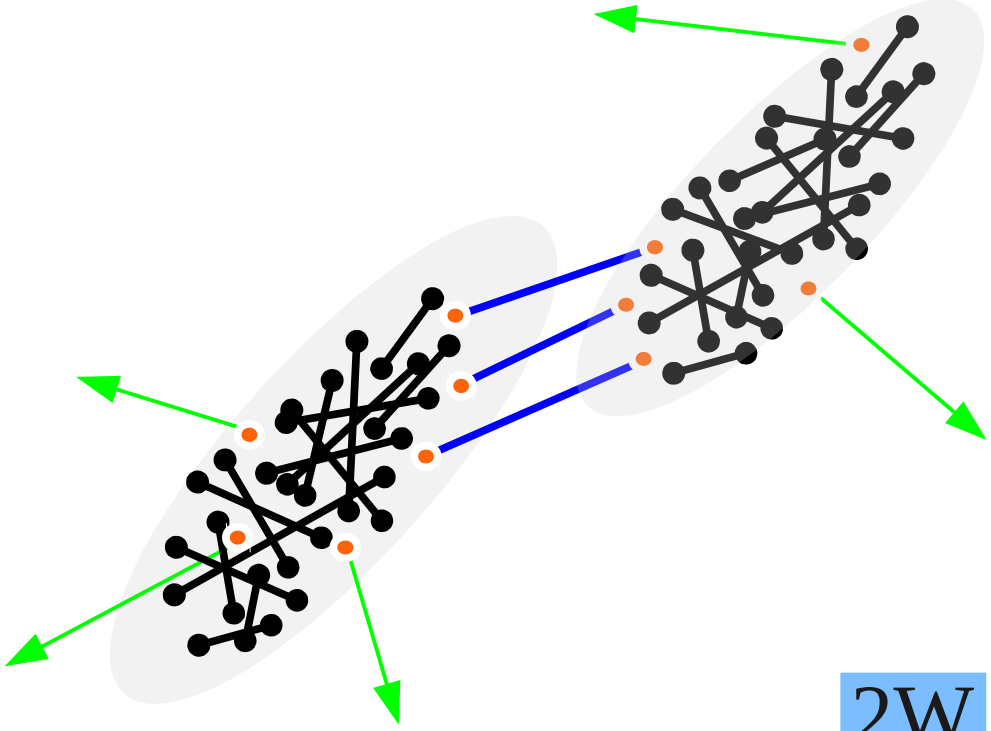
Supercluster Identification



Reads which originate from single repeat are frequently split into multiple cluster during clustering phase – we need to identify such clusters

Supercluster Identification

Identification of related clusters from presence of paired reads



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

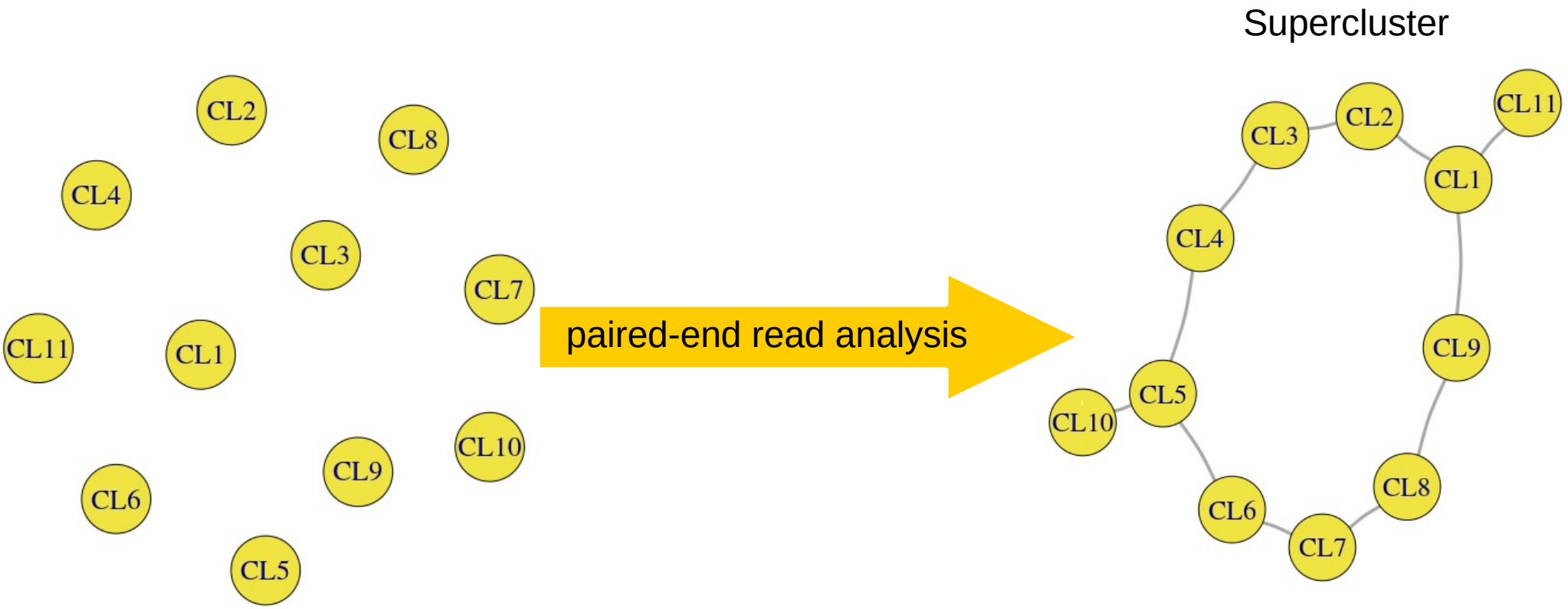
W number of reads pairs shared between clusters x and y
n_x and n_y is number of reads in cluster *x* and cluster *y* with absent read mate within the same cluster respectively

Suitable $k_{x,y}$ cutoff 0.05 – 0.2

full connection: $k_{x,y} = 1$

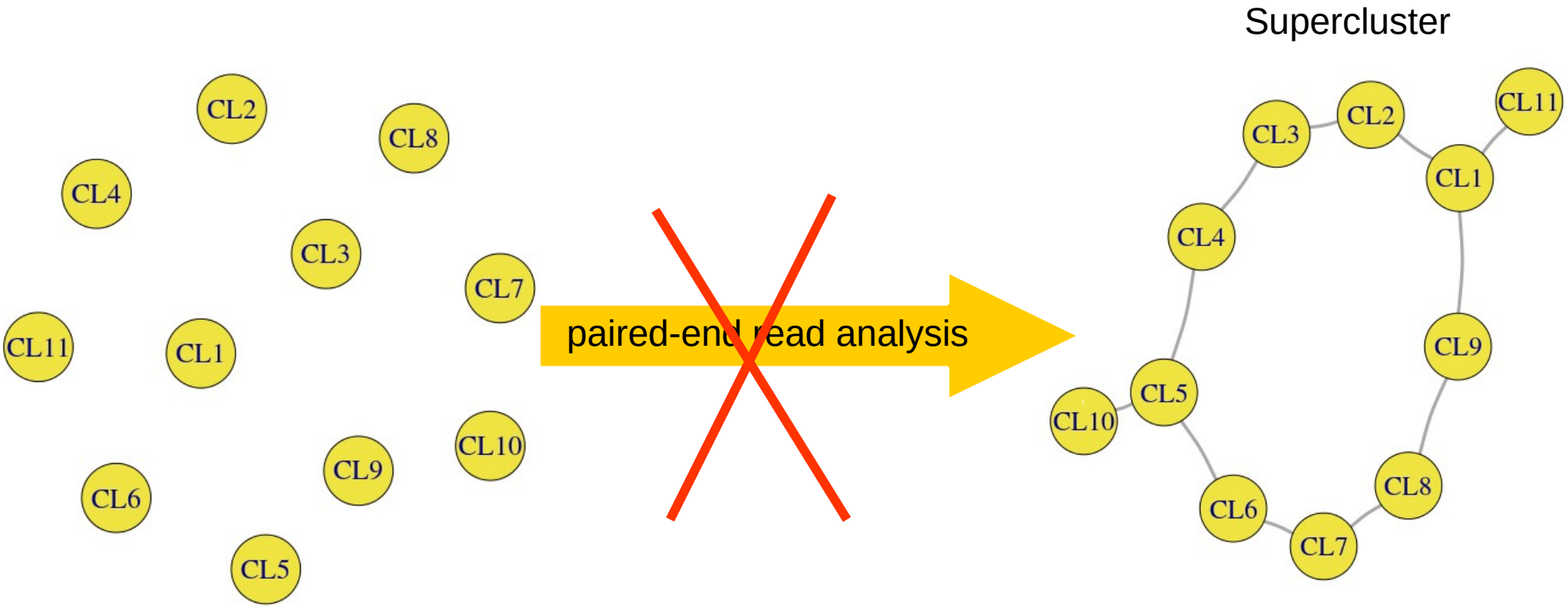
no connection $k_{x,y} = 0$

Supercluster Identification

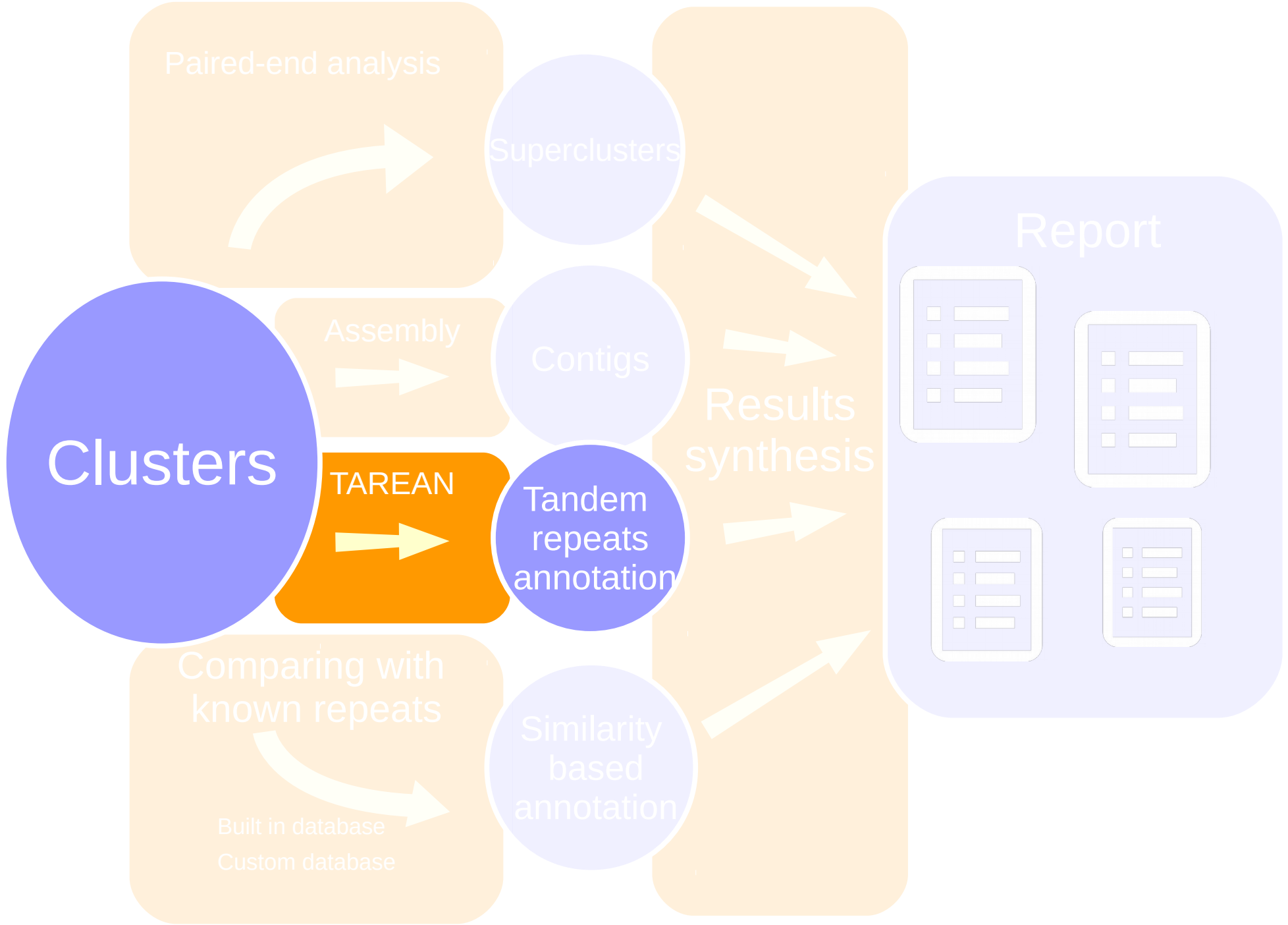


Supercluster Identification

In the absence of paired-end reads
clusters are equivalent to superclusters

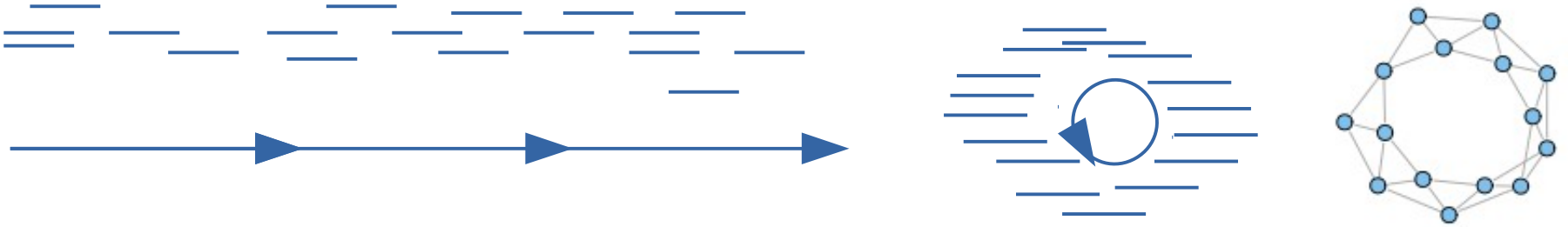


Cluster centered analysis

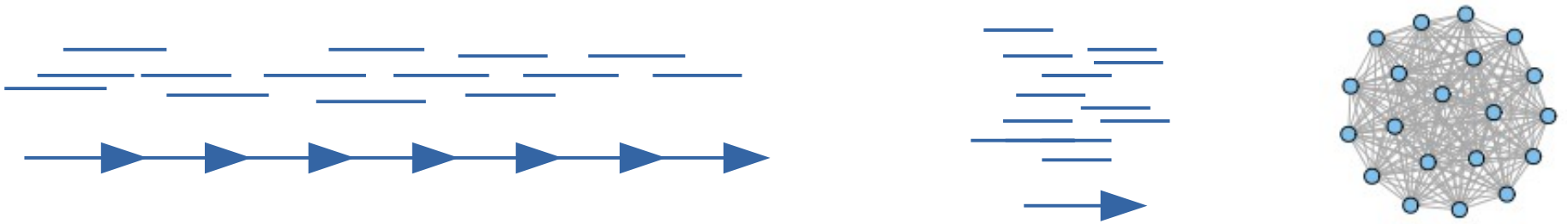


Tandem Repeat Analyzer - TAREAN

Read length << monomer

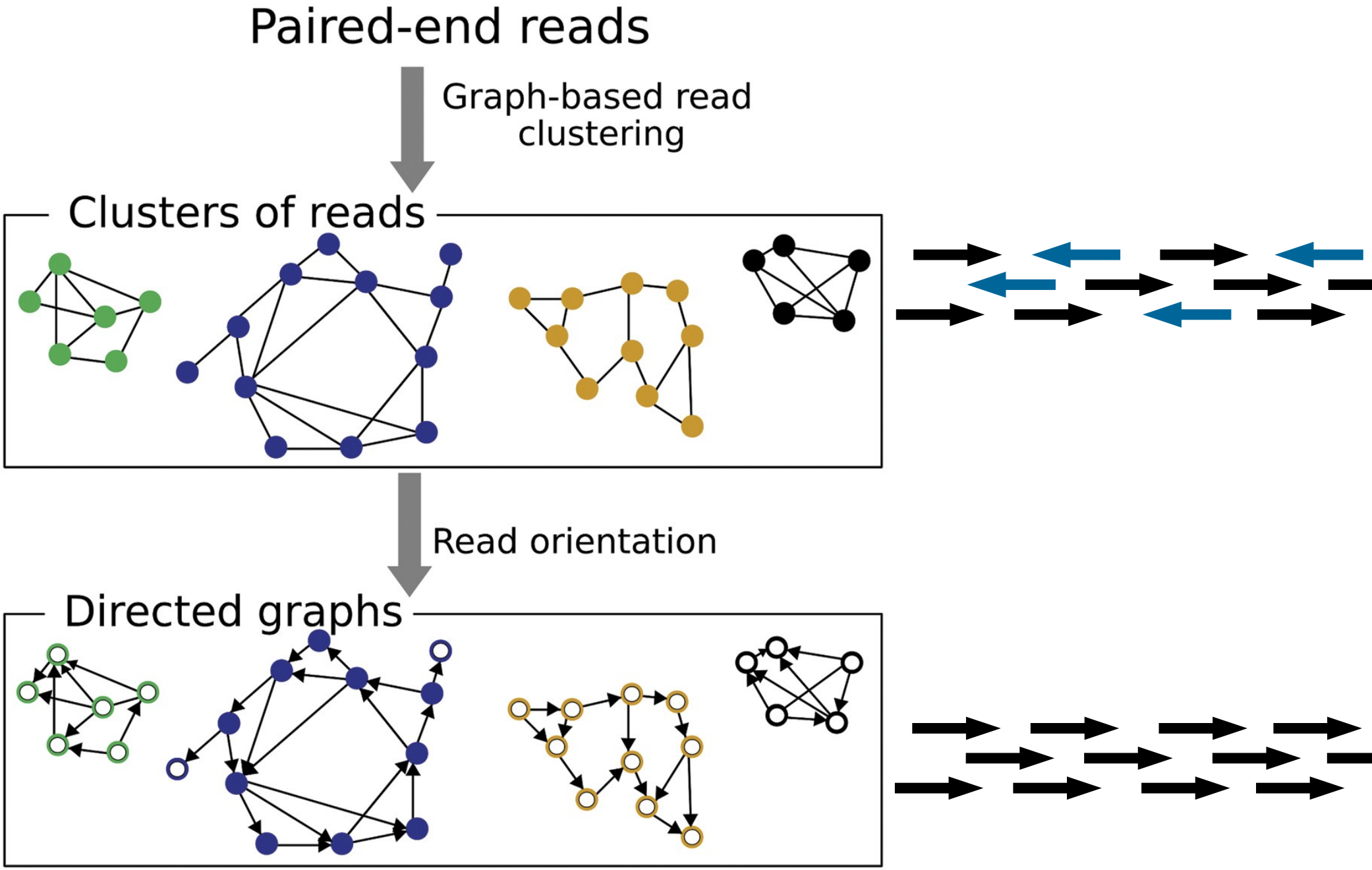


Read length >= monomer

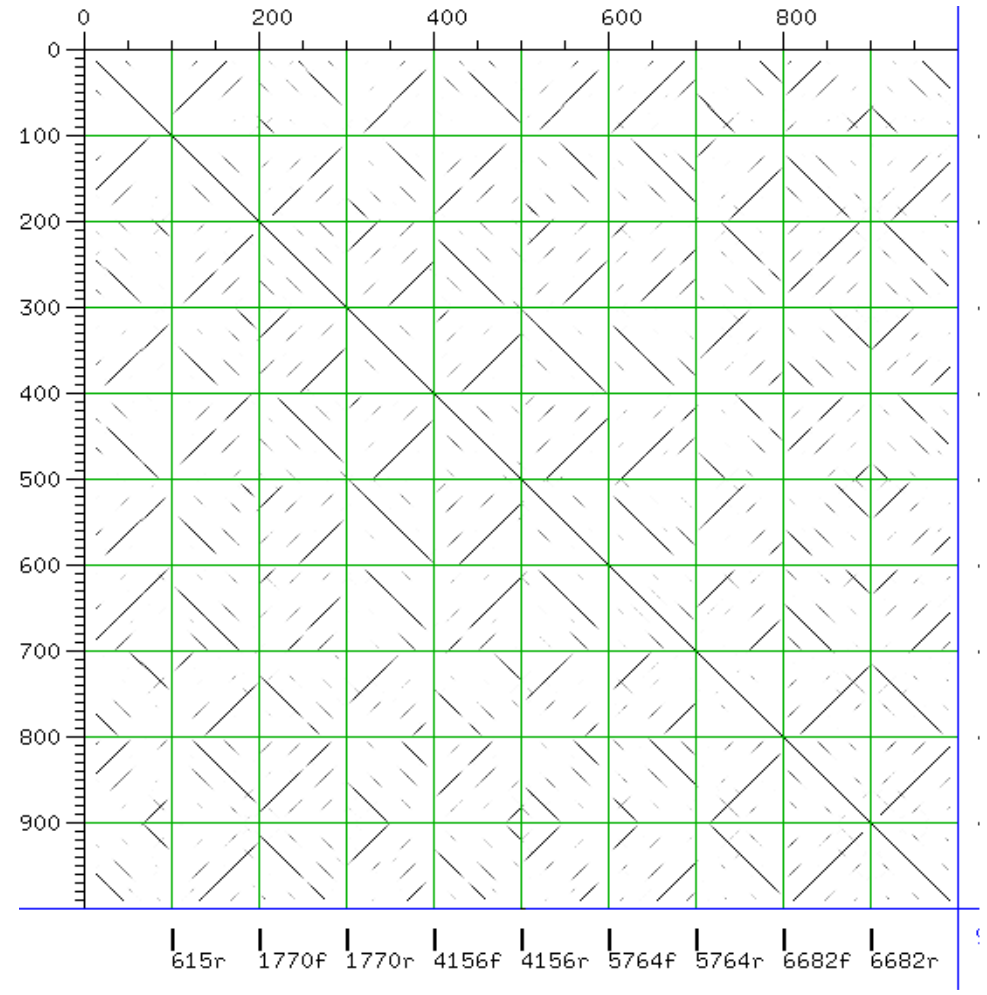


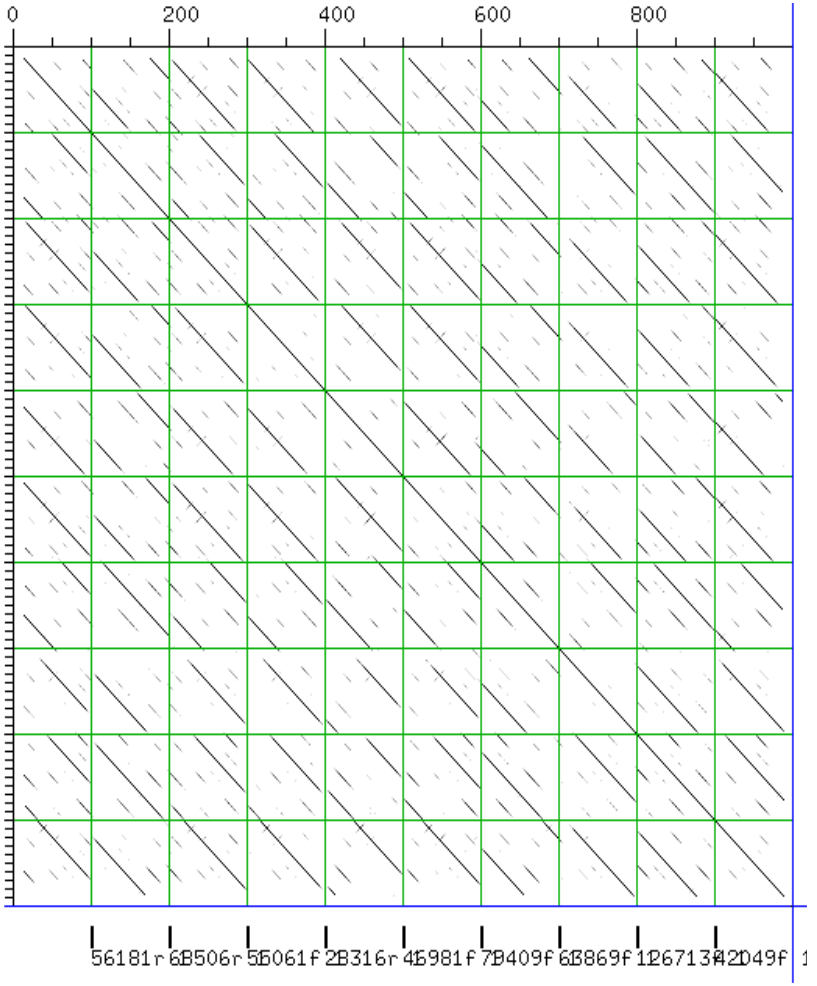
TAREAN calculates **graph layout** and provide automatic analysis of **graph topology** with the aim to identify **tandem repeats**

Tandem Repeat Analyzer - TAREAN



Tandem Repeat Analyzer - TAREAN

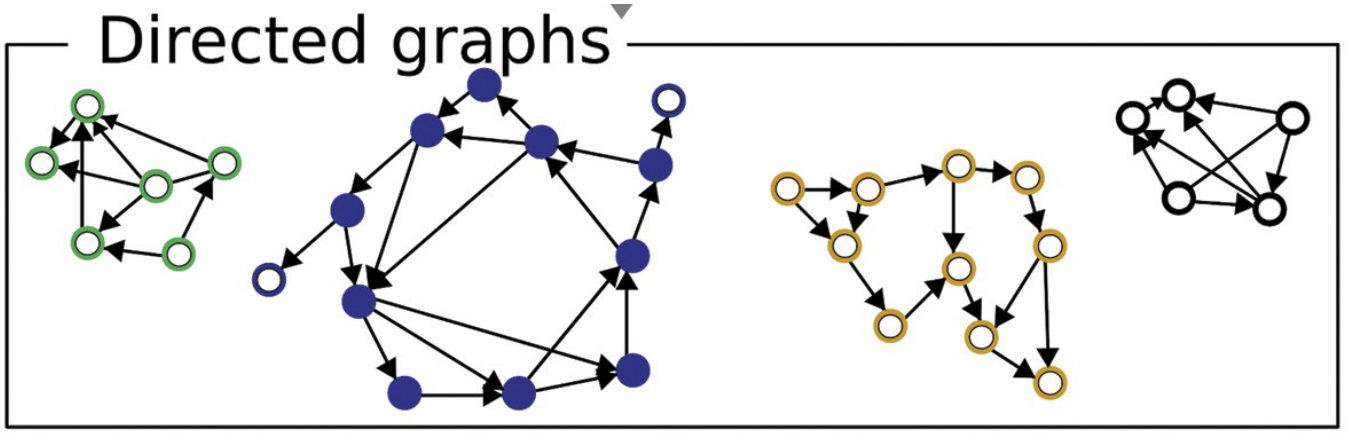


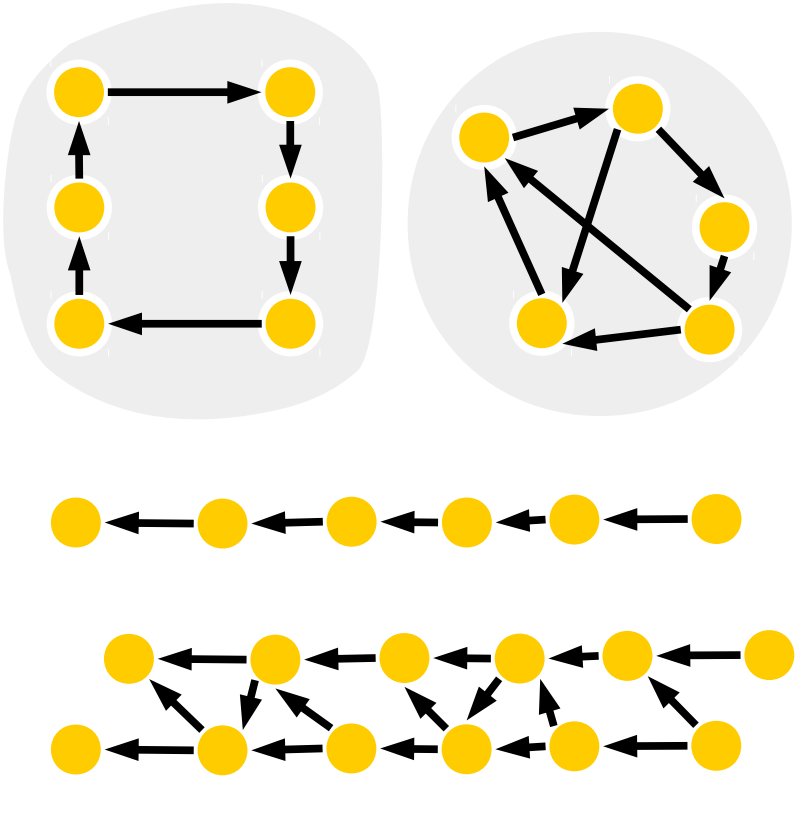


Original reads

Oriented reads

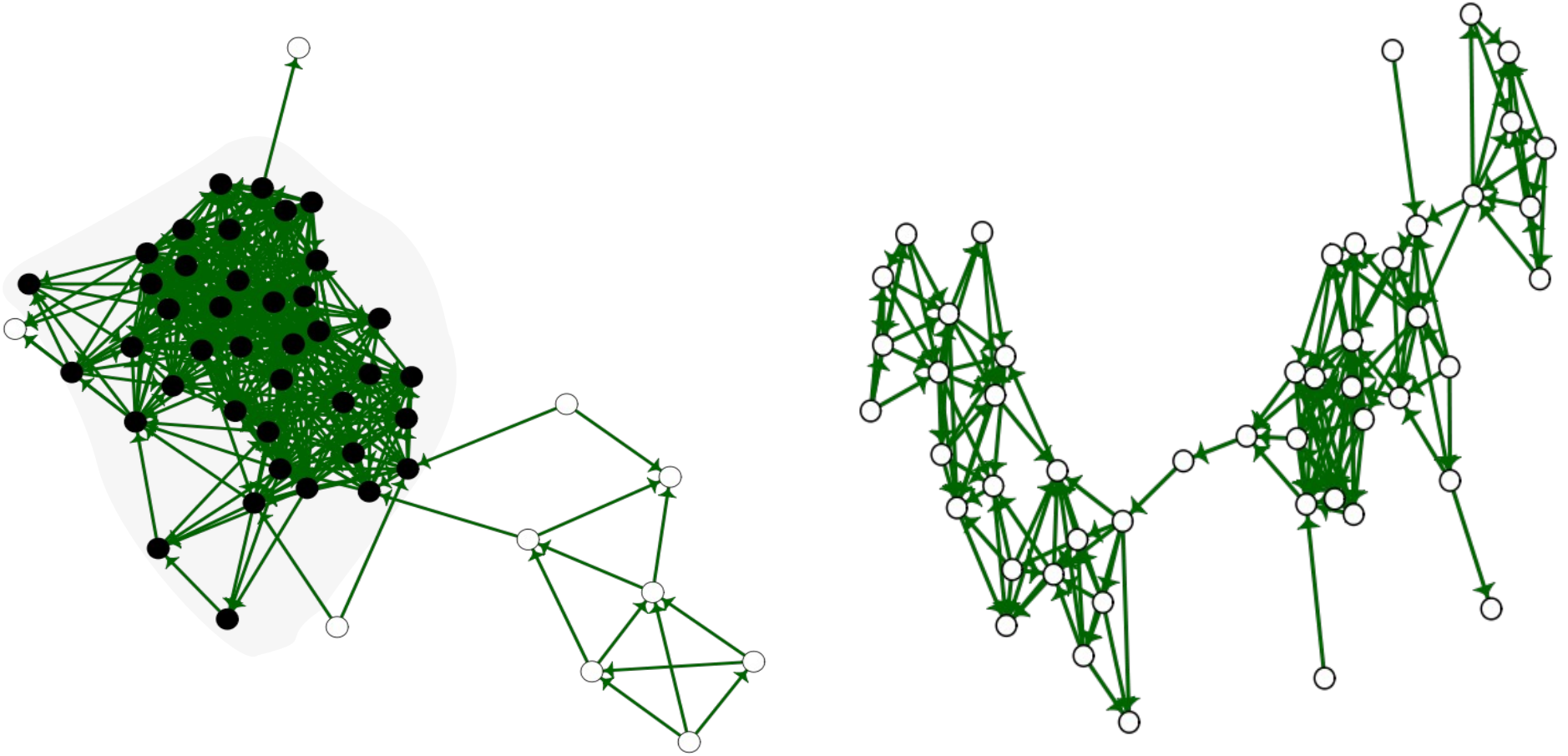
Tandem Repeat Analyzer - TAREAN





A directed graph is called **strongly connected** if every vertex is reachable from every other vertex

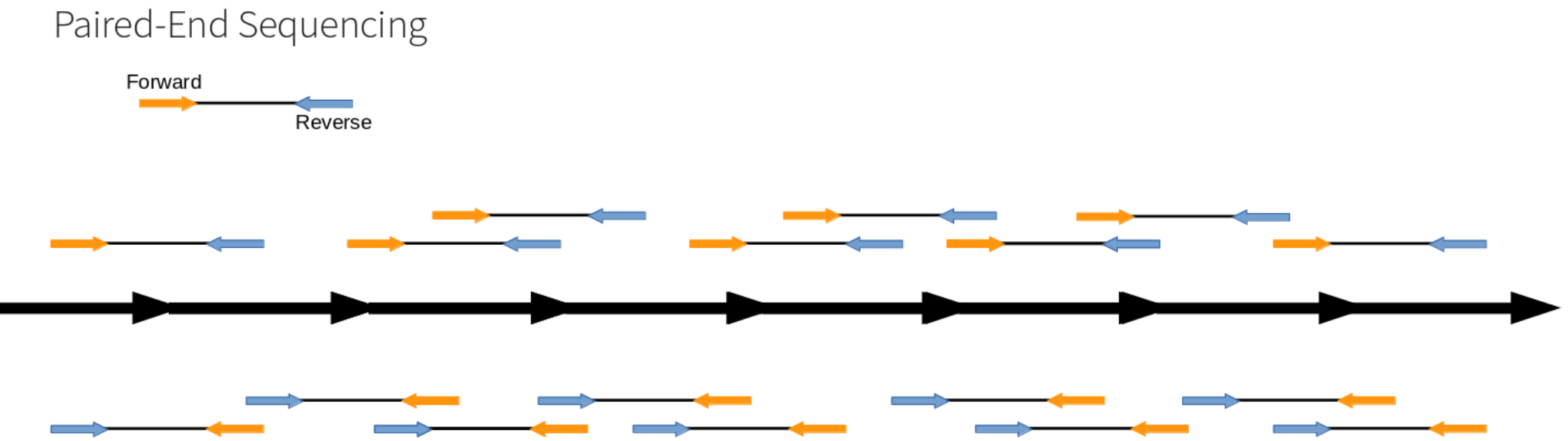
Tandem Repeat Analyzer - TAREAN



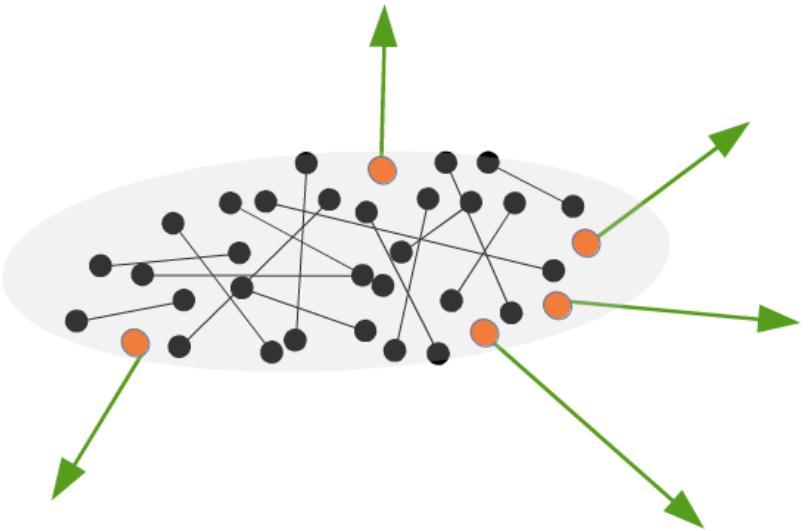
$$\mathbf{C} = \frac{\text{size of the largest strongly connected components}}{\text{Total graph size}}$$

C – Connected component index

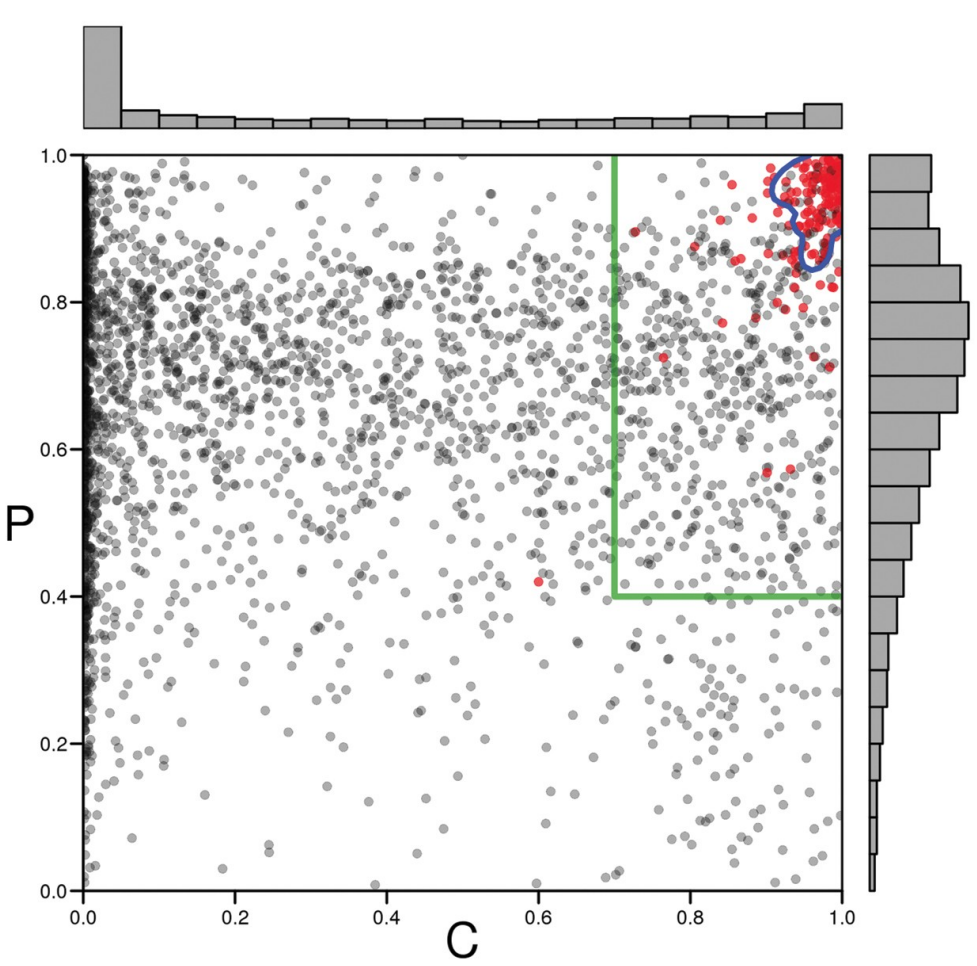
Tandem Repeat Analyzer - TAREAN

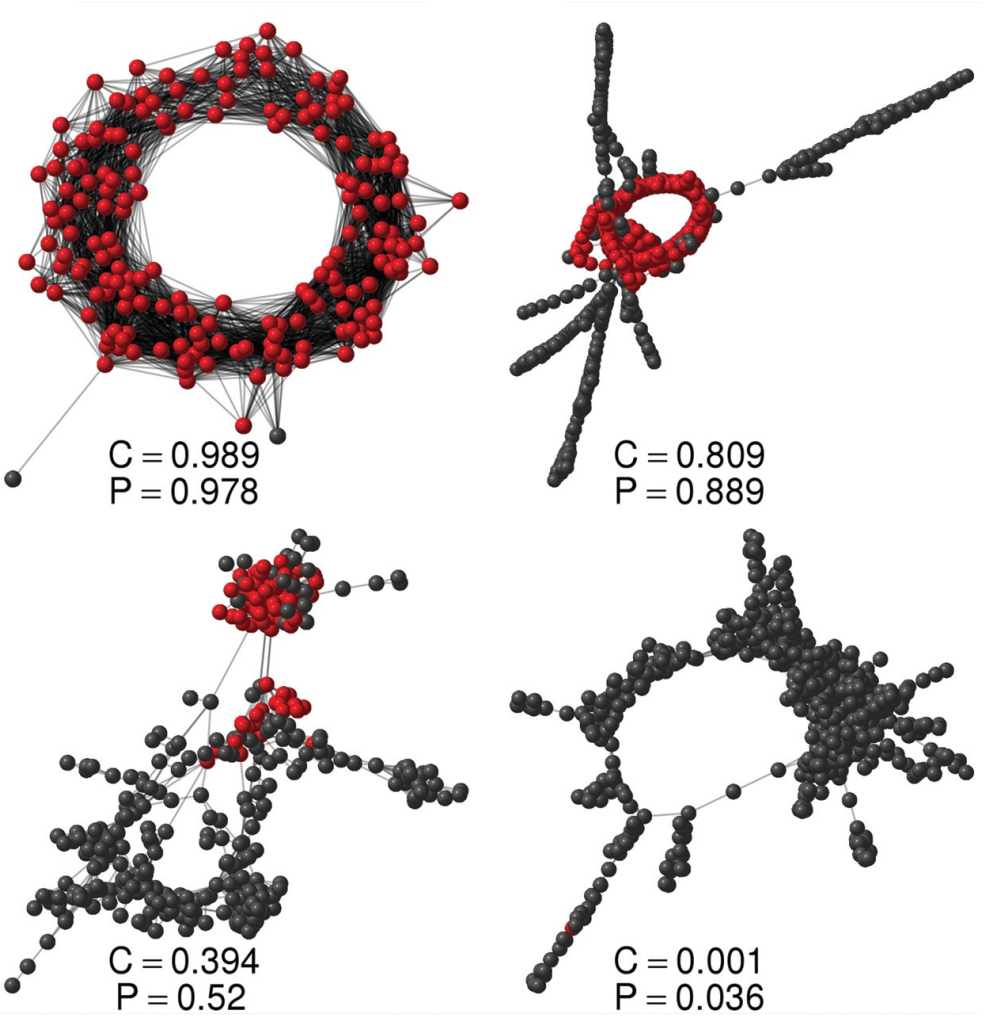


Pair completeness = fraction of complete pairs in cluster

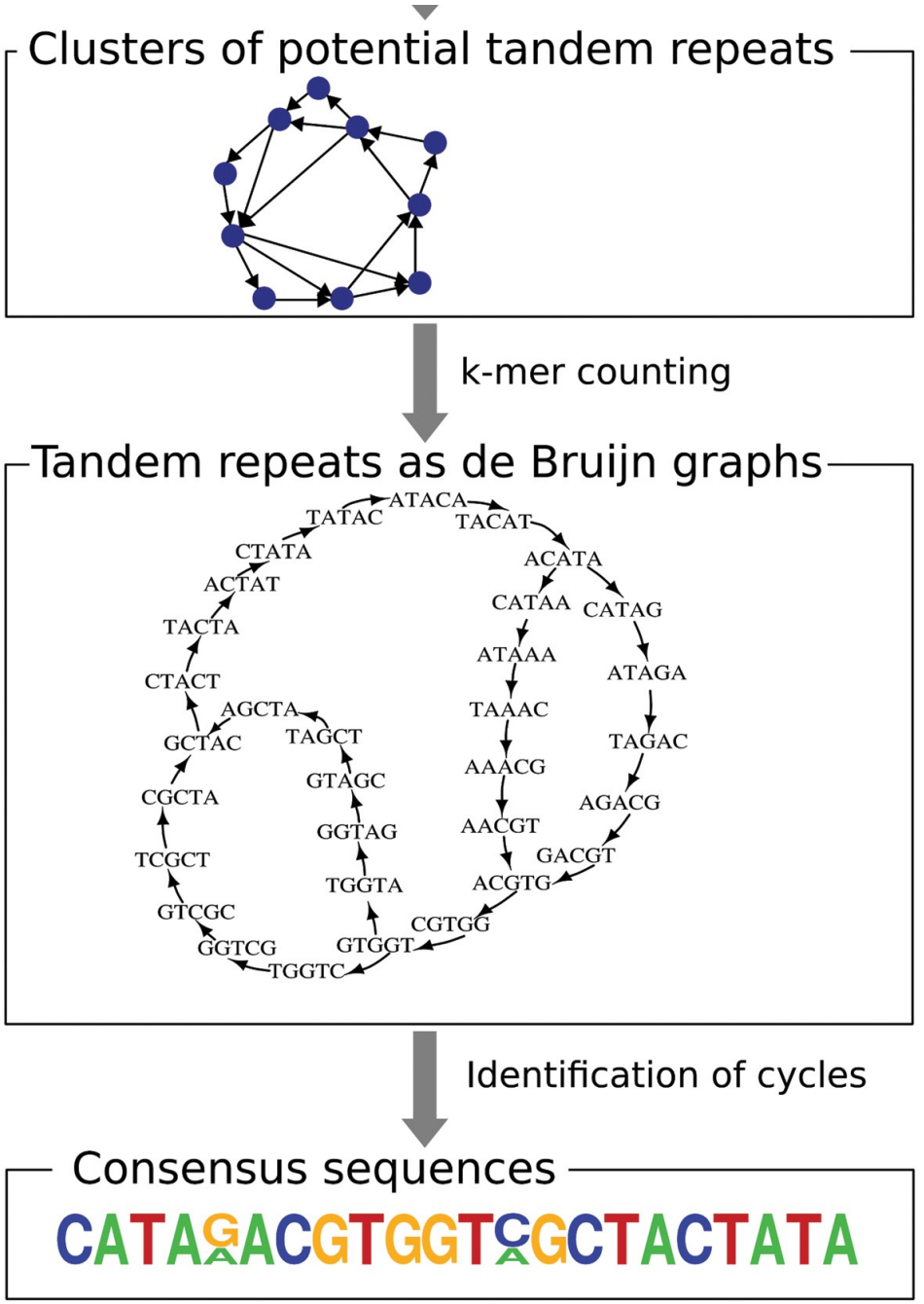


Tandem Repeat Analyzer - TAREAN





Tandem Repeat Analyzer - TAREAN

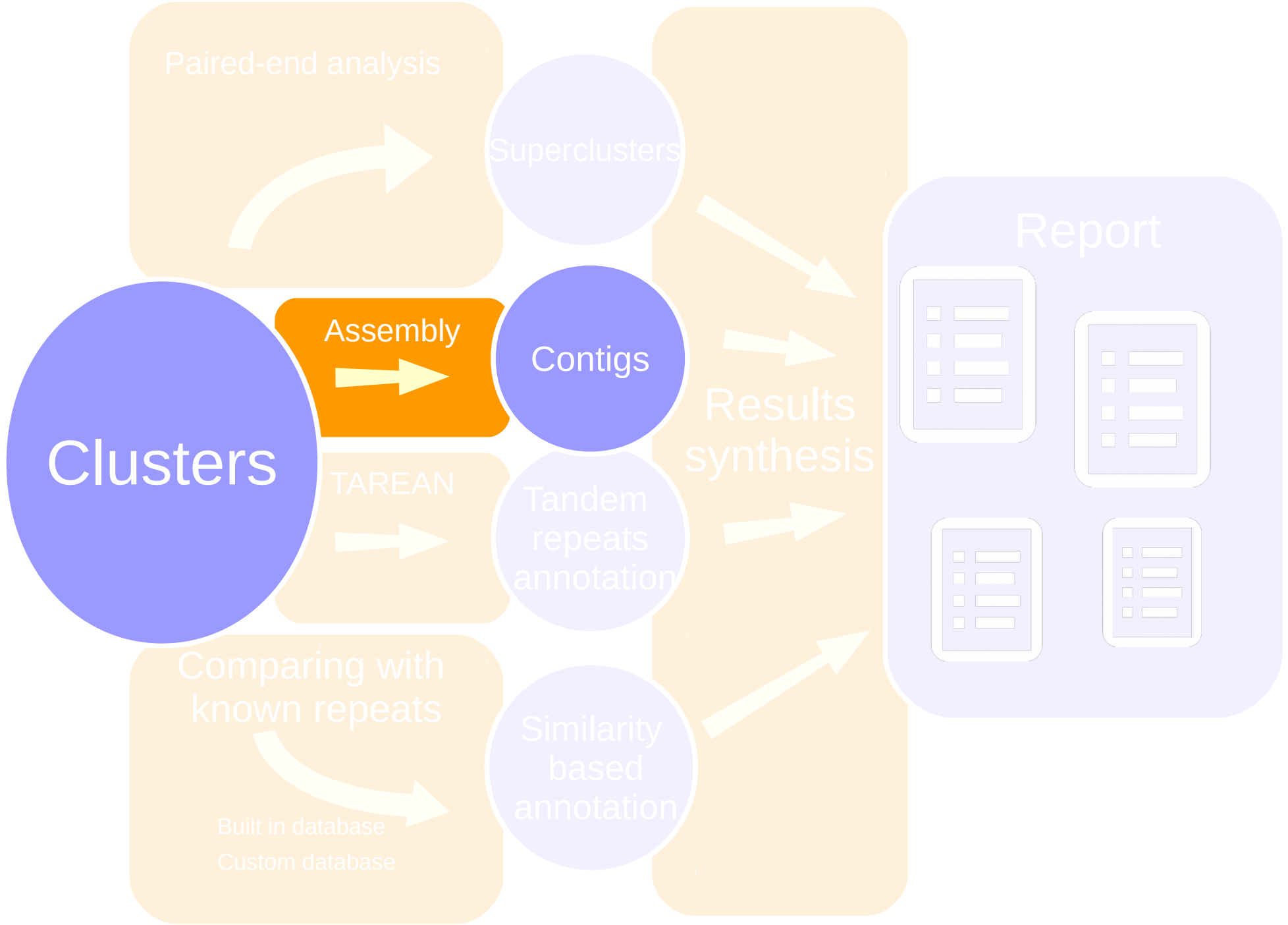


Tandem Repeat Analyzer - TAREAN

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



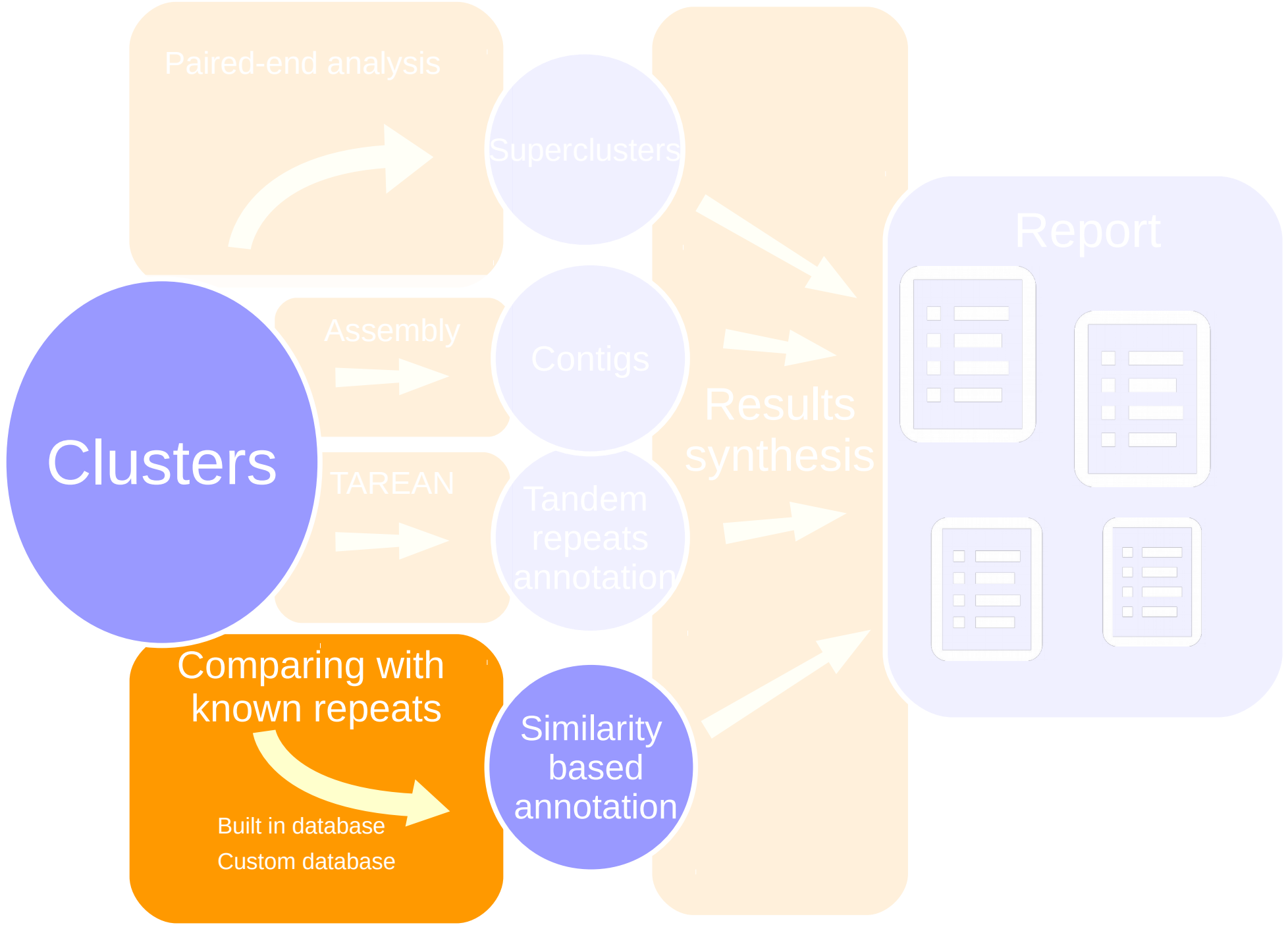
Assembly

Reads are assembled by CAP3 program, each cluster separately:

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

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

Cluster centered analysis



Similarity based annotation

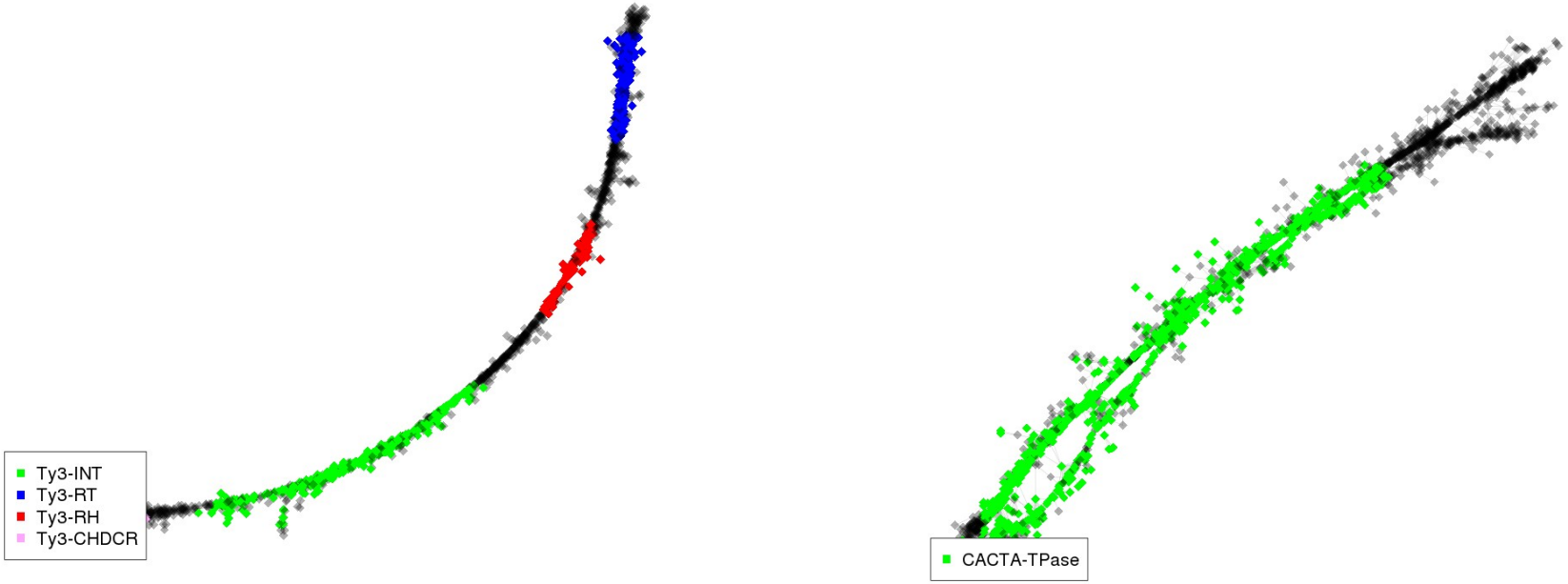
All reads are compared with:

- 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

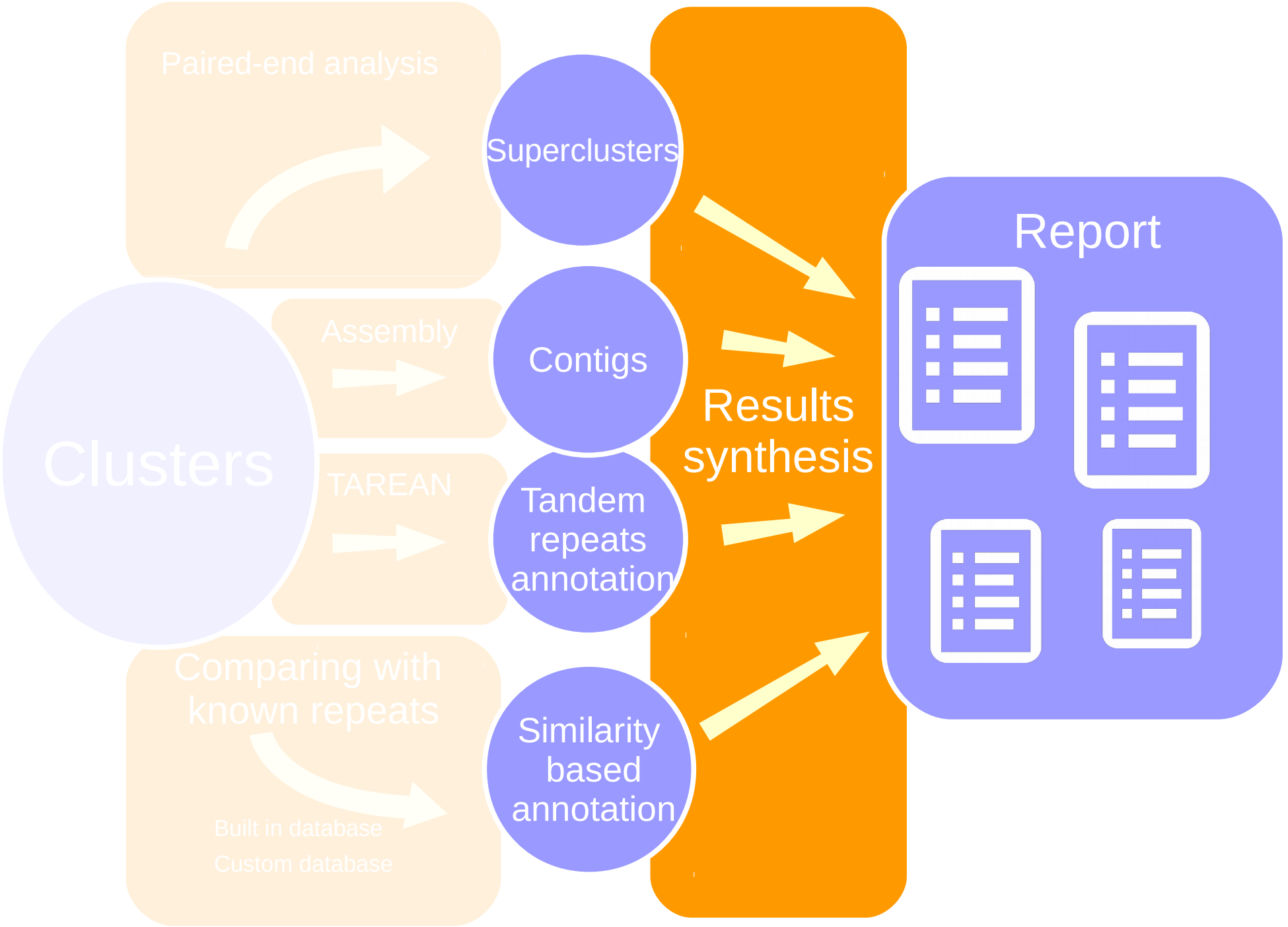
All reads are compared with:

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

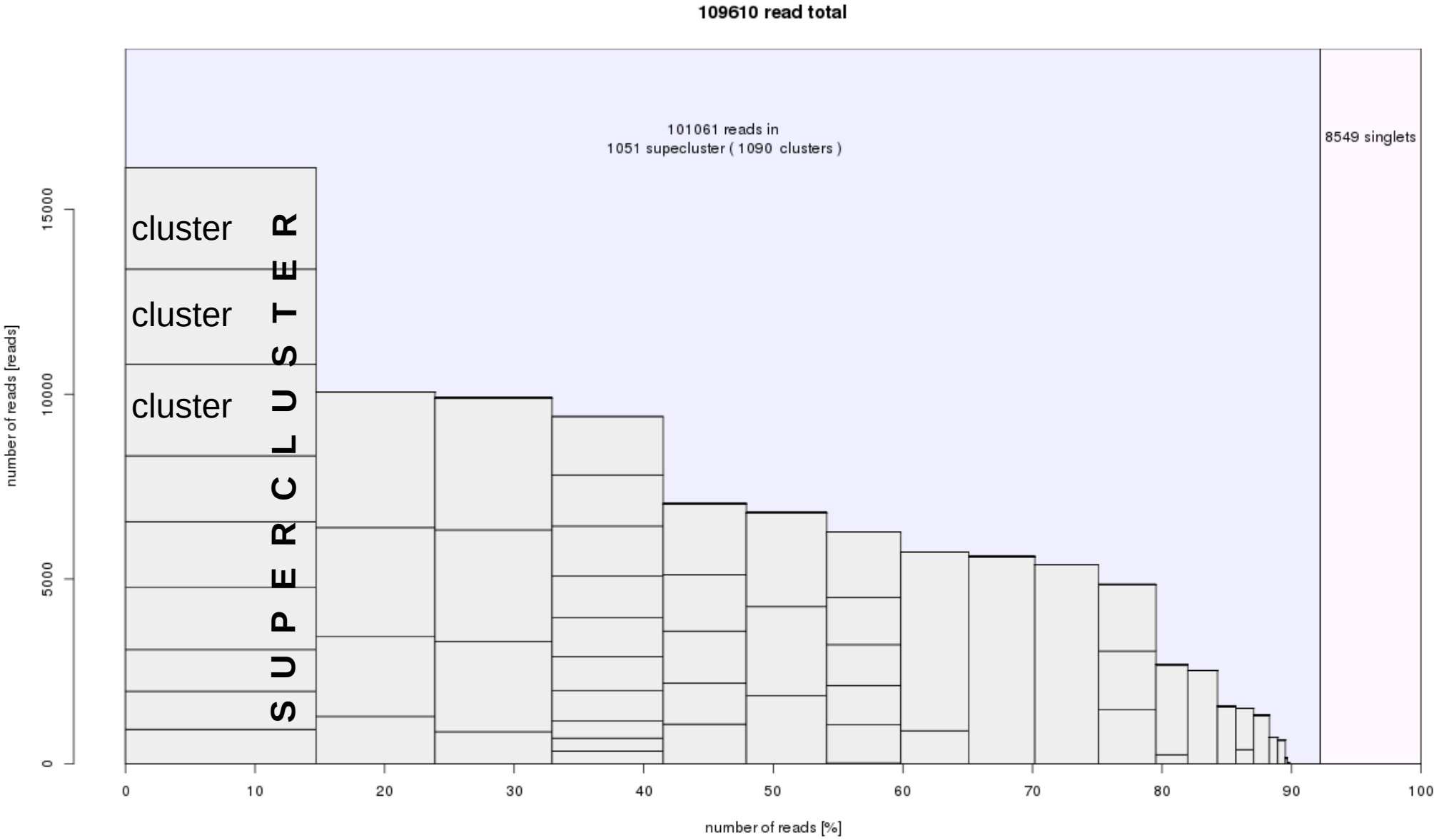
  Library of repeats as DNA sequences in fasta format. The required format for IDs in a custom library is :

  >reapeatname#class/subclass

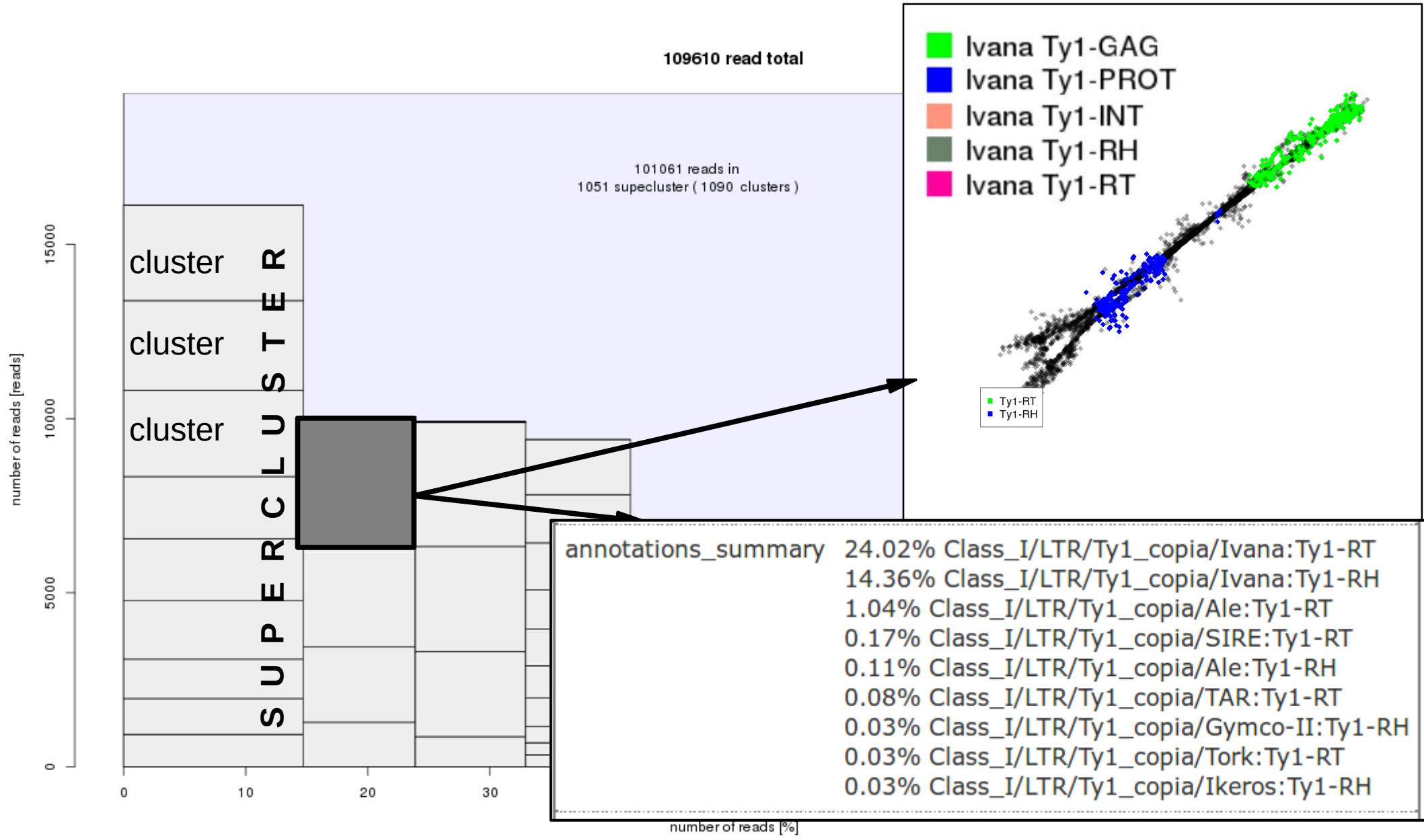
Reporting


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

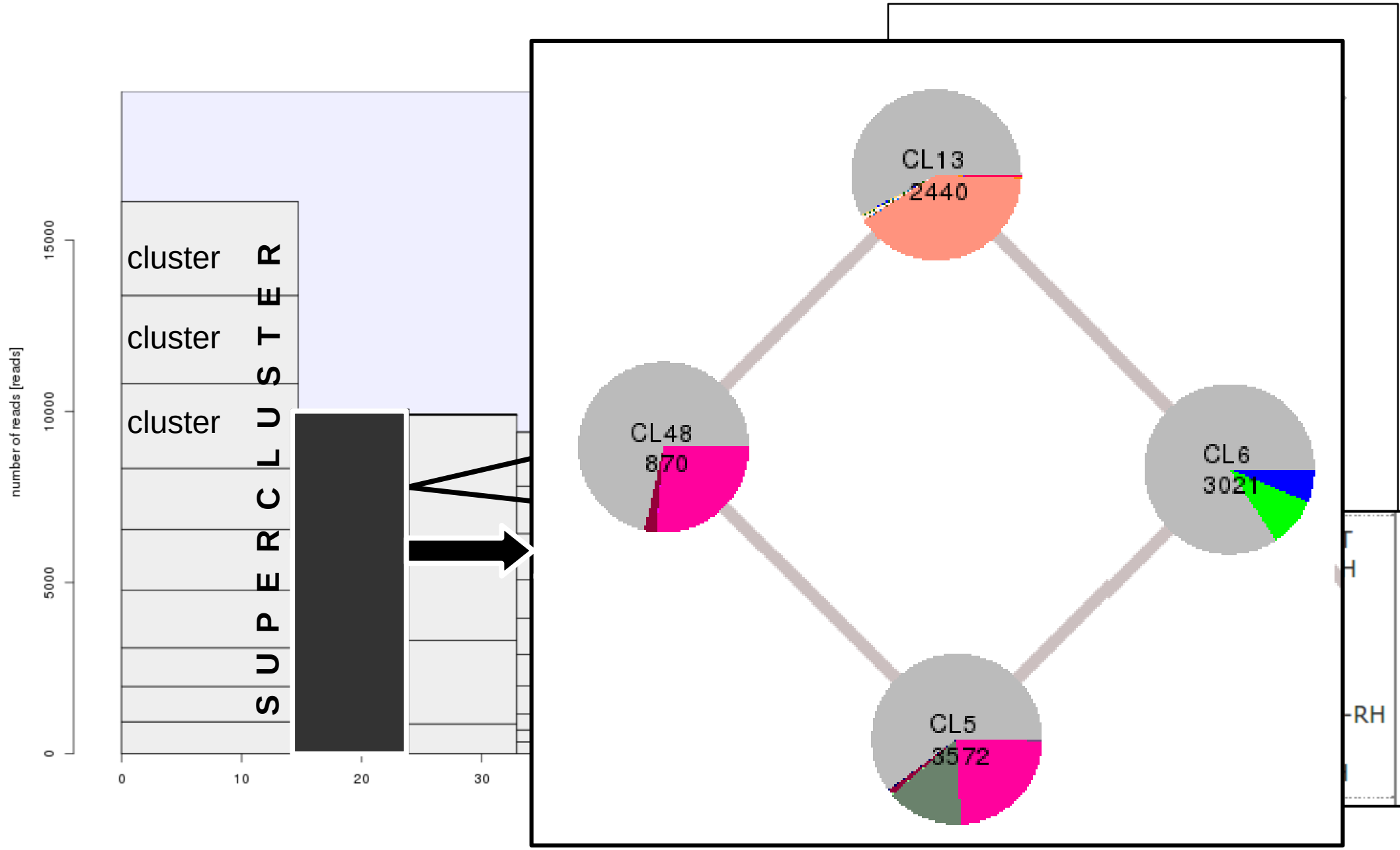
Reporting


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 [%]

Reporting


109610 read total
101061 reads in
1051 supecluster (1090 clusters)
cluster
cluster
cluster
SUPERCLUSTER
number of reads [reads]
0
5000
10000
15000
0
10
20
30
number of reads [%]
Ivana Ty1-GAG
Ivana Ty1-PROT
Ivana Ty1-INT
Ivana Ty1-RH
Ivana Ty1-RT
Ty1-RT
Ty1-RH
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


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

Reporting

	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),

Ivana

number of reads [reads]

5000

0

SUPERCL

0

10

20

30

870

CL6

3021

CL5

3572

-RH

Reporting - Repeat annotation summary

	Genome_proportion[%]	Nsuperclusters	Nclusters	Nreads
Unclassified_repeat	0	0	0	0
├--rDNA	0	0	0	0
│ ├--45S_rDNA	14.71	1	9	16125
│ │ ├--18S_rDNA	0	0	0	0
│ │ ├--25S_rDNA	0	0	0	0
│ │ └--5.8S_rDNA	0	0	0	0
│ └--5S_rDNA	2.3	1	1	2524
├--satellite	20.11	4	8	22047
└--mobile_element	0	0	0	0
├--Class_I	0	0	0	0
│ ├--SINE	0	0	0	0
│ ├--LTR	0	0	0	0
│ │ ├--Ty1_copia	0	0	0	0
│ │ │ ├--Ale	0	0	0	0
│ │ │ ├--Alesia	0	0	0	0
│ │ │ └--Angela	0	0	0	0

	Genome_proportion[%]	Nsuperclusters	Nclusters	Nreads
organelle	0	0	0	0
├--plastid	8.57	1	10	9396
└--mitochondria	0	0	0	0

	Genome_proportion[%]	Nsuperclusters	Nclusters	Nreads
Unclassifed	6.92	4	4	7582

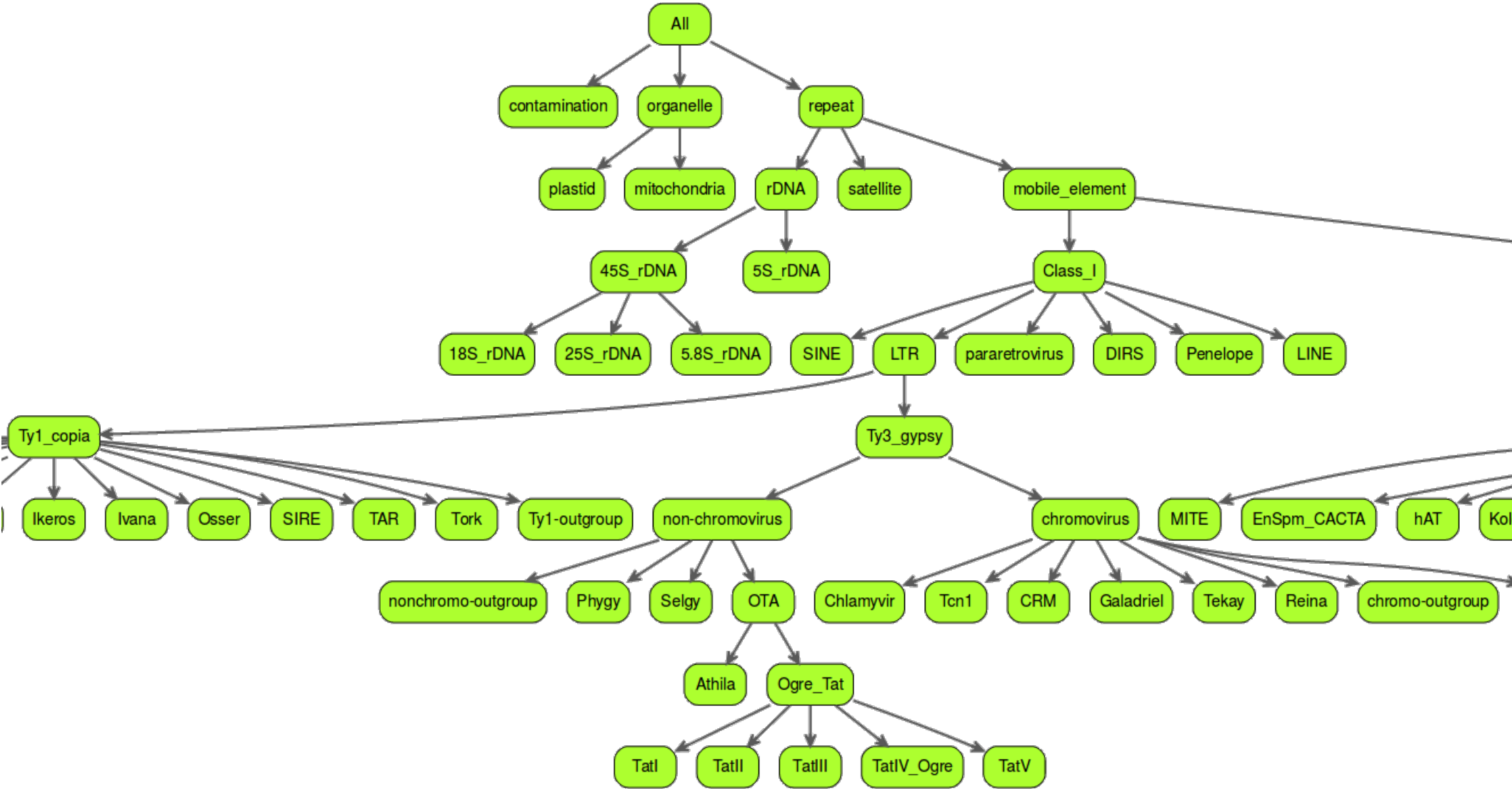
	Genome_proportion[%]	Nsuperclusters	Nclusters	Nreads
contamination	6.42	1	5	7040

		├--TatII	0	0	0	0
		├--TatIII	0	0	0	0
		├--TatIV_Ogre	0	0	0	0

0
0
0
0
11407
0
0
0
4850
0
0
0
0
0
0
0
0
0
0

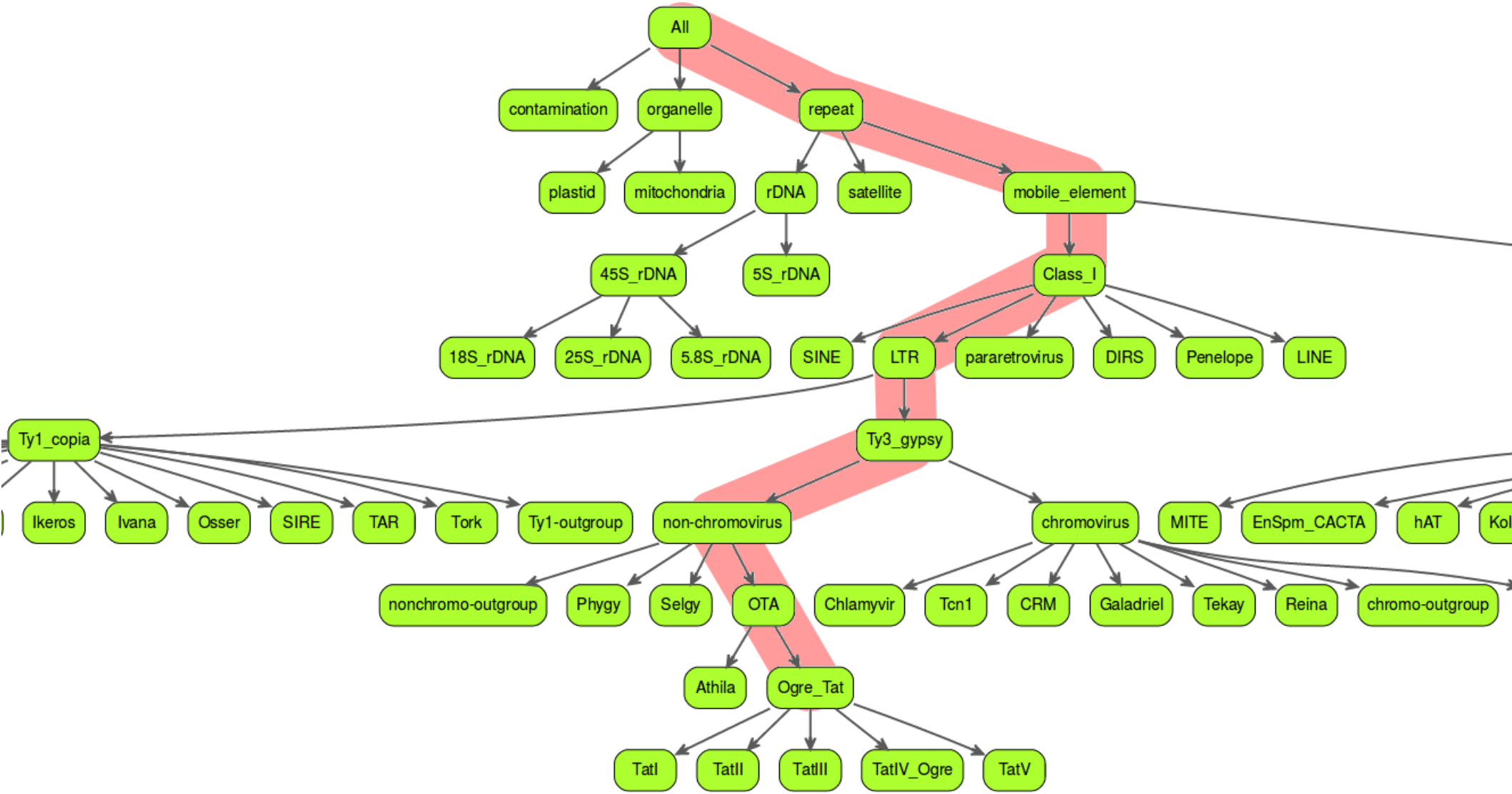
What is the best hit?

Top-down classification

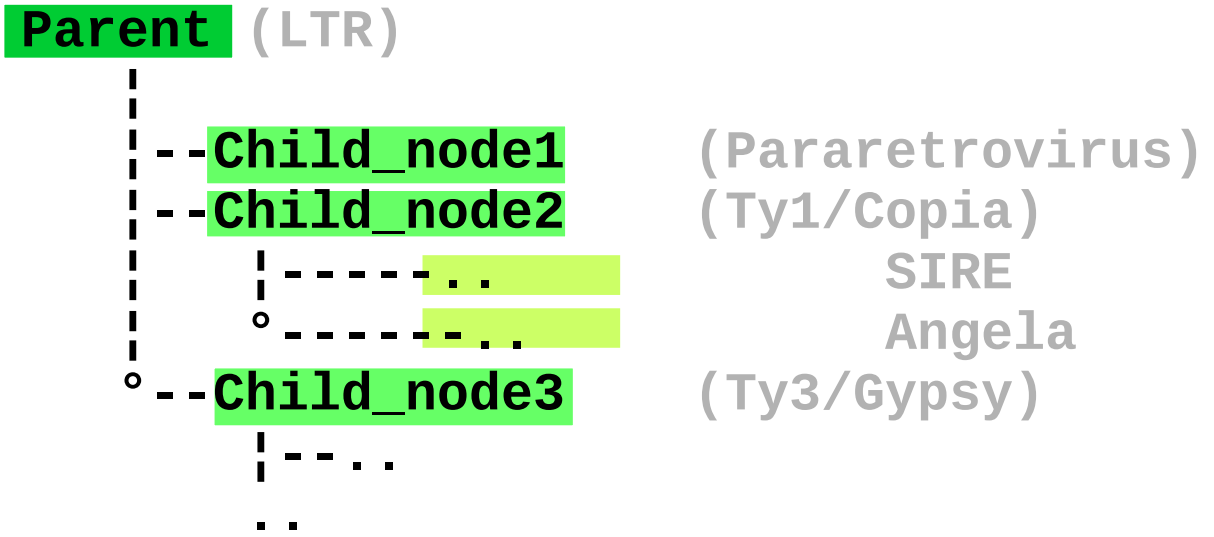


What is the best hit?

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

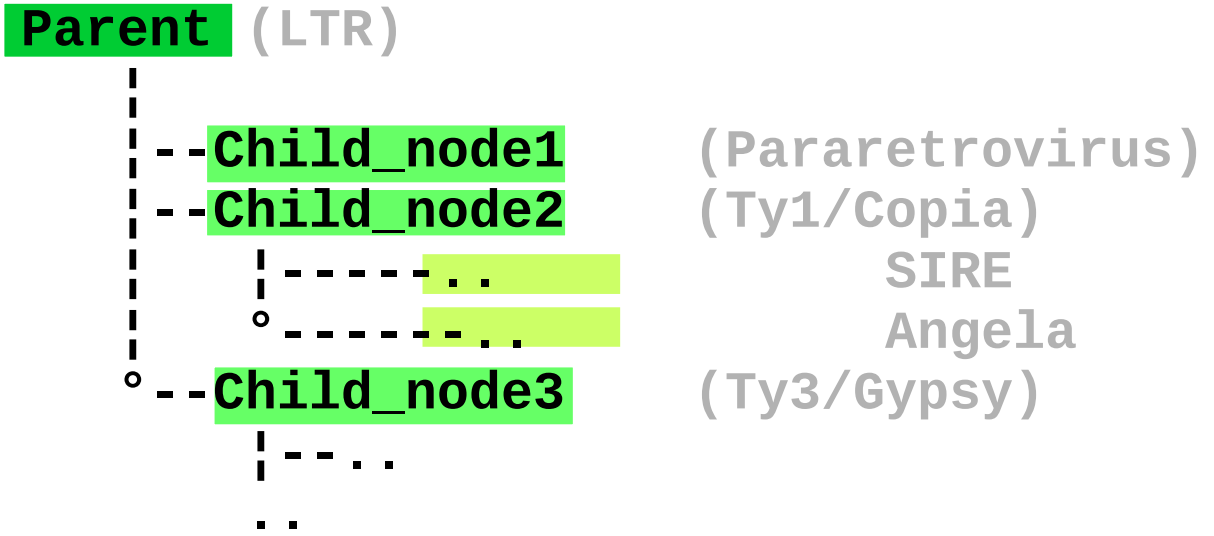
H_p number of hits in parent node

$H_{c,x}$ 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: $\frac{H_{c,1}^2}{N} > 2.5$

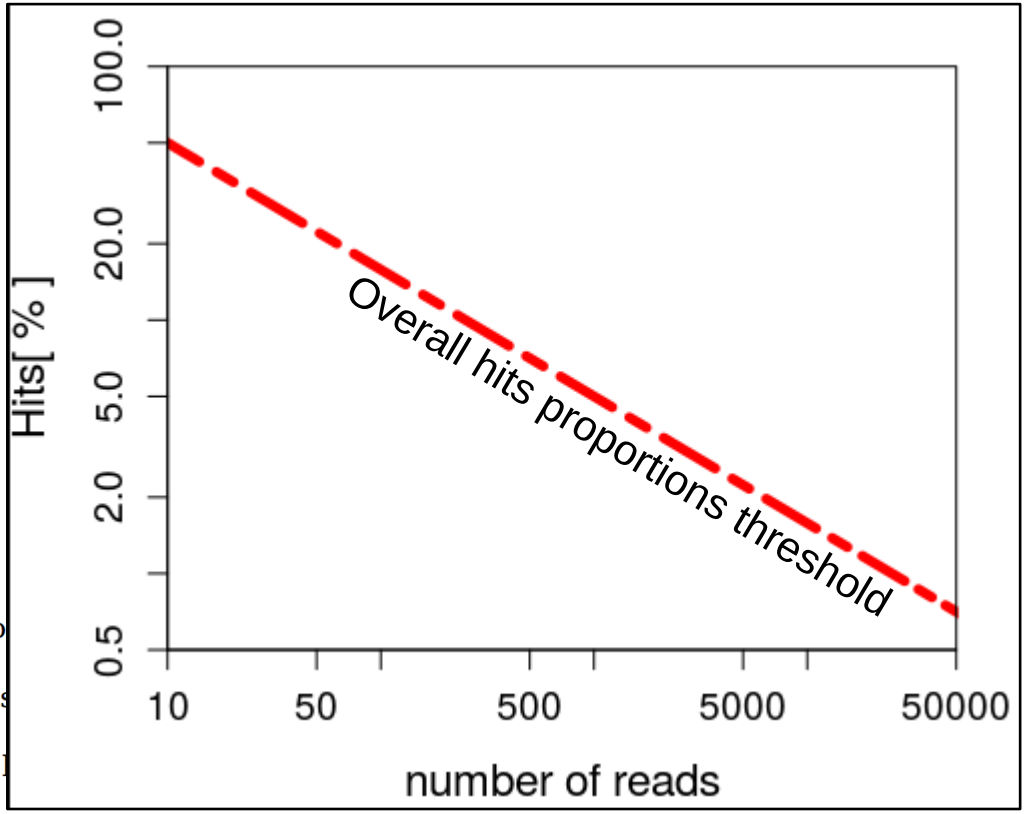
N number of reads in supercluster

H_p number of hits in parent node

$H_{c,x}$ number of hits in children node x, nodes are sorted by number o

$H_{c,1}$ number of hits in child node with the highest number of hits (bes

$H_{c,2}$ number of hits in child node with the second highest number of



Full Repeat Analysis vs. Tandem Repeat Analysis

Clustering can be run in two different modes:

- **Full Repeat Analysis**

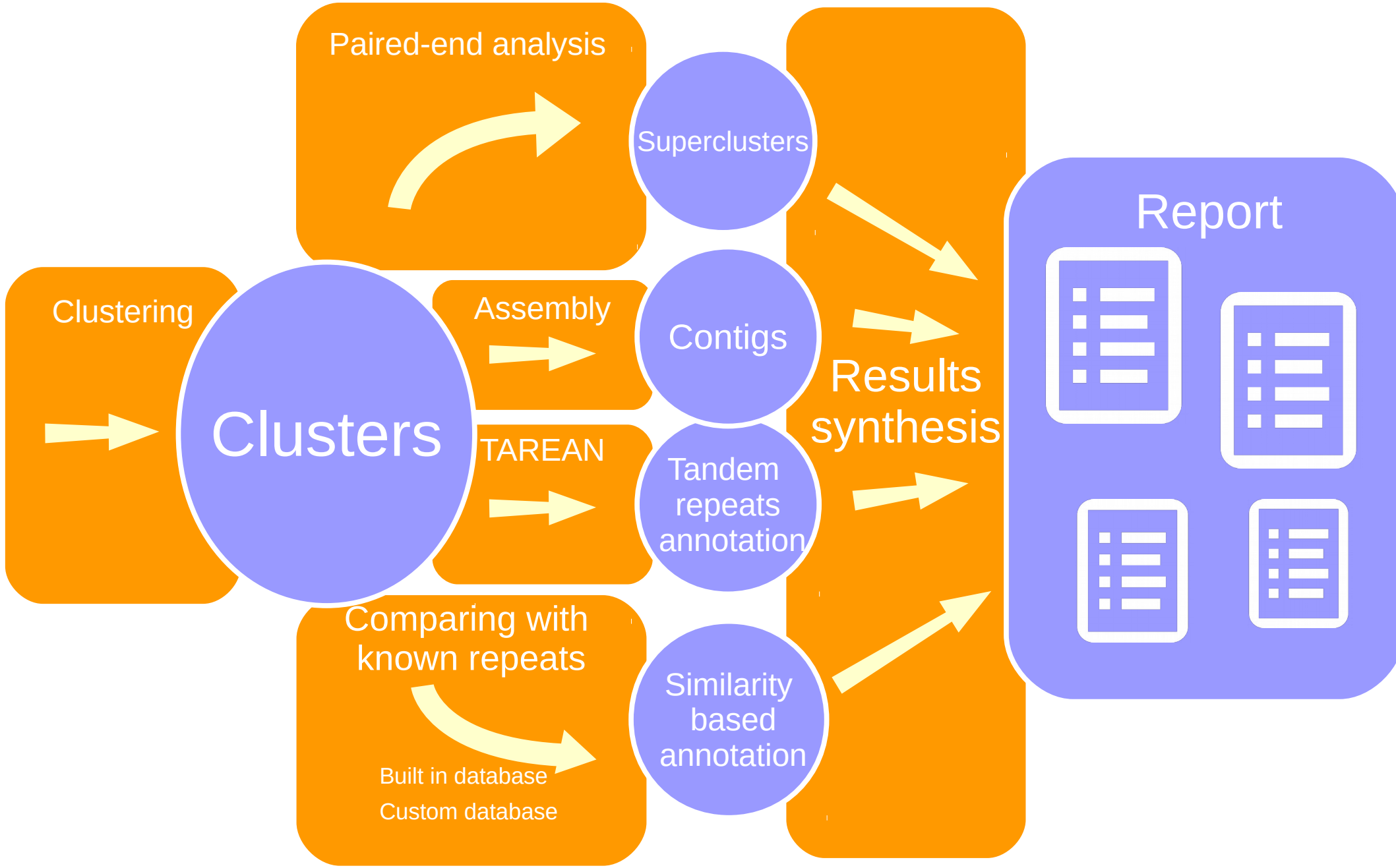
  Focus on all types of repeat but less sensitive satellite detection

- **Tandem Repeat Analysis**

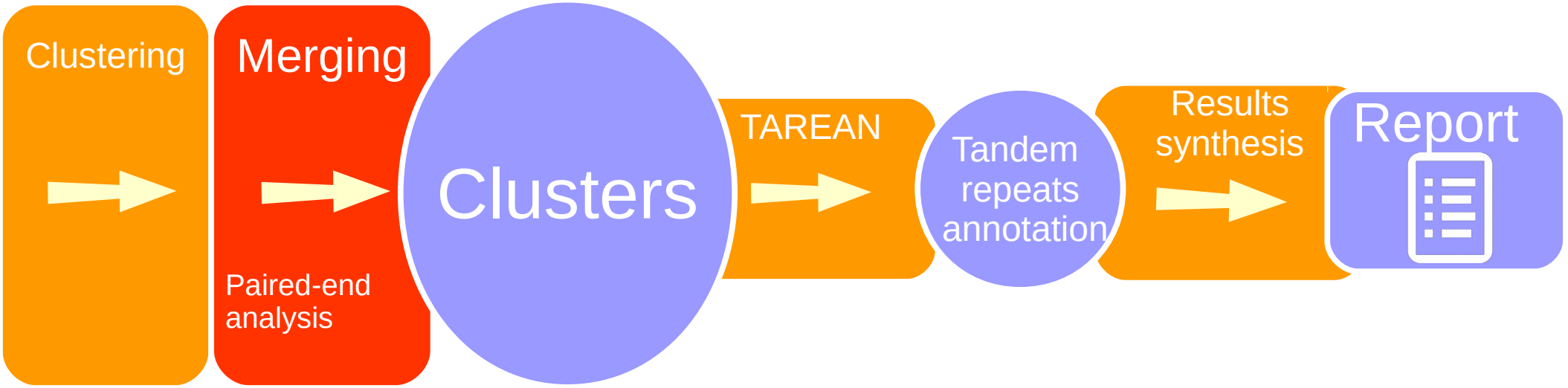
  Focus on tandem repeat detection only

  Better sensitivity of satellite identification

Full Repeat Analysis



Tandem Repeat Analysis



- 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

COFFE BREAK

RepeatExplorer 2.0

Discover repeats in your next generation sequencing data