# Tandem Repeat Analyzer - TAREAN

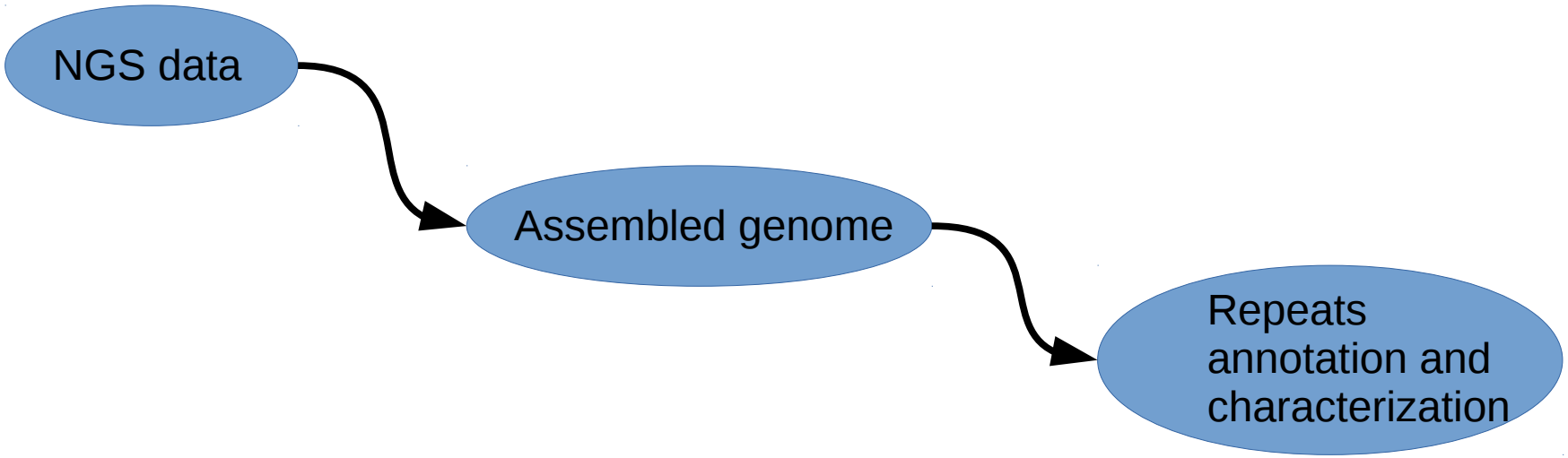
## Extension of RepeatExplorer clustering

Automated identification of clusters derived from tandem repeats and reconstruction of monomer consensus sequence

Petr Novák
Laboratory of Molecular Cytogenetics
Institute of Plant Molecular Biology
Biology Centre ASCR

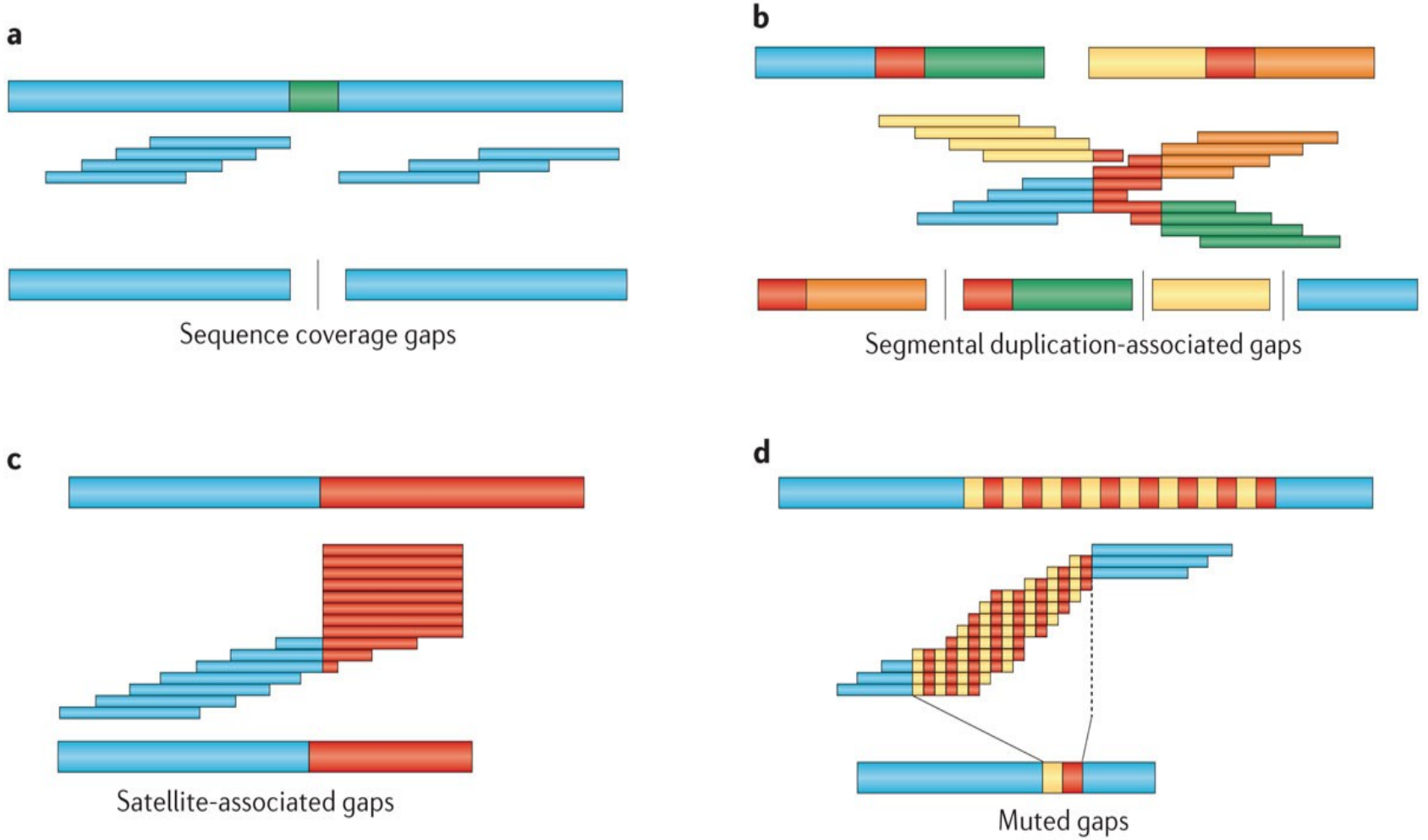
# Tandem Repeat Analyzer - TAREAN

Current bioinformatics focus on NGS data, genome assembly and utilize assembled genome as reference.



# Tandem Repeat Analyzer - TAREAN

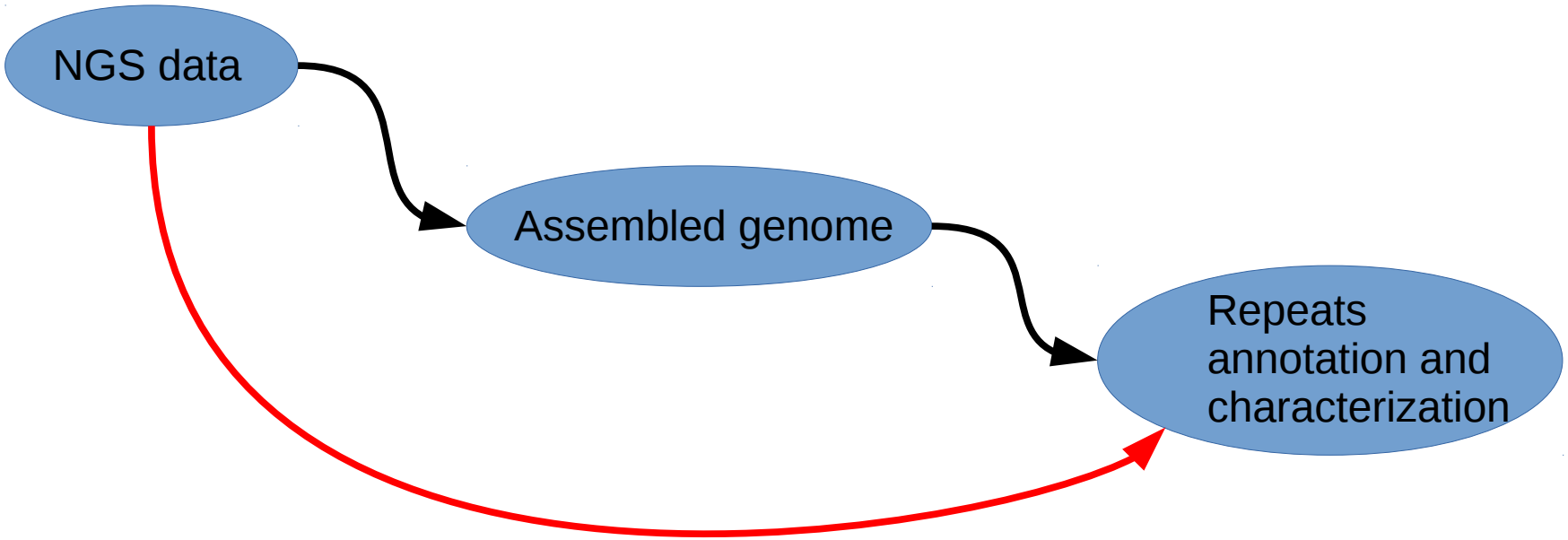
Types of genome assembly gaps.



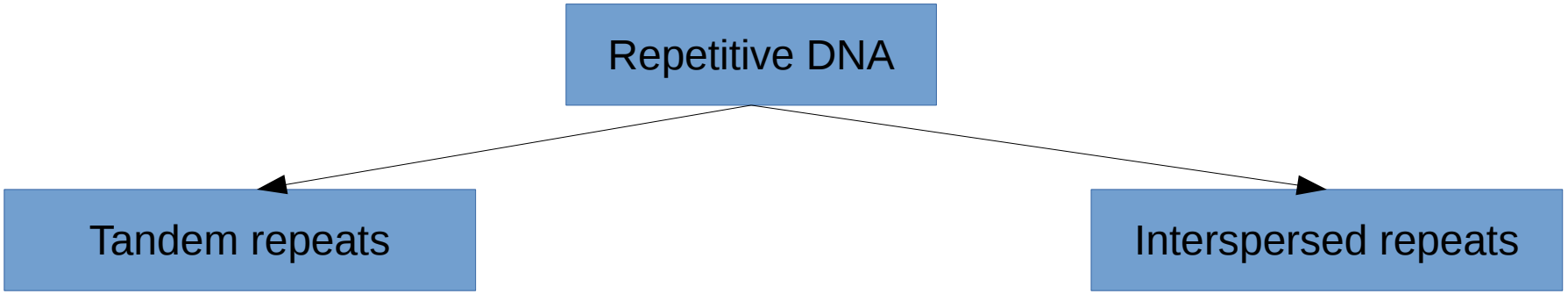

Nature Reviews Genetics, 16, 627–640 (2015) doi:10.1038/nrg3933

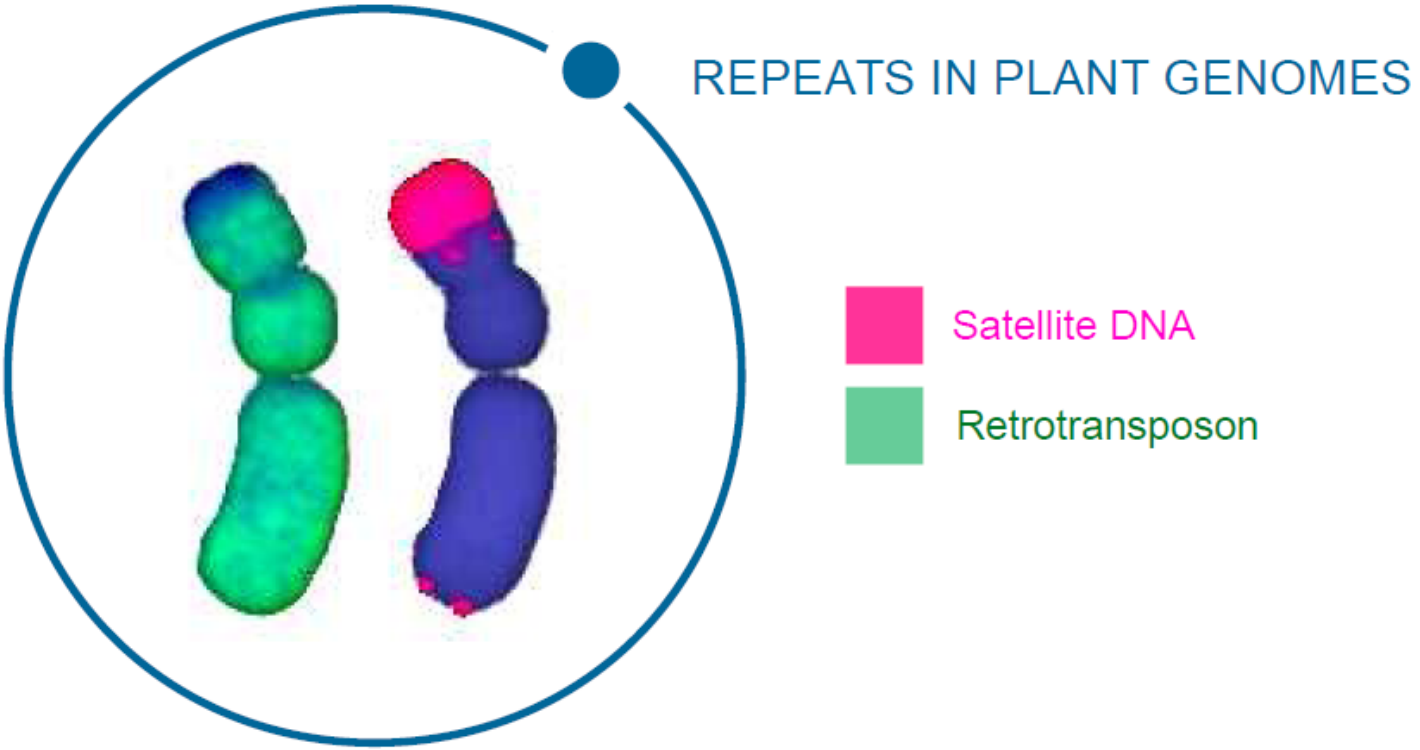
# Tandem Repeat Analyzer - TAREAN

Current bioinformatics focus on NGS data, genome assembly and utilize assembled genome as reference.

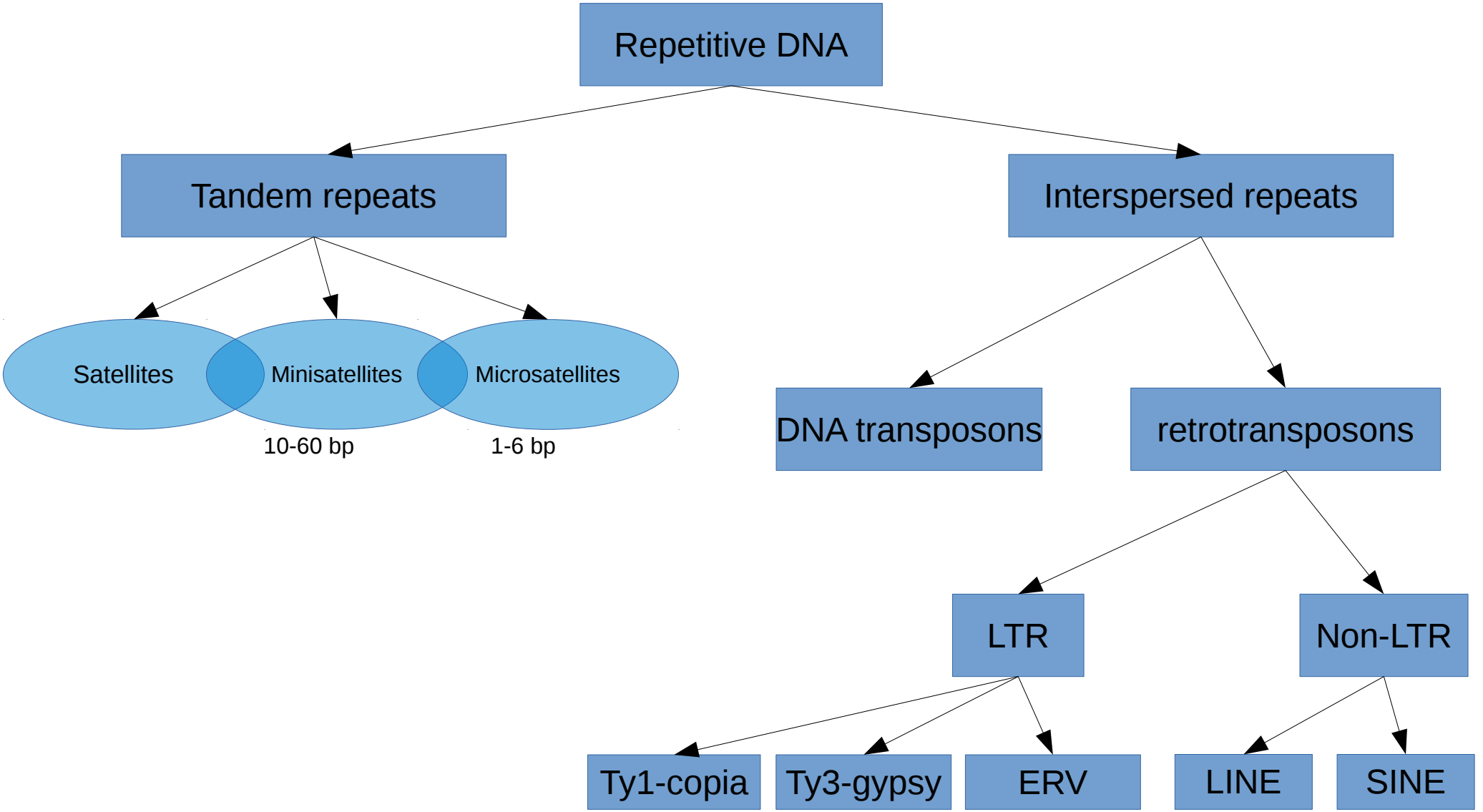


# Tandem Repeat Analyzer - TAREAN



# Tandem Repeat Analyzer - TAREAN



# Tandem Repeat Analyzer - TAREAN

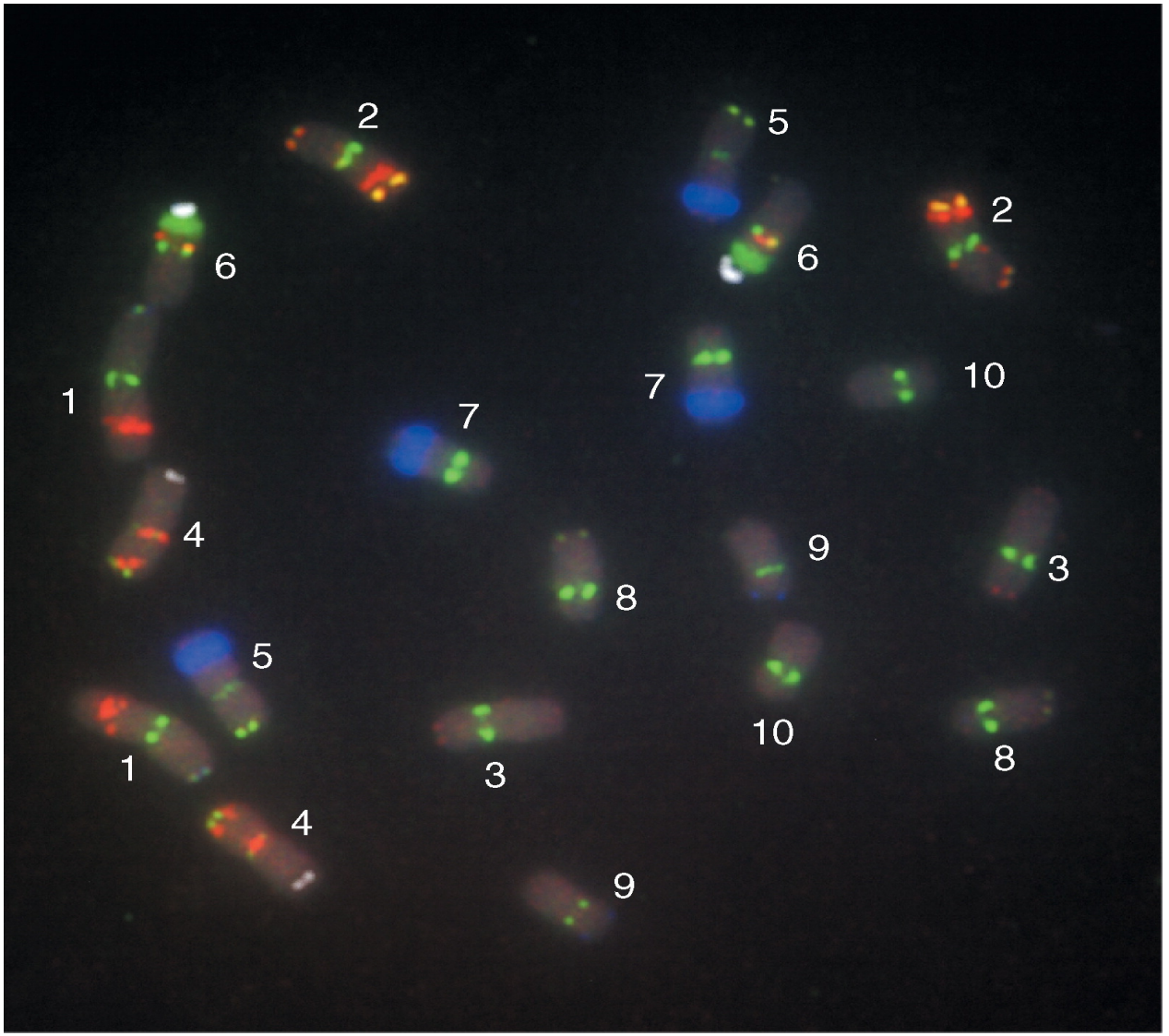
## Tandem Repeats

- Significant portion of eukaryotic genomes
- Formation and maintenance of chromatin structure
  - Centeromeric, pericentromeric regions
  - Subtelomeric
  - Heterochromatin
- Chromosome pairing and segregation
- Gene expression

- Important marker in cytogenetic research

# Tandem Repeat Analyzer - TAREAN

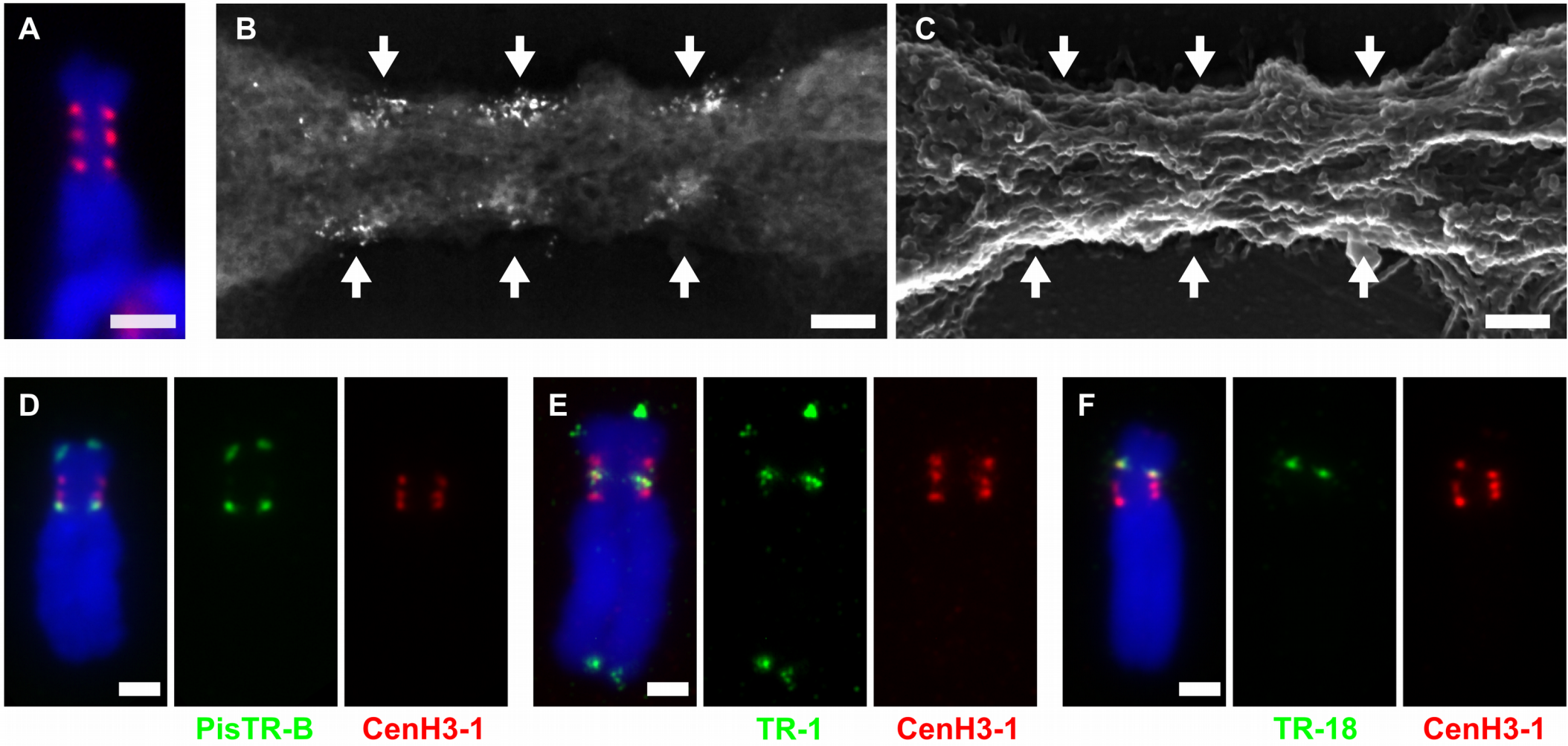
Identification of the ten pairs of somatic chromosomes of maize inbred line B37 using nine fluorochrome labeled DNA probes




Knob 180bp repeat
5SrDNA
CentC satellite
...

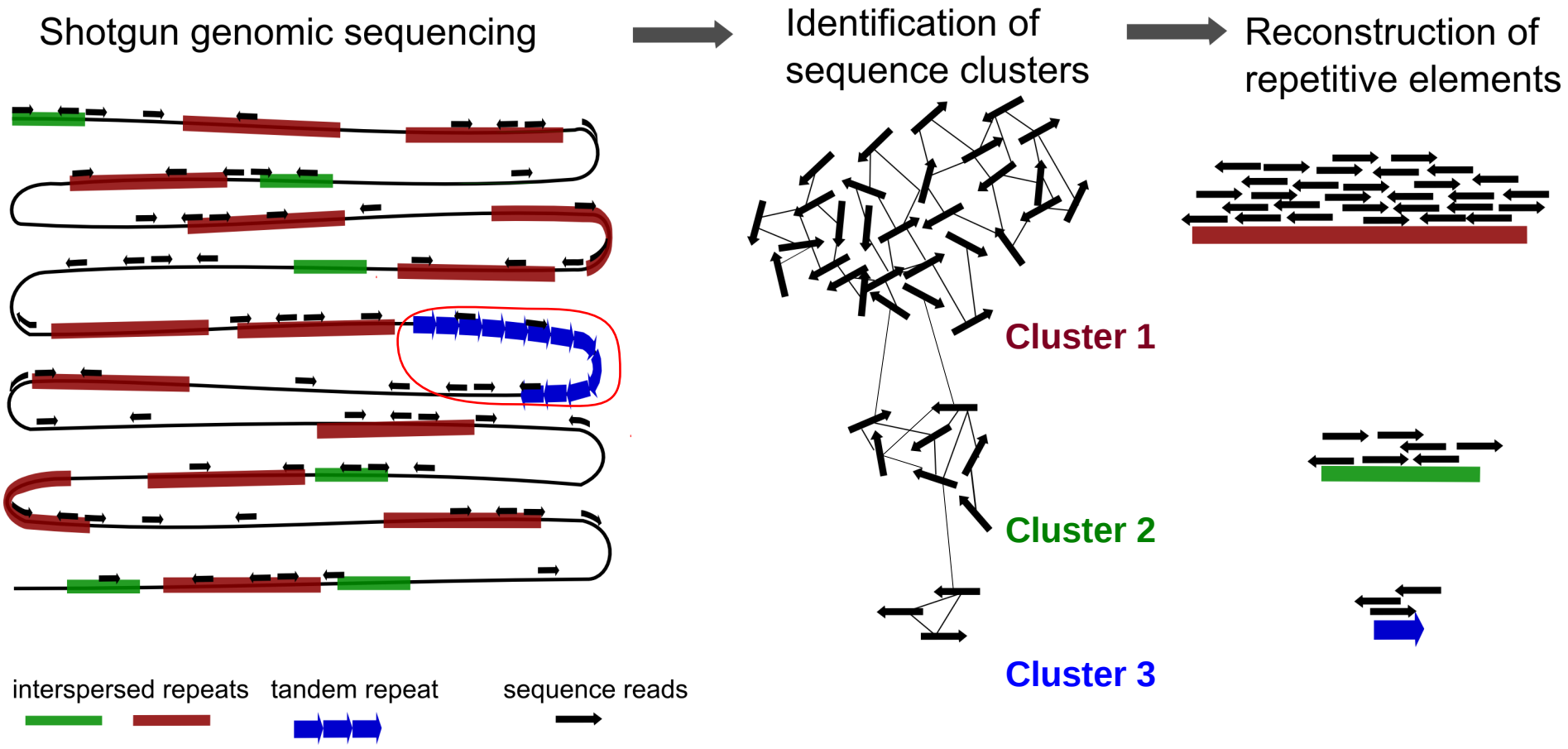
From Current Opinion in Plant Biology ,Volume 8, Issue 2, April 2005, Pages 148–154

# Tandem Repeat Analyzer - TAREAN



# Tandem Repeat Analyzer - TAREAN

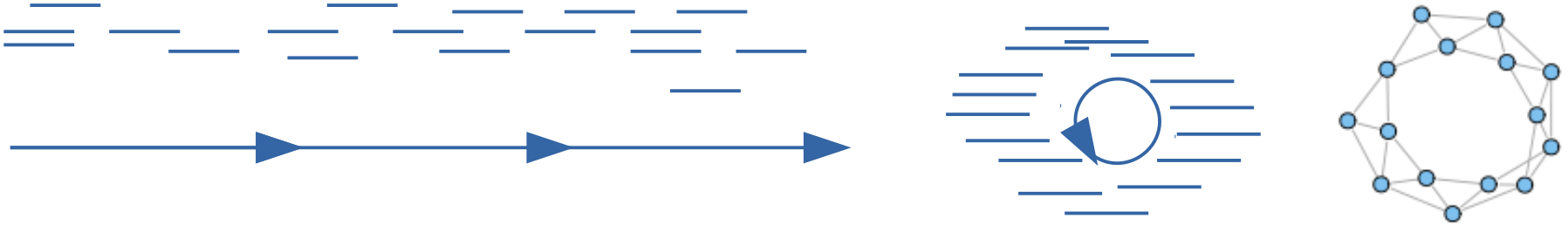
## Principle of tandem repeat identification



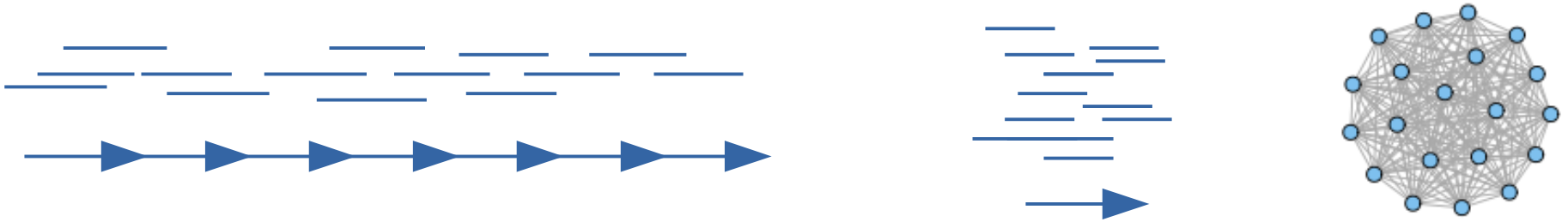
# Tandem Repeat Analyzer - TAREAN

## Principle of tandem repeat identification

Read length << monomer



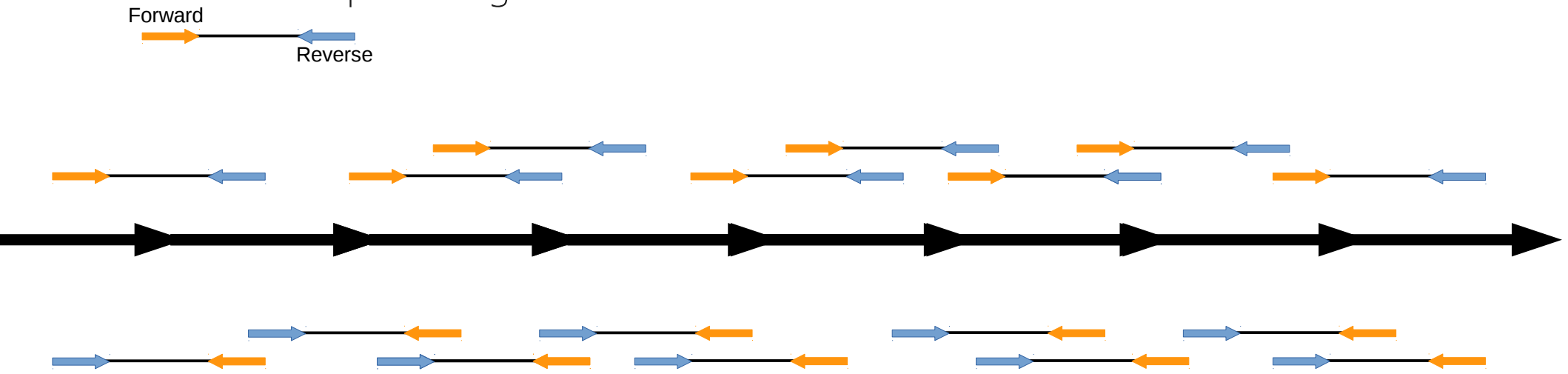
Read length >= monomer



# Tandem Repeat Analyzer - TAREAN

## Principle

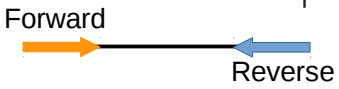
### Paired-End Sequencing

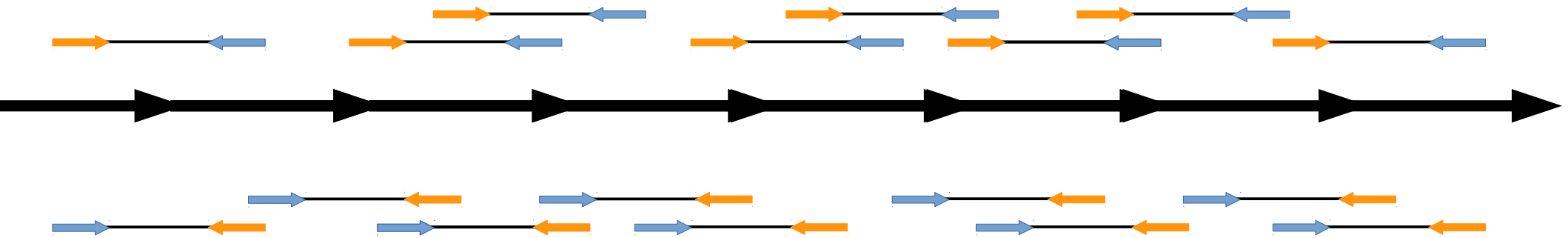


# Tandem Repeat Analyzer - TAREAN

## Principle

### Paired-End Sequencing



### All-to-all comparison - blast

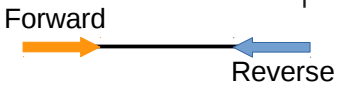
| query | subject | pid | e.value | strand |
|---|---|---|---|---|
| 66706r | 130626f | 93.22 | 3e-20 | + |
| 65114r | 95482r | 95.35 | 4e-16 | - |
| 32080f | 22009f | 93.65 | 3e-23 | + |
| 85661f | 22009f | 97.14 | 9e-14 | - |
| 66706r | 9071r | 95.74 | 2e-18 | + |
| 130626f | 9071r | 95.83 | 5e-43 | + |
| 66706r | 147917r | 95.74 | 2e-18 | - |

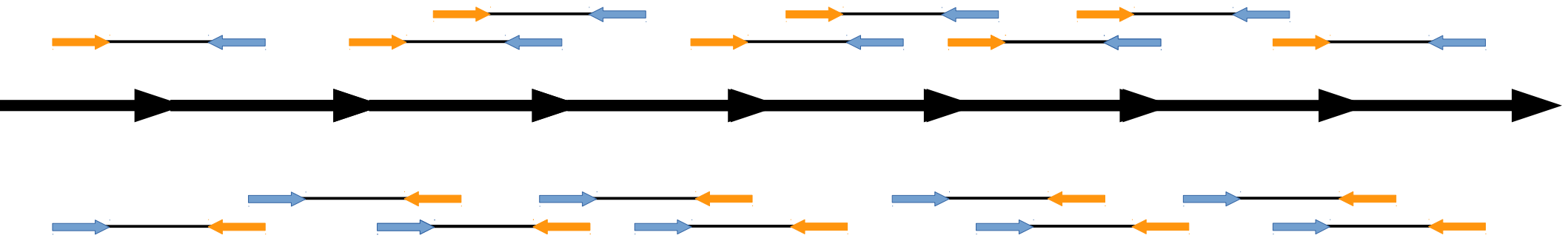
...

# Tandem Repeat Analyzer - TAREAN

## Principle

### Paired-End Sequencing



### All-to-all comparison - blast

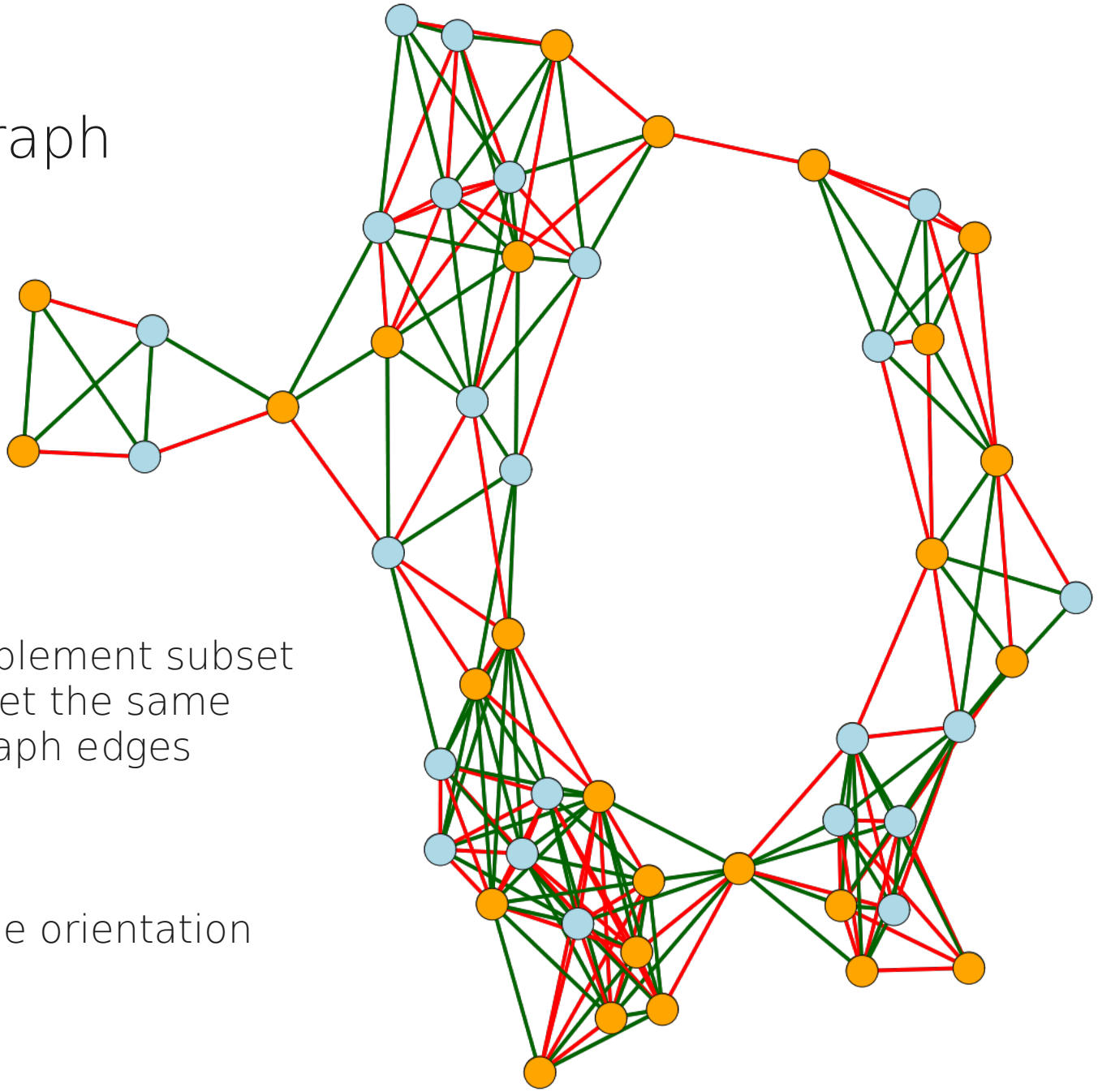
| query | subject | pid | e.value | strand |
|---|---|---|---|---|
| 66706f | 130626r | 93.22 | 3e-20 | + |
| 65114f | 95482f | 95.35 | 4e-16 | - |
| 32080r | 22009r | 93.65 | 3e-23 | + |
| 85661r | 22009r | 97.14 | 9e-14 | - |
| 66706f | 9071f | 95.74 | 2e-18 | + |
| 130626r | 9071f | 95.83 | 5e-43 | + |
| 66706f | 147917f | 95.74 | 2e-18 | - |

...

# Tandem Repeat Analyzer - TAREAN

Principle

Signed graph



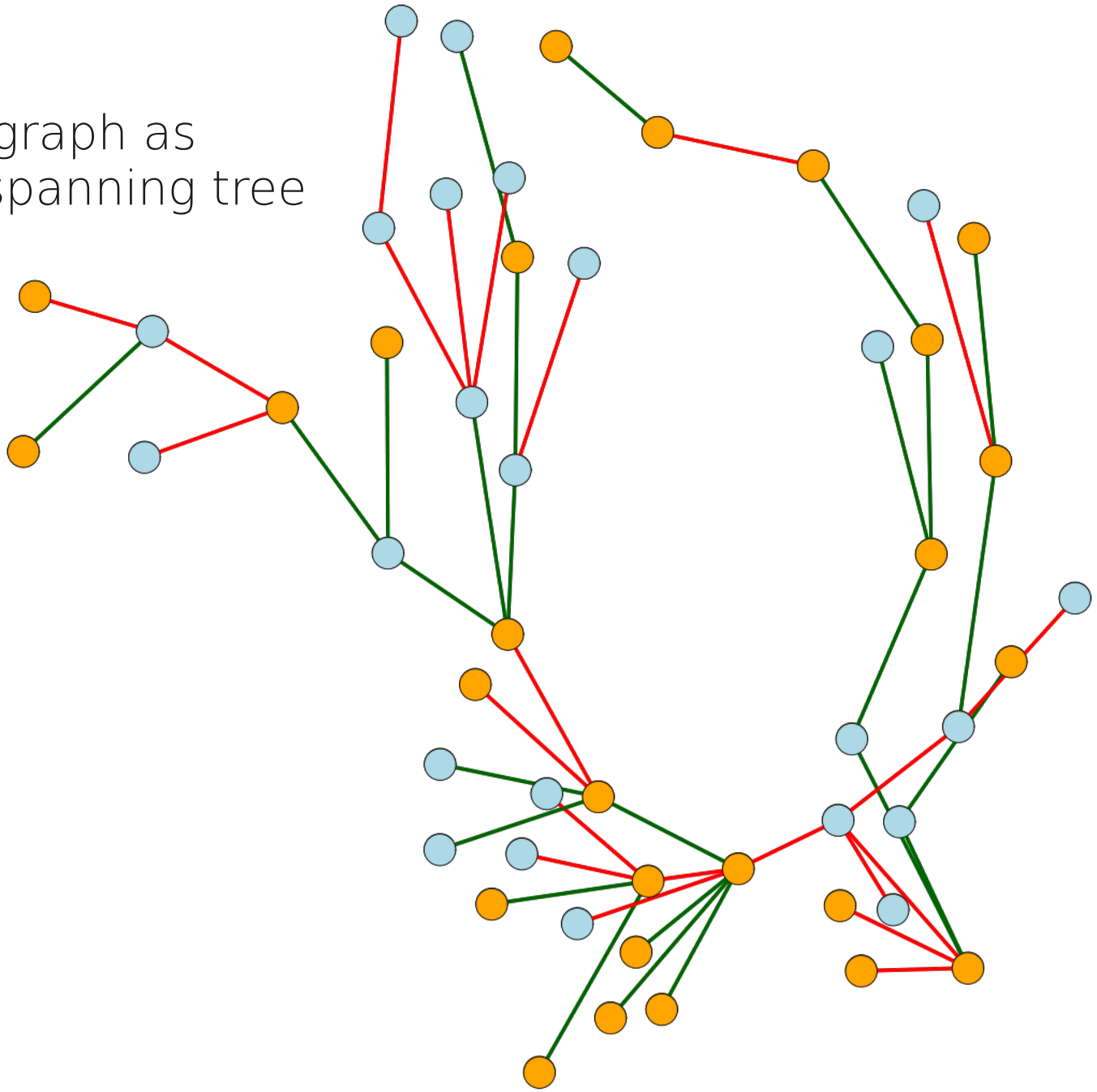
Reverse complement subset of reads to get the same sign on all graph edges

All reads same orientation

# Tandem Repeat Analyzer - TAREAN

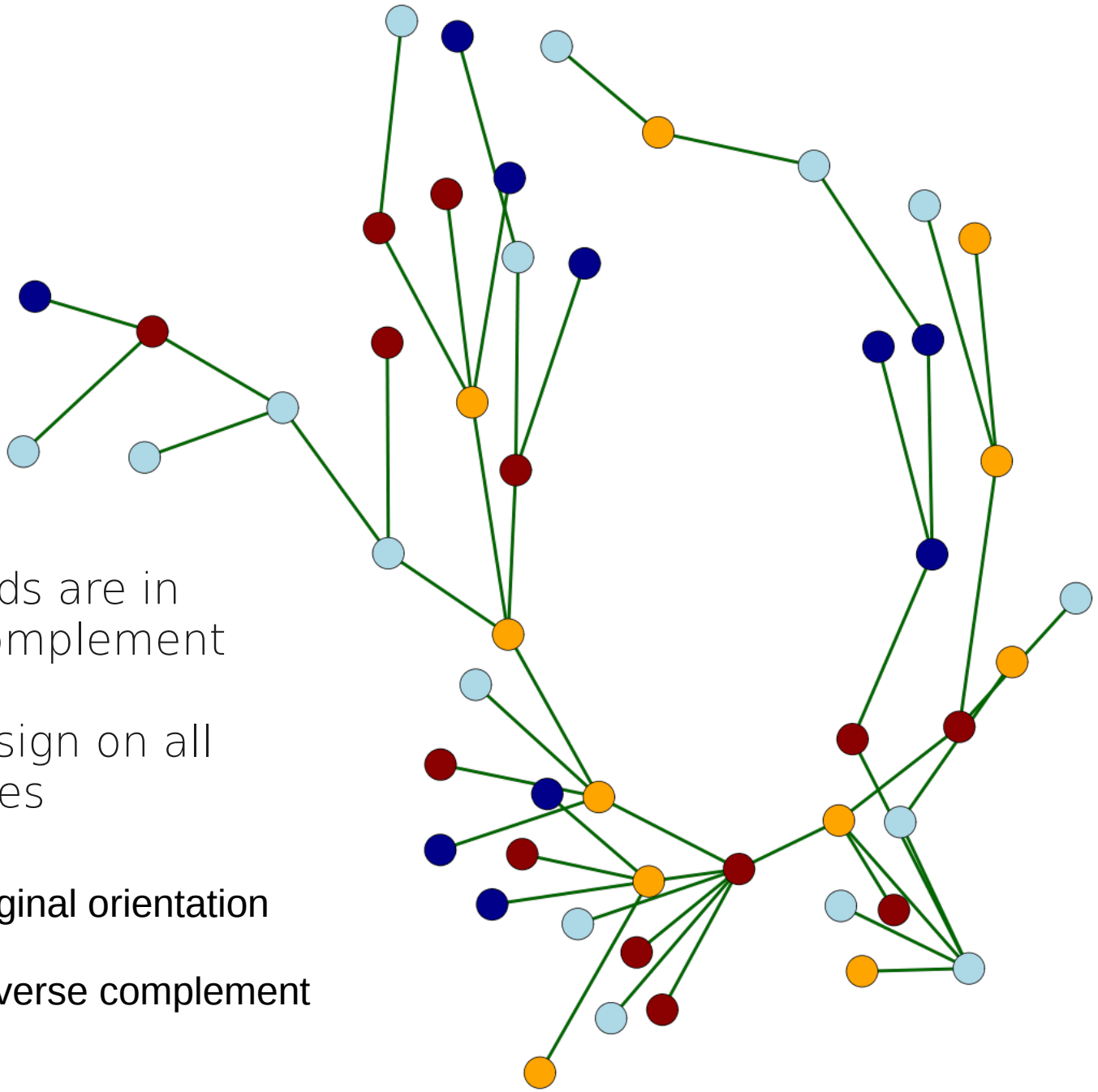
Principle

Simplified graph as
minimum spanning tree



# Tandem Repeat Analyzer - TAREAN

## Principle



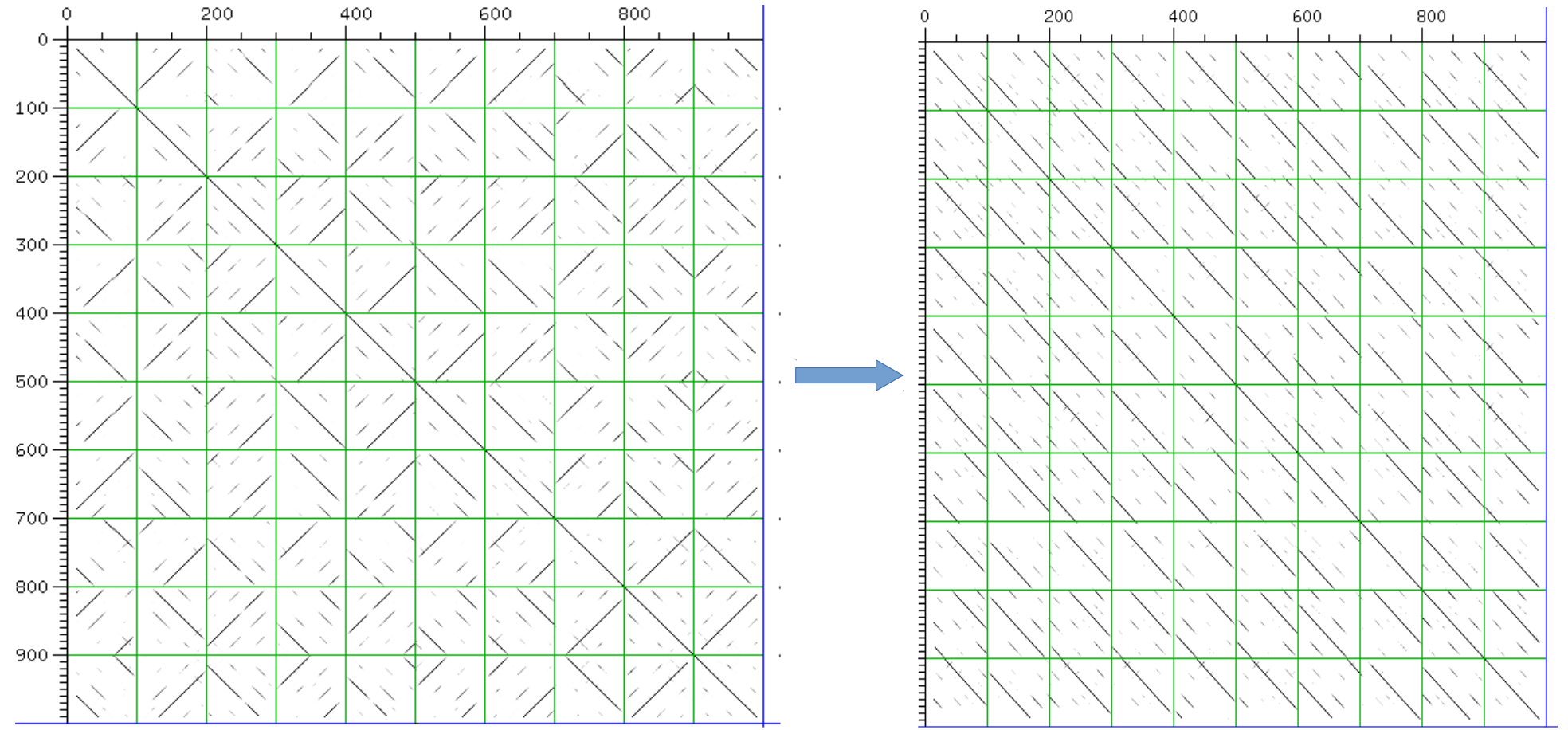
- ~50% reads are in reverse complement
- the same sign on all graph edges

Original orientation

Reverse complement

# Tandem Repeat Analyzer - TAREAN

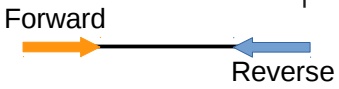
## Principle

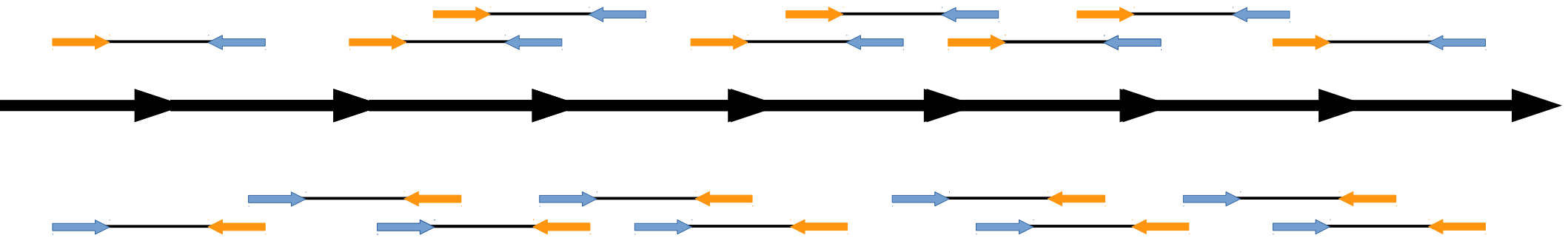


# Tandem Repeat Analyzer - TAREAN

## Principle

### Paired-End Sequencing



### All-to-all comparison - blast

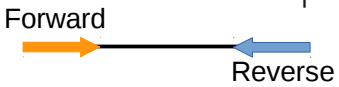
| query | subject | pid | e.value | strand |
|---|---|---|---|---|
| 66706f | 130626r | 93.22 | 3e-20 | + |
| 65114f | 95482f | 95.35 | 4e-16 | - |
| 32080r | 22009r | 93.65 | 3e-23 | + |
| 85661r | 22009r | 97.14 | 9e-14 | - |
| 66706f | 9071f | 95.74 | 2e-18 | + |
| 130626r | 9071f | 95.83 | 5e-43 | + |
| 66706f | 147917f | 95.74 | 2e-18 | - |

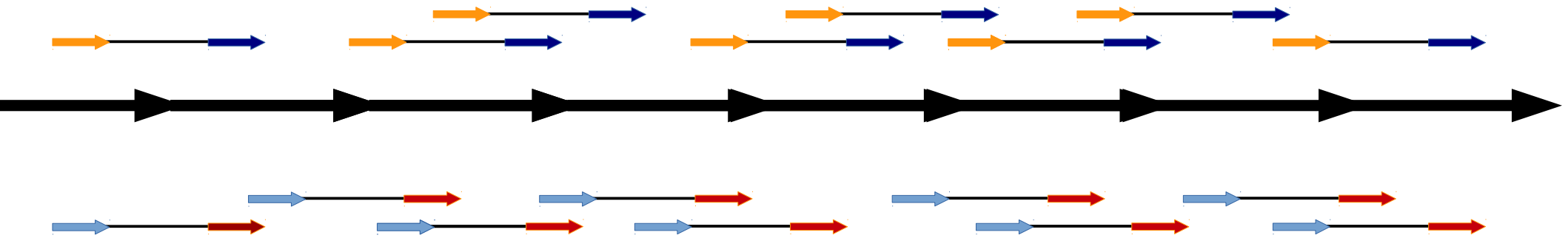
...

# Tandem Repeat Analyzer - TAREAN

## Principle

### Paired-End Sequencing



### All-to-all comparison - blast

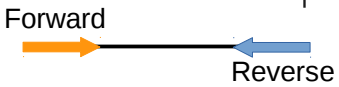
| query | subject | pid | e.value | strand |
|---|---|---|---|---|
| 66706f | 130626r | 93.22 | 3e-20 | + |
| 65114f | 95482f | 95.35 | 4e-16 | - |
| 32080r | 22009r | 93.65 | 3e-23 | + |
| 85661r | 22009r | 97.14 | 9e-14 | - |
| 66706f | 9071f | 95.74 | 2e-18 | + |
| 130626r | 9071f | 95.83 | 5e-43 | + |
| 66706f | 147917f | 95.74 | 2e-18 | - |

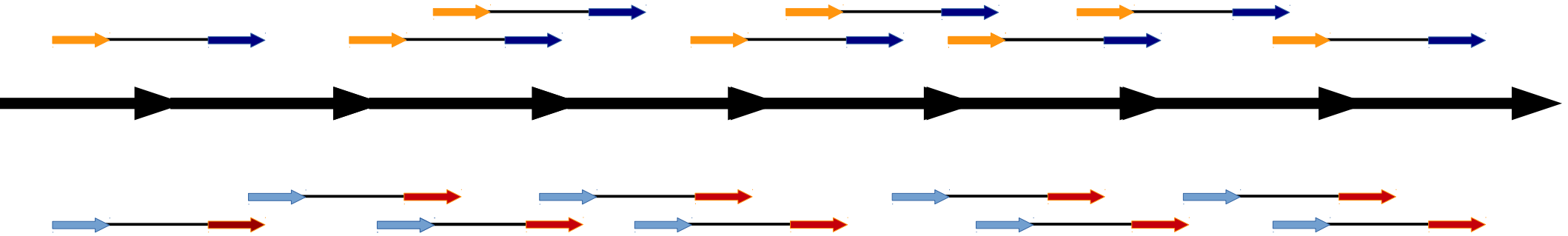
...

# Tandem Repeat Analyzer - TAREAN

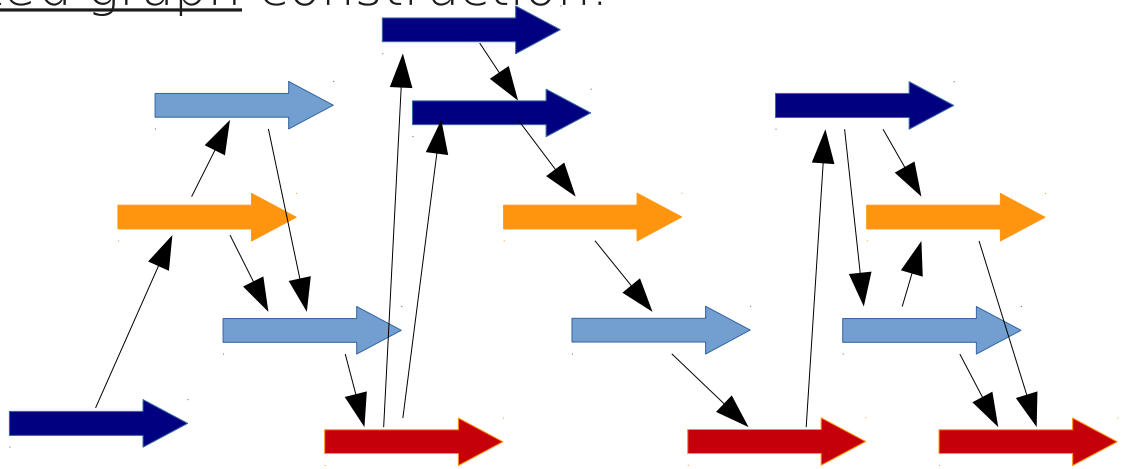
## Principle

### Paired-End Sequencing



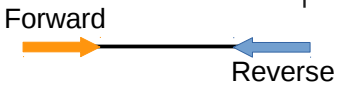
All-to-all comparison – blast – position of alignment and same orientation enable directed graph construction:

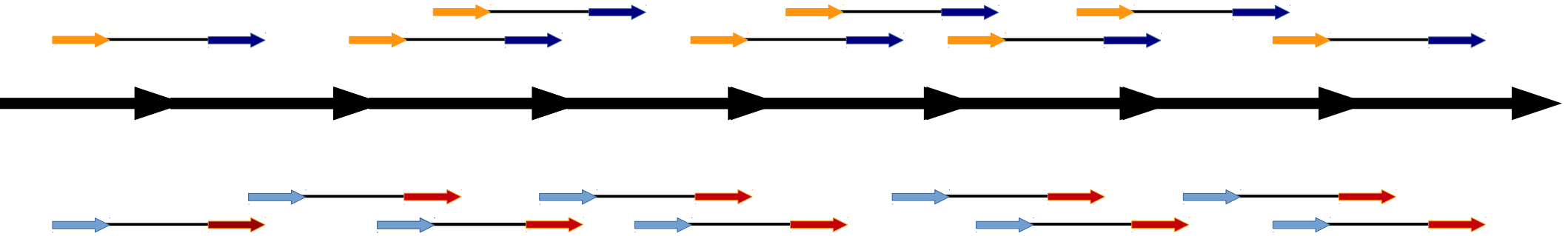


# Tandem Repeat Analyzer - TAREAN

## Principle

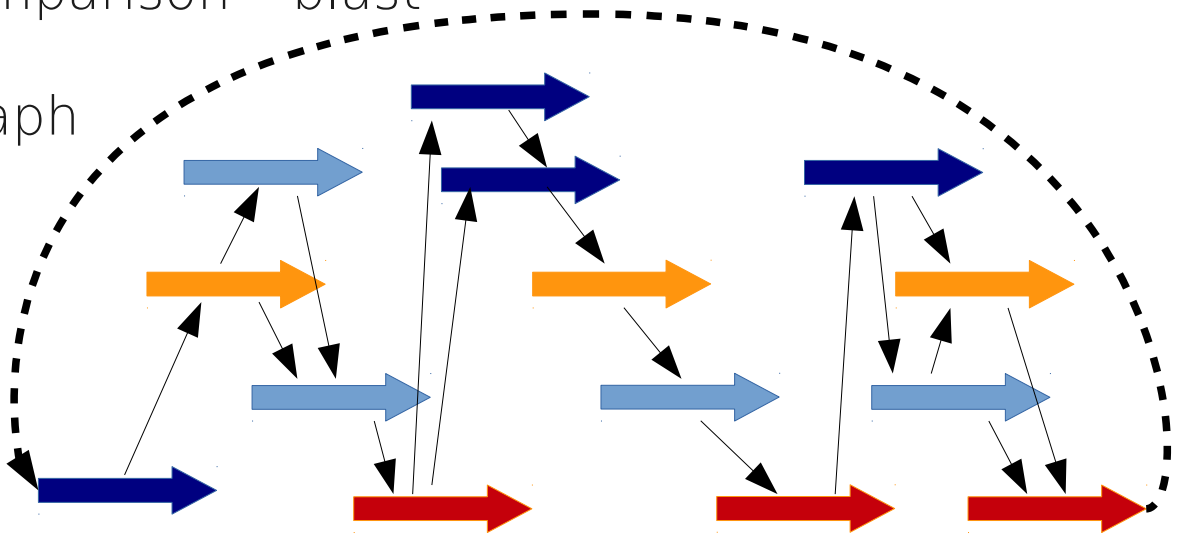
### Paired-End Sequencing



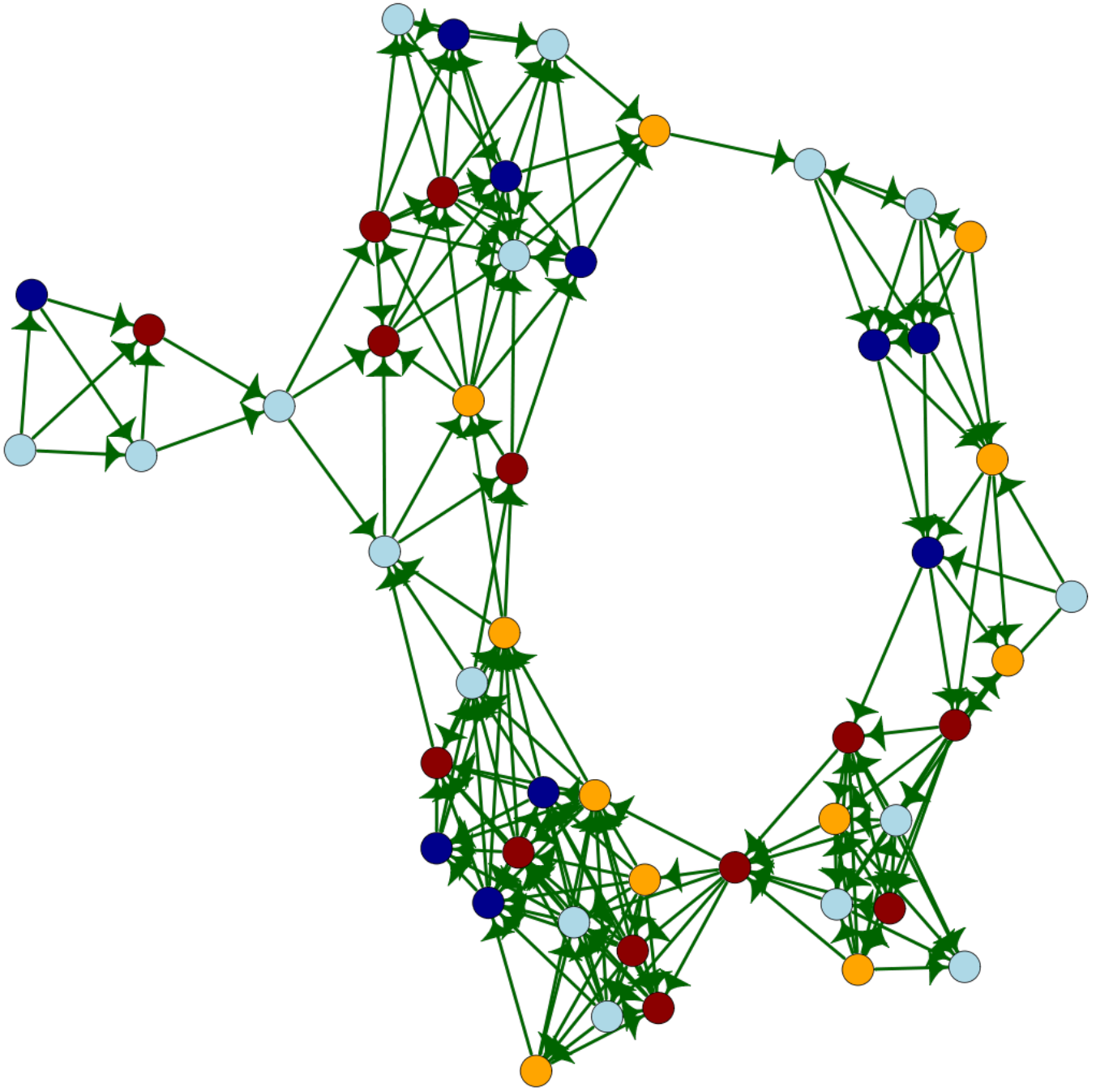
### All-to-all comparison – blast

### Directed graph



# Tandem Repeat Analyzer - TAREAN

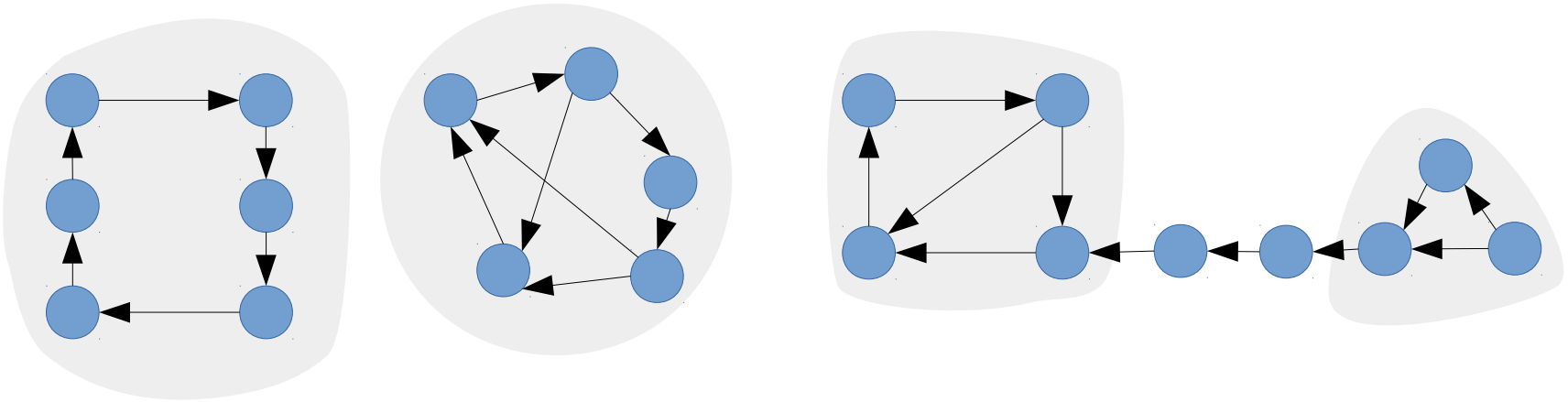
Principle

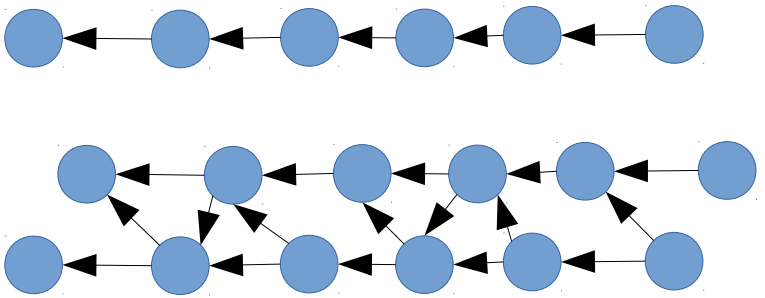


# Tandem Repeat Analyzer - TAREAN

## Principle

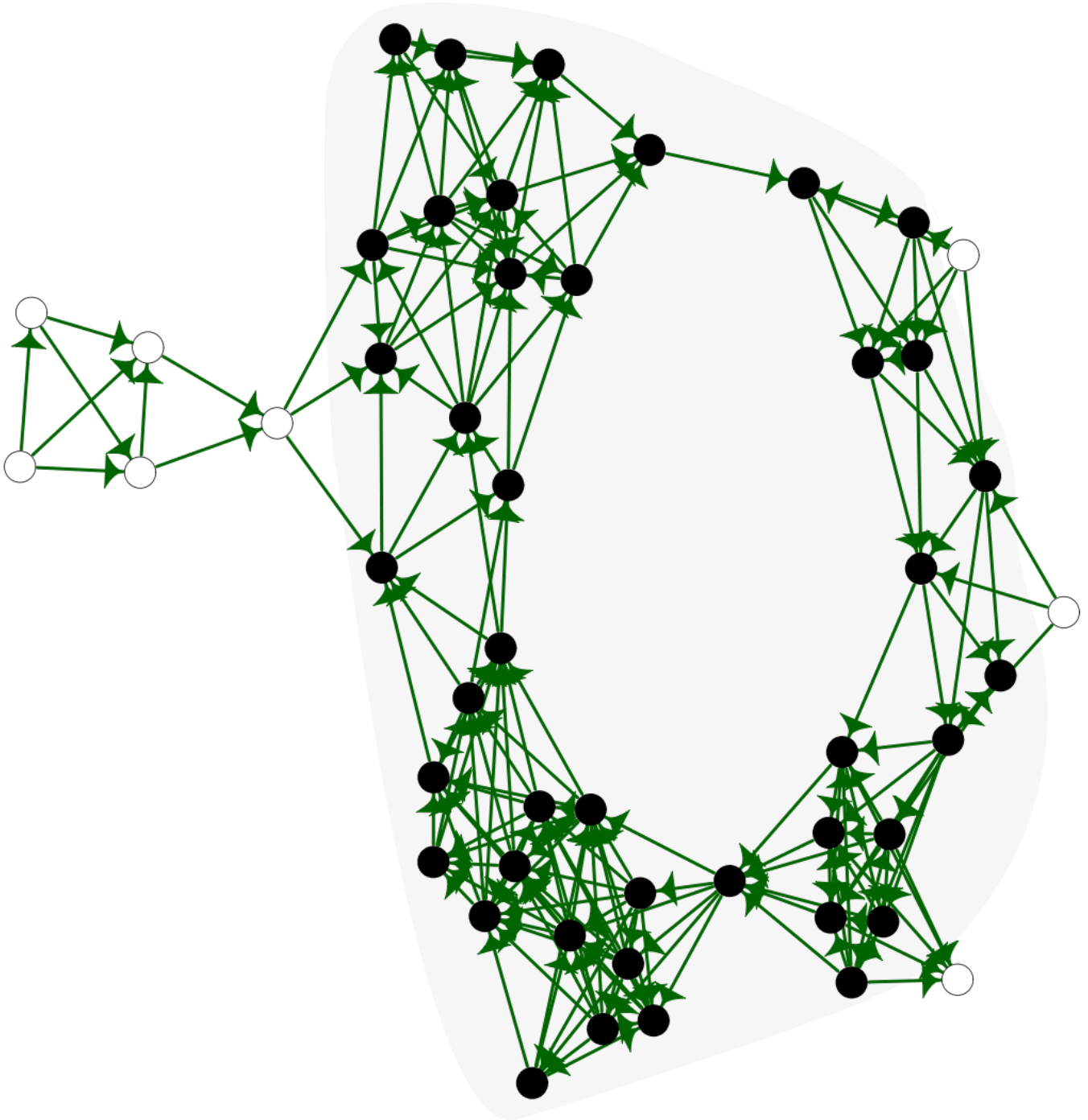
A directed graph is called **strongly connected** if there is a path in **each direction** between each pair of vertices of the graph.





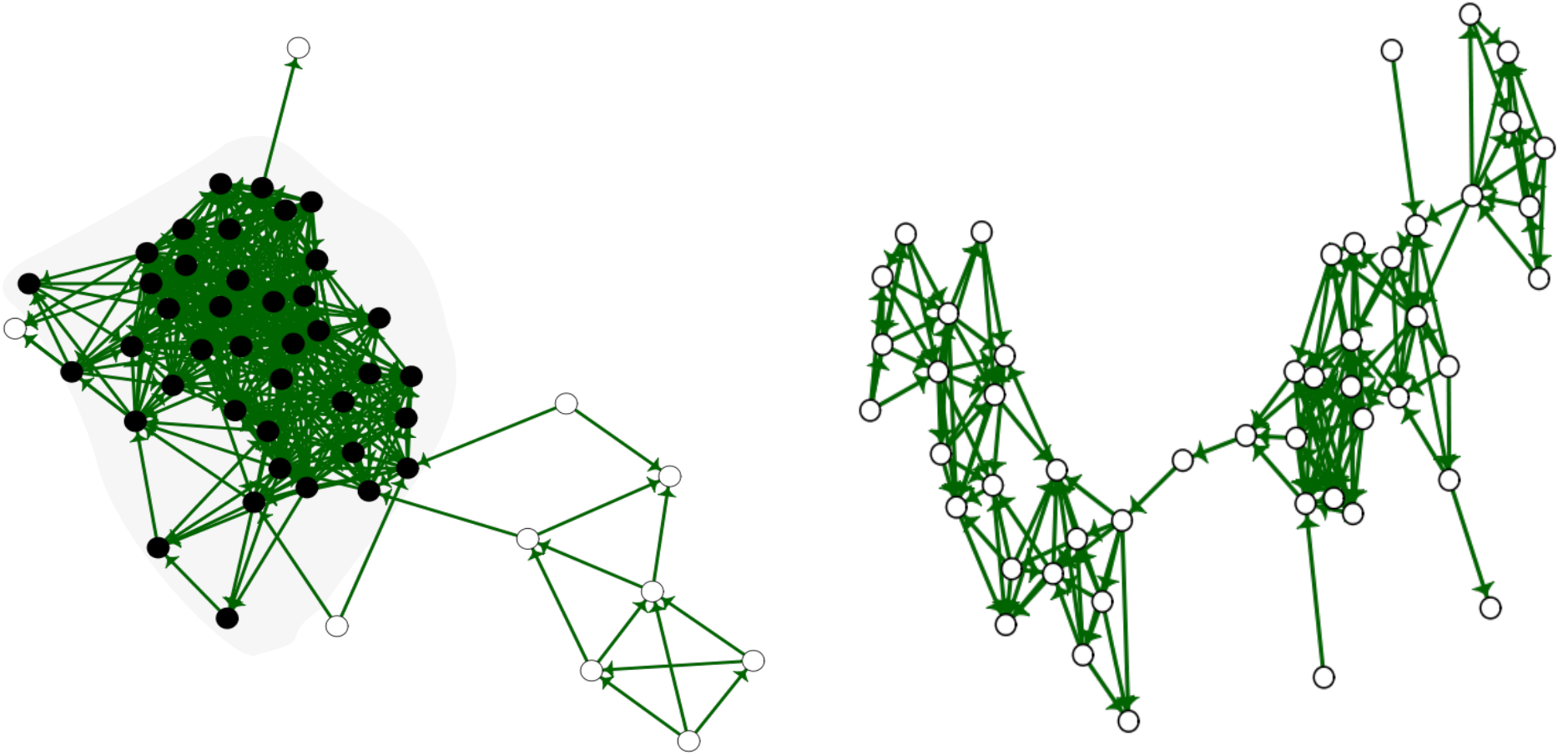
# Tandem Repeat Analyzer - TAREAN

Principle



# Tandem Repeat Analyzer - TAREAN

## Principle

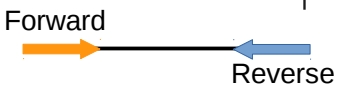


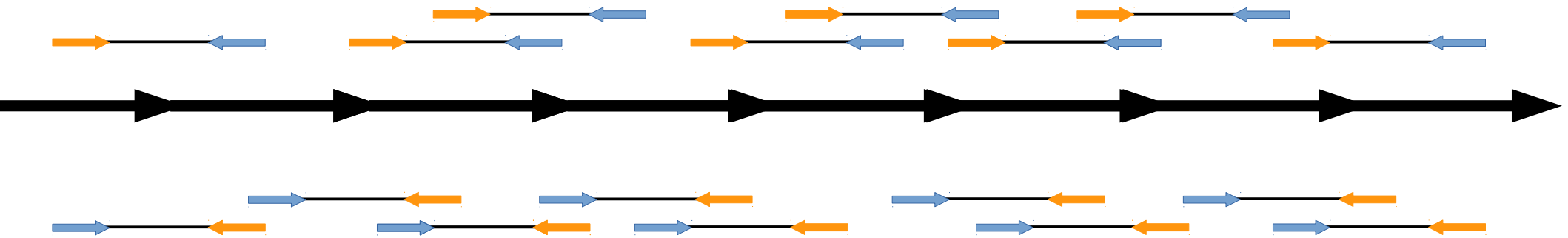
$$\textbf{Loop index} = \frac{\text{size of the largest strongly connected components}}{\text{Total graph size}}$$

# Tandem Repeat Analyzer - TAREAN

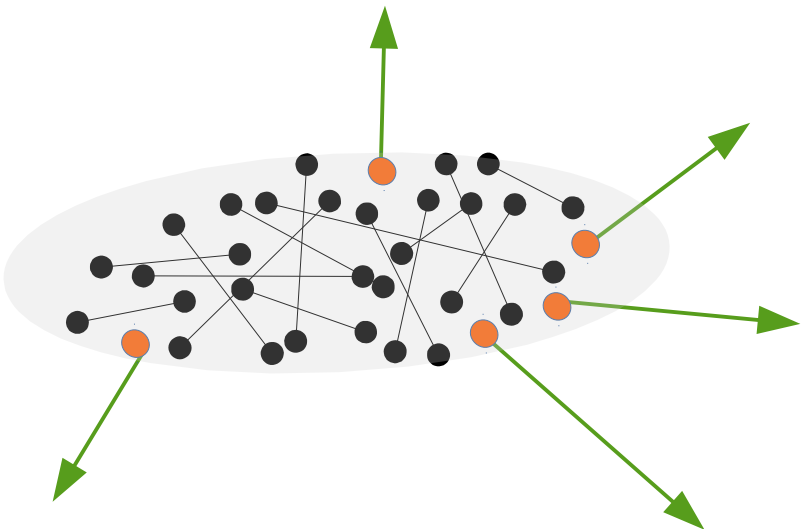
## Principle

### Paired-End Sequencing



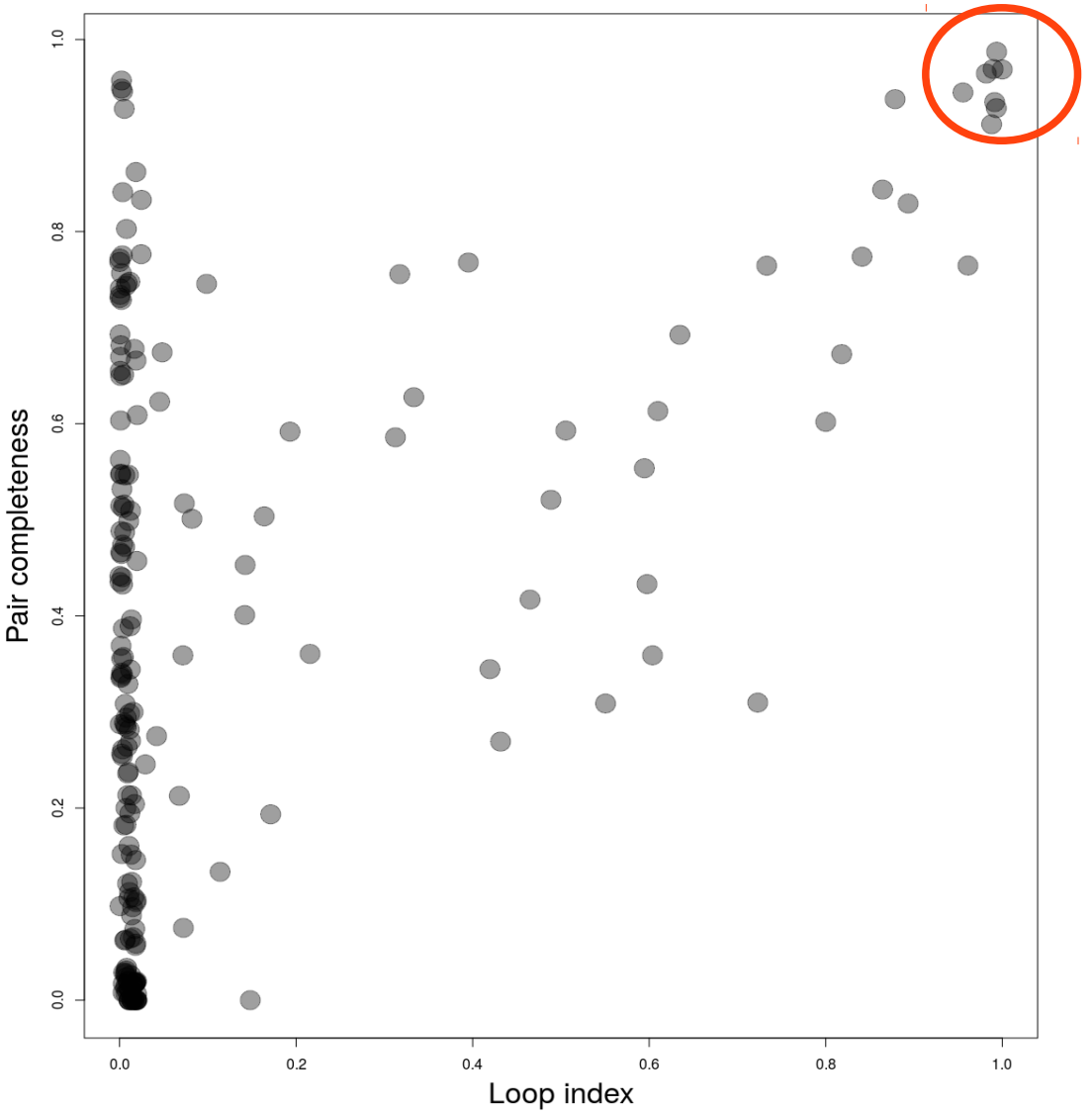
**Pair completeness** = fraction of complete pairs in cluster



# Tandem Repeat Analyzer - TAREAN

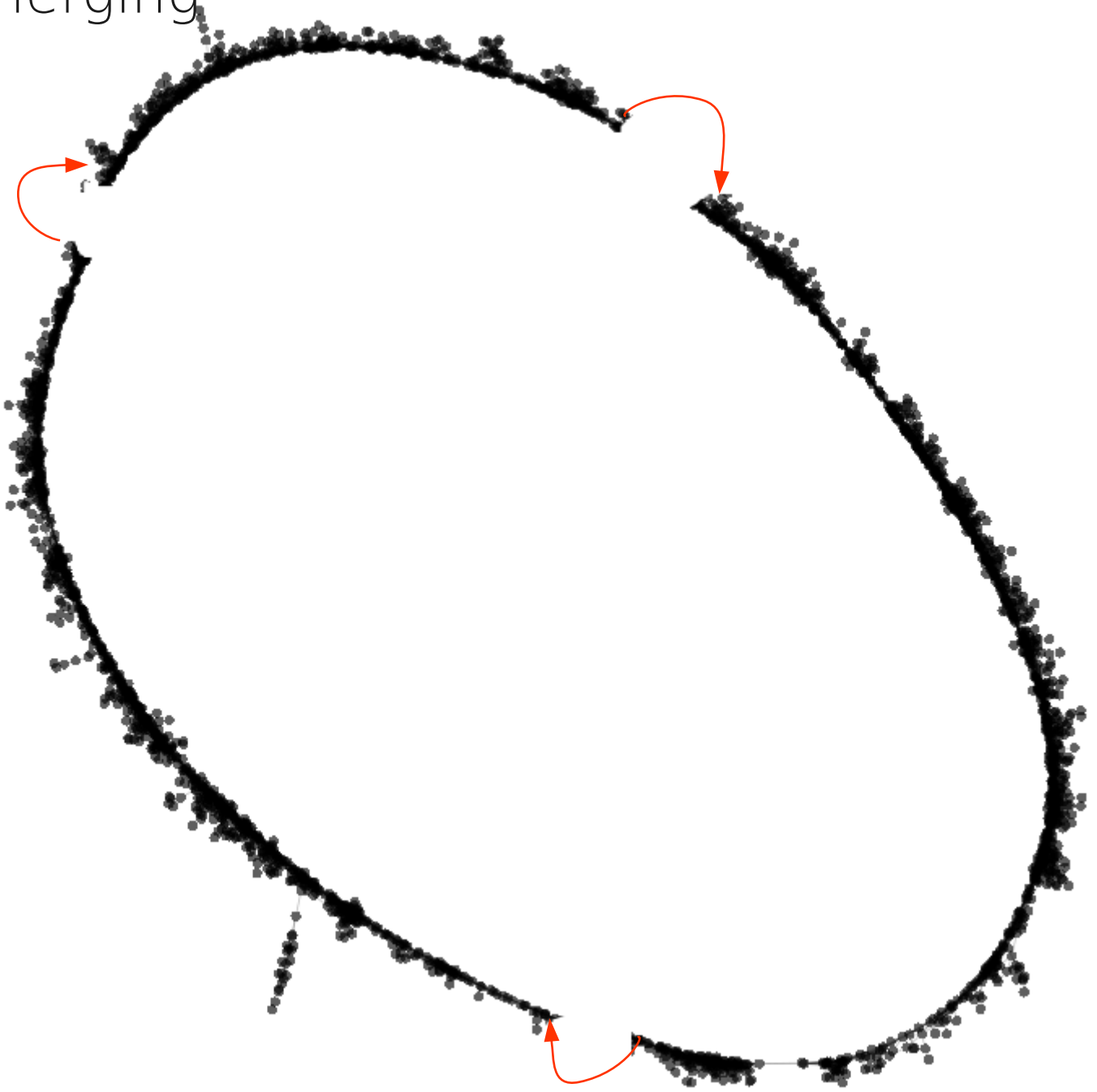
## Principle

Satellites tend to have pair completeness and loop index close to 1



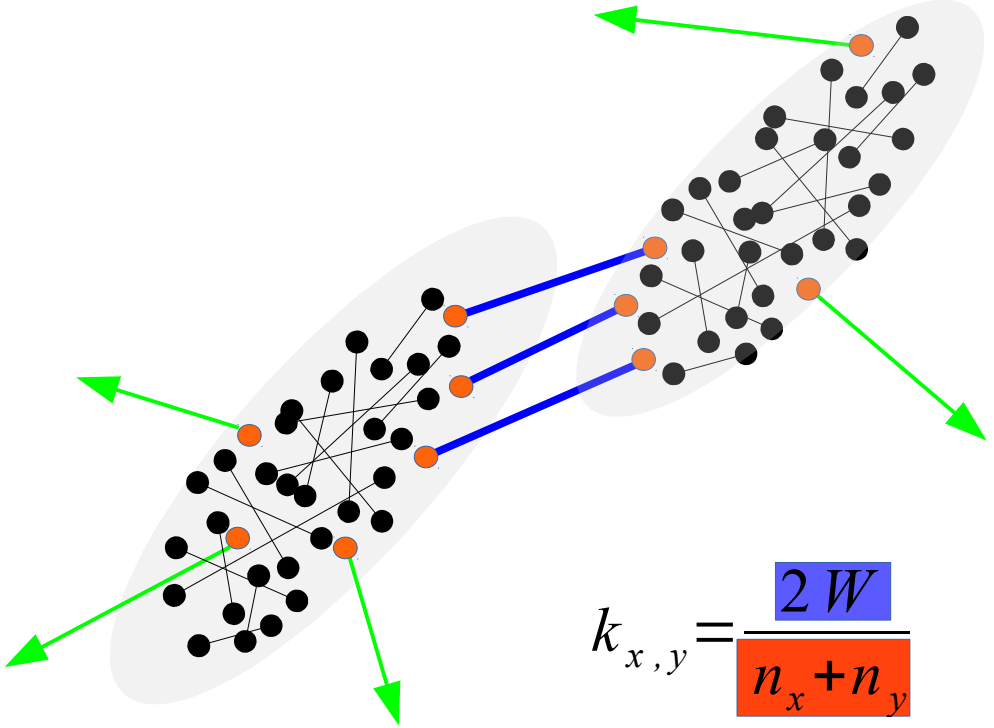
# Tandem Repeat Analyzer - TAREAN

Cluster merging



# Tandem Repeat Analyzer - TAREAN

## Cluster merging



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

**Merging of related clusters**

$W$ 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 $k_{x,y} = 0$

# Tandem Repeat Analyzer - TAREAN

kmer based tandem repeat monomer consensus reconstruction:

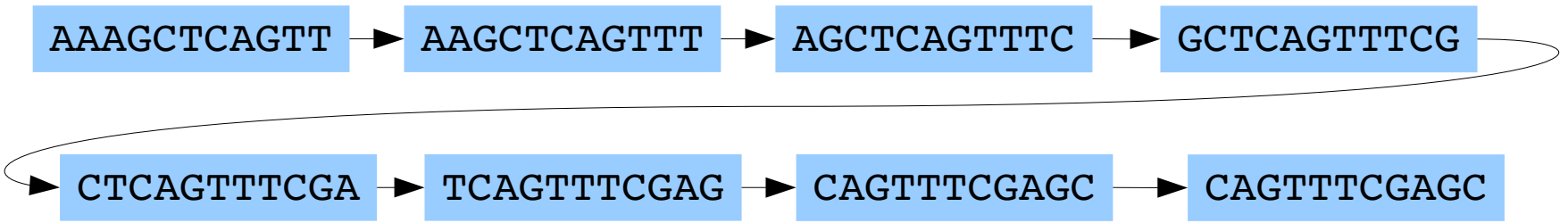
Sequence read:

```
AAAGCTCAGTTTCGAGCCAGAGACCAGAAAGTGTGGGAGCTTACAGCGCAACTTCAGCAAGAGCGGAG
```

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

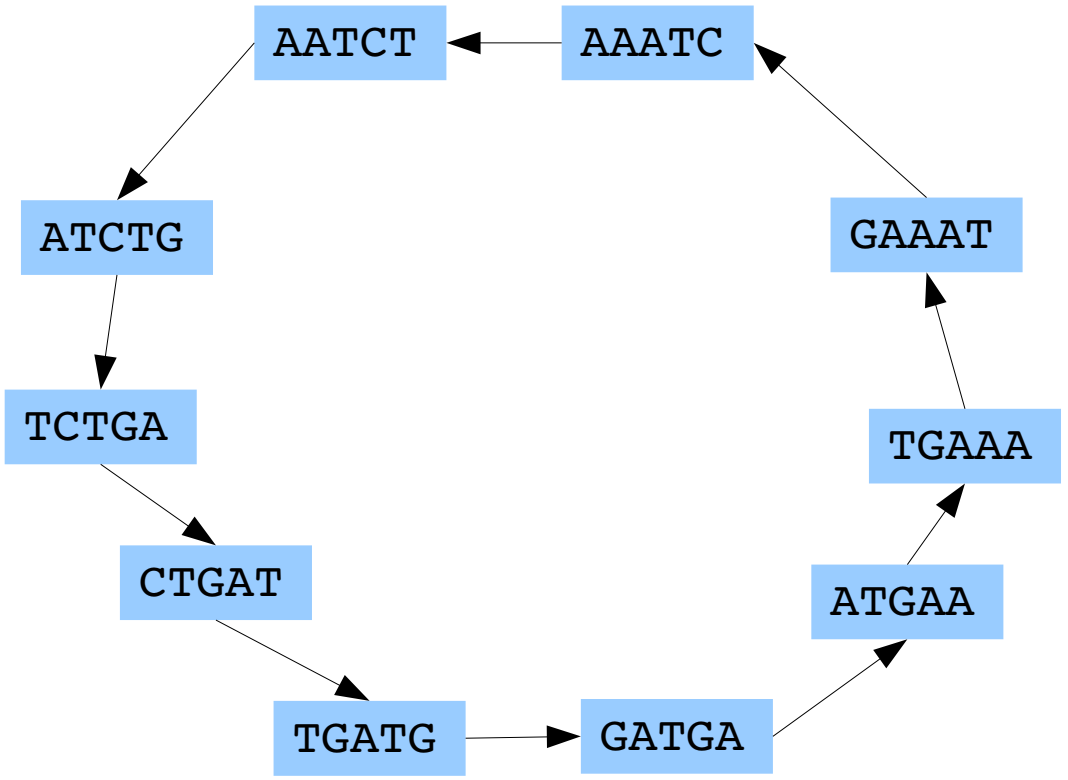
11mers

de Bruijn graph



# Tandem Repeat Analyzer - TAREAN

Consensus reconstruction:

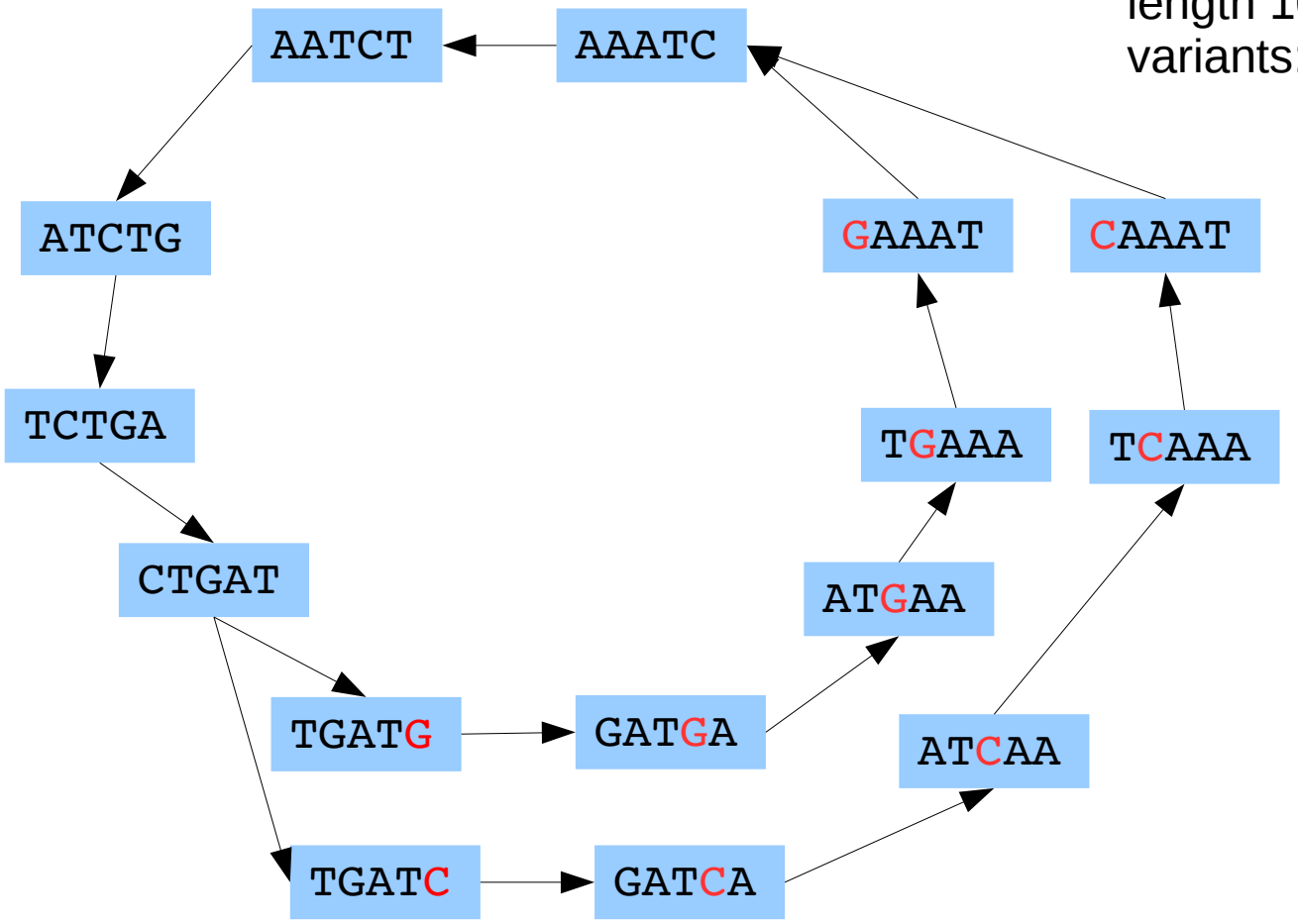



Tandem repeat with monomer length 10nt:

AATCTGATGA

# Tandem Repeat Analyzer - TAREAN

Consensus reconstruction:




Tandem repeat with monomer length 10nt, two sequence variants:

AATCTGAT[C/G]A

# Tandem Repeat Analyzer - TAREAN

## Results: 500k paired reads *P. sativum*

## ~ 0.012x genome coverage

### Run statistics:

Number of input sequences: 4525544

Number of sequences in clusterings: 500000

Paired sequences: TRUE

Number of putative satellites: 8

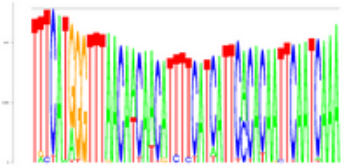
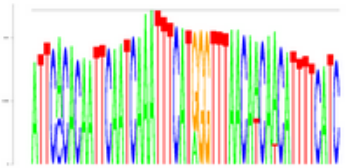
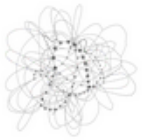
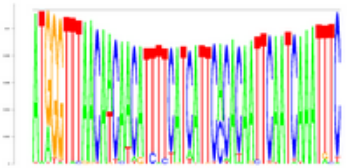
Proportion of sequences in analyzed clusters : 66 %

### Putative satellites:

| | Cluster | Genome Proportion[%] | Size real | Satellite probability | Consensus length | Consensus | Kmer analysis | Graph layout | Loop index | Pair completeness | Kmer coverage | \|V\| | \|E\| | Pbs score | The longest ORF length | Similarity based annotation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | CL14 | 1.300 | 6470 | 0.9439340866644385 | 50 | TTTCATGGGTTTAACAACAACATTTTCATCATTCACACAATTCAATCAAA | report |  | 0.9913446676970634 | 0.9348086124401914 | 0.50 | 6470 | 6256255 | 0 | 82 | 0.02% organelle/plastid% |
| 43 | CL43 | 0.200 | 996 | 0.9441168259917678 | 65 | CAAAACAAGAACACTAATAACACTATCAATCTTCTCATTTCTTGATAACAATATACAAAATTTAG | report |  | 0.9879518072289156 | 0.9117082533589251 | 0.65 | 996 | 406309 | 0 | 43 | |
| 51 | CL51 | 0.150 | 768 | 0.9935709746605441 | 54 | TACGGACGAGTATTGACAAAGAAATGATTAACACGGACGAGTGTTGAAAATCAA | report |  | 0.9895833333333334 | 0.9692307692307692 | 0.51 | 768 | 205652 | 0 | 71 | |
| 58 | CL58 | 0.120 | 622 | 0.9863672152615018 | 164 | ATGAAATGTGGTAAAGTTGTAAAGTAGGTACCACATTTATCCTAAAATATGGTATTTGTATACATGTTTTACCCGCATTTATTAAGAAATGCGGTAAAATTACAAAGAAAGTGCAACATTTGTCTTAAAATGGATGTGTTATAATTGTTTTTCATGCATTTATT | report |  | 0.9935691318327974 | 0.9872204472843451 | 0.58 | 622 | 42830 | 0 | 56 | |
| 84 | CL84 | 0.059 | 297 | 0.9439340866644385 | 960 | GTCCAAACCGTTGTGTTCATCGTCTTCCTCATGATGTTGGTACCAACCGCTACAATCACCGTCATCATCGTCGTCATCATTCTGGCAGCAGTTATAATCATCATCCTCTTCATTATTTCTTGAAACTTCGTCATCATCAACTTCTGCAAAATCAAACAAAATATAAGTTTCACTAAACTAAAAGTGAAAATTTAAAATAAGCTACAAAGATCTATTATTACCCTGACTACTTAGATTTTCCATATGGCTATCTATGACATCCATTTCCCTCTTATAGTCTTCCAGCTTTTGTTCCAATTTGGTTATCTTTTTCTGTATTAGACGTTGTTTGTCTTCCAATTCCTCAATGAGGAGGCTACAAAAAAAATGATAAACAGAAATTAAAAACAAATACAATGTGAGGAAAGATCATAGTAATGTAAGAACTGTAACATTTAGTAATTCCAAAACACCTAGAAAAATCAGTGAAAATGTAGTAATACAATGGCTAAAAATGTAACTATCAACATGAAAAACCTATCAAGAAAAAATGATTAACAGCGTGAGGAATTTGCAAATTCAAGAAAAGGCAATTAACATAAAGATCGTTAAAATCATCATTATCTAACGCAAGAAAACTATAAAAATGCAAGCATGACAAACTTGCAAATTCAAGAAAAAGAAGATCATATAAATAAGCAGTTTCTAAAACAAGAACACGAGAAAATGCACAAAAACATGCATGATTATGGTTATGGTTATCTGCAACAAGAAAGTTTTGTATAGGATGATGGTTATTTGCAGCAAGAAAGTTCTGTATAGGATCATCGTTATCTGCAGCAAGAAGGTTCTCAGTATGAGCAACTTCCTCGGGCTGGTTCTGGTTCAGGTTCTTTGAGGGAGCTTCATTCAATGTCAGGTTCAGTGCATGATCATGGTTATCTGCAGCAAGATAAAGGTTCTGTACGGGAGCATACTATTG | report |  | 0.9932659932659933 | 0.9285714285714286 | 0.40 | 297 | 5761 | 0 | 191 | |
| 89 | CL89 | 0.055 | 277 | 0.9812880760410964 | 280 | AGAAGTATTTAAAAAATCGCAAAAAAAGAGGTGCAACACAAGGACTTCCCAGGAGGTCACCCATCCTAGTACTACTCTCGCCCAAGCACGCTTAACTGCGGAGTTCTGATGGGATCCGGTGCATTAGTGCTGGTATGATCGCACCTGATATCTCATGCGCTTTGATTGTATATATTTTATTTCTTTCACTTATTAACGGCAAAACAAACATAATTAACGCTGAAATGCTTGTCTTCTCCTACTTTTTCCGACTCTTTTACGCCGATTAACAGAAATAAAA | report |  | 0.9819494584837545 | 0.9645390070921985 | 0.74 | 277 | 10488 | 0 | 61 | 56.32% 5S_rDNA/5S_rDNA% |

# Tandem Repeat Analyzer - TAREAN

## Results: 500k paired reads, *P. sativum*

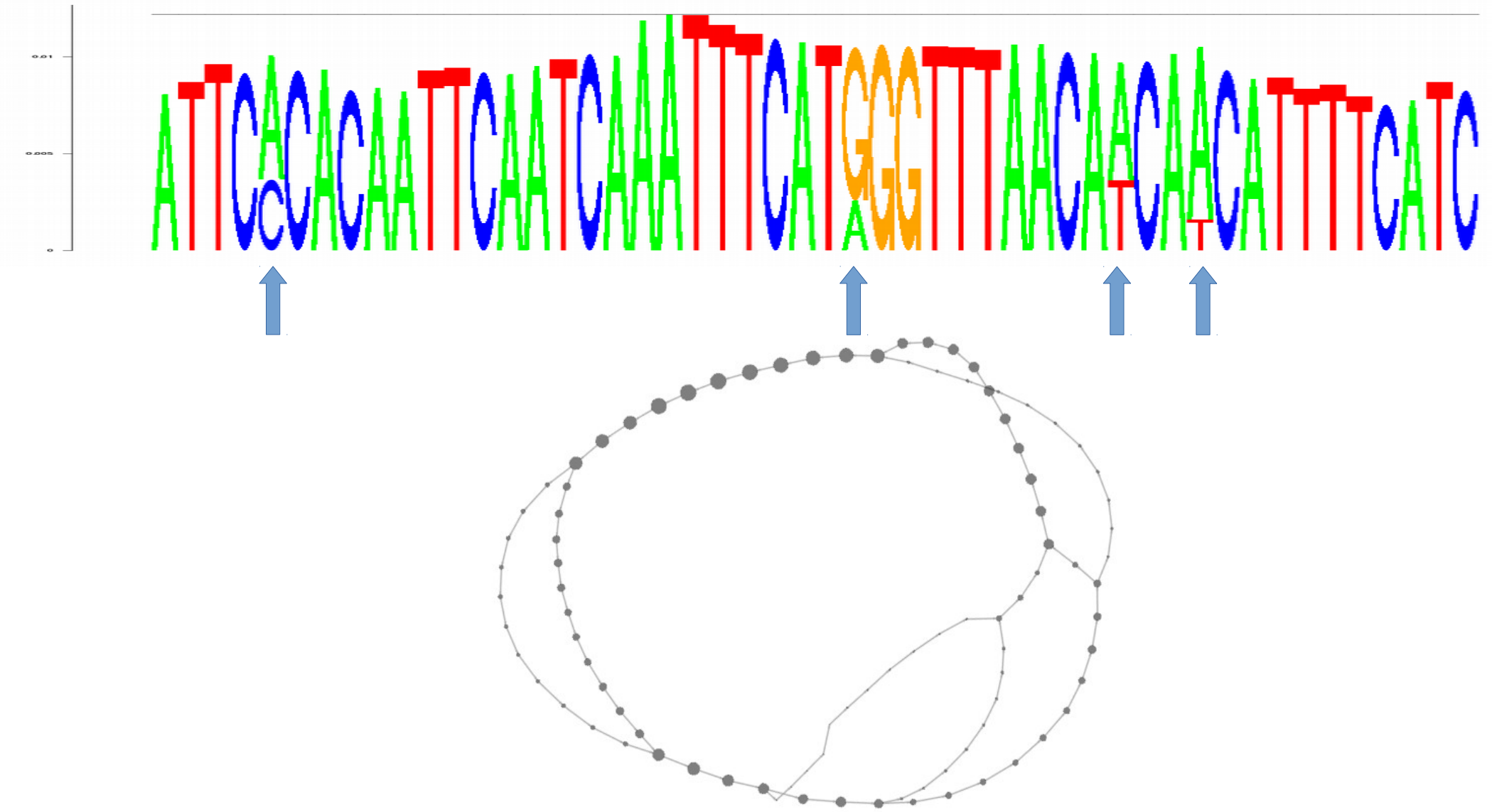
| | kmer | variant | total_score | monomer_length | score_bn | consensus | graph_image | logo_image |
|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 2 | 0.503 | 50 | 8.3e-03 | TTTCATGGGTTTAACAACAACATTTTCATCATTCACACAATTCAATCAAA |  |  |
| 2 | 11 | 1 | 0.486 | 50 | 7.5e-03 | ATTCACACAATTCAATCAAATTTCATGGGTTTAACAACAACATTTTCATC |  |  |
| 3 | 19 | 2 | 0.479 | 50 | 8.4e-03 | ATGGGTTTAACAACAACATTTTCATCATTCACACAATTCAATCAAATTTC |  |  |

PisTRB

# Tandem Repeat Analyzer - TAREAN

## Results: 500k paired reads, *P. sativum*

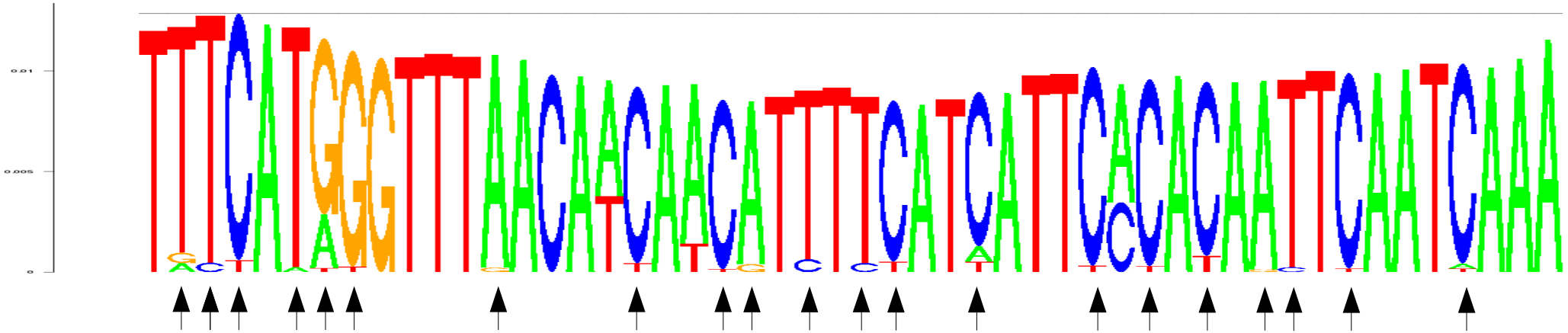
PisTRB – consensus based on 11mers

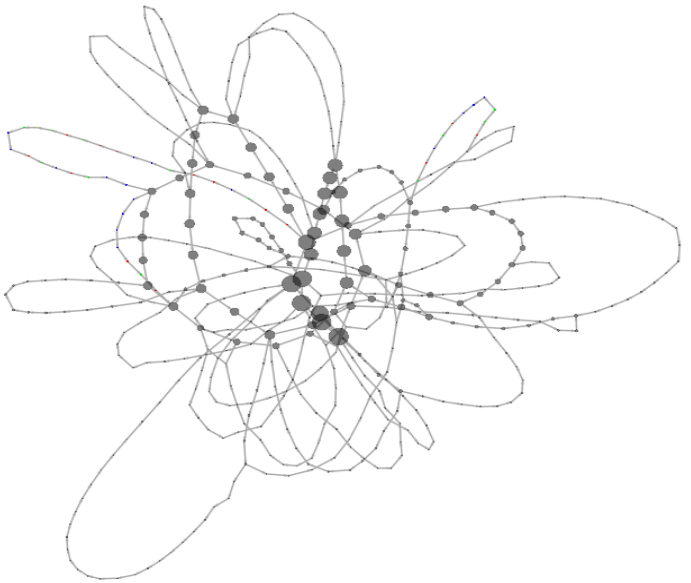


# Tandem Repeat Analyzer - TAREAN

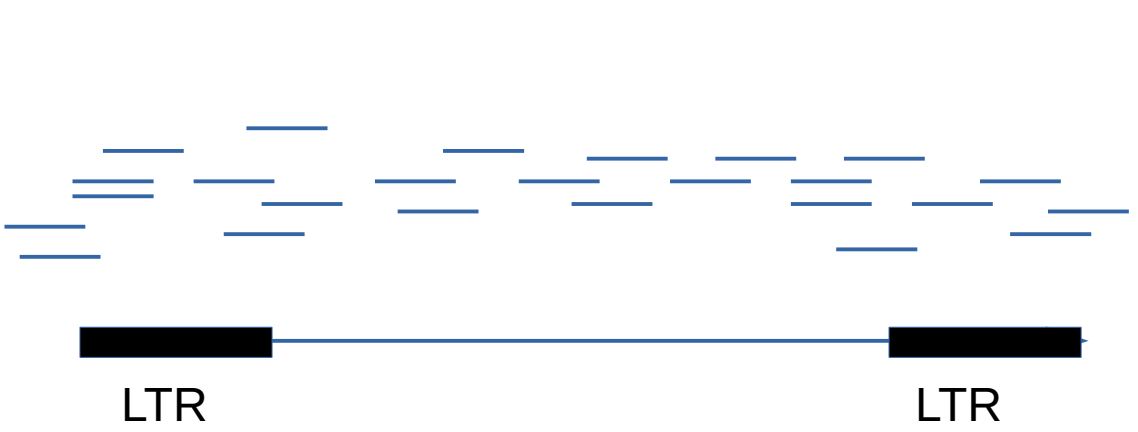
## Results: 500k paired reads, *P. sativum*

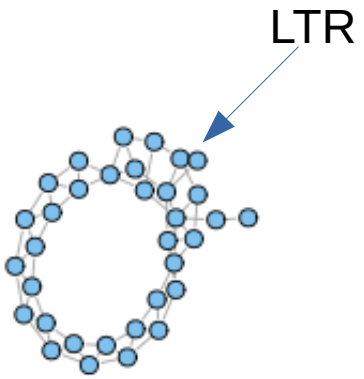
PisTRB – consensus based on 15mers



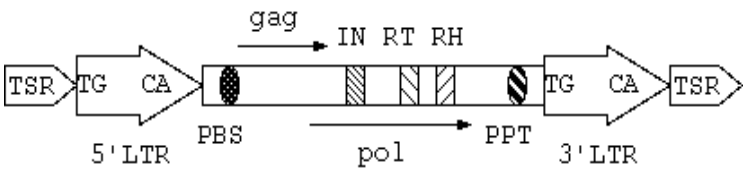


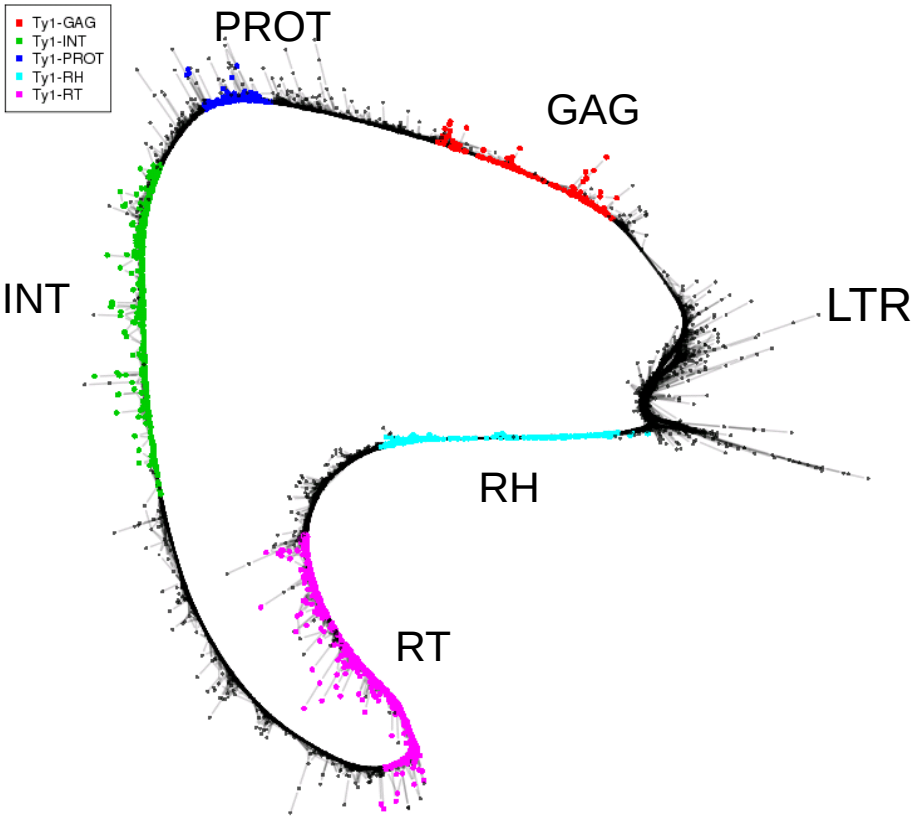
# Tandem Repeat Analyzer - TAREAN




LTR-retrotransposons




- Circular layout, loop index → 1
- frequently high pair completeness

- detection of long ORF
- Indetification of PBS

# Tandem Repeat Analyzer - TAREAN

## Results: 500k paired reads, *P. fulvum*

### Run statistics:

Number of input sequences: 5015824
Number of sequences in clusterings: 500000
Paired sequences: TRUE
Number of putative satellites: 14
Proportion of sequences in analyzed clusters : 71 %

### Putative satellites:

| | Cluster | Genome Proportion[%] | Size real | Satellite probability | Consensus length | Consensus | Kmer analysis | Graph layout | Loop index | Pair completeness | Kmer coverage | \|V\| | \|E\| | Pbs score | The longest ORF length | Similarity based annotation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | CL3 | 4.400 | 21863 | 0.7483552307194397 | 6264 | TTTAGAAAAATCCCATTTTTGGTAAAAAGGTTATTTTCACGTGAAAAGGGAAAAGAGGCAGAAAGGGCAAAAAGCCAGAG<br>AACGAAGGTTGAAGGAAGGAGAAGCTTGGAGCTCGGAAGCTGCCGGATTAACTCAGGTAAGGGGGGTTTATCATCGTTTA<br>ATGGGTATTATGGGTTAACATGTAATGGGTAGTGATAAGCCGTTGAATTGACCCTAATTCGGATTATGAATGCTGAAAAT<br>TGTGATGGATGAGTTATGTTAAAAACTGAAATTGAATCGATAGTAGTGTGAGTCATAATTTGCTGGAAGTGTAGCTTTTT<br>ACGGAATTGGAATCGGAGGTCCGGAAGTCCTCCAACGGCGGAAAATGCGGAGAATTCTGCATTCTGCTTTGTGTTAGCGC<br>AGGAACAGCTTTCTGTCTTGCGTTAACCGGTTAACCCAGGGCGTTAACCGGTTAACACTGTTGAAAATTGGAAAAATTGA<br>TATTTTAATGTTGTGTACCTAATTGGTGGTTGGCCTATATCGGGTGTATGATATAGTAGGGATTATTTCCCGTTGTTTTGA<br>GTAGTATAGGTATTAGTAGAGTGTGCTAATACTGTGATTGATTATGTGGCATGATATGATATGCTGTTGTGATAAAATAT<br>GTTGATGATGTGTGATGATATGCATAATGCTGTGAATGTATGTATTATGTATGCATTGTGAATAGACTGTTTTATGGCTT<br>AGAGTGTGAGCATATGTCTATTCTTGATATTGTTGTTGATGTTGCATTTGCTAGGTGATTAGCTTGCACATTGTGGCCCA | report |  | 0.9568677674610072 | 0.8537391894183484 | 0.4514519480519480 | 21863 | 1787265 | 11.279959999999999 | 1491 | |
| 14 | CL14 | 0.780 | 3882 | 0.9935709746605441 | 7482 | AGCCGTTTCCAGGGTGGGCAGGCTGTTAAACAGAAAAGATAACTCTTTCCGGGGCCCCCGCCAACGTCTCCGGACTCCCT<br>AACGTTGCCGTCAGCCACCACGTCCCGGTTCAGGAATTTTAACCCGATTCCCTTTCGGTGTACGCGCTTAGAGCGCTATC<br>AGACGGGCTTCCCCCGTCCCTTAGGATCGACTAACCCATGTGCAAGTGCCGTTCACATGGAACCTTTCCCCTCTTCGGCC<br>TTCAAAGTTCTCATTTGAATATTTGCTACTACCACCAAGATCTGCACCGACGACCGCTCCGCCCAGGCTCGCGCCCTGGG<br>TTTTGCAGCGACCGCCGCGCCCTCCTACTCATCAGGGCTTGGCCCTTGCCCCCAACGGCCGGGTATAGGTCGCGCGCTTTA<br>GCGCCATCCATTTTCGGGGCTAGTTGATTCGGCAGGTGAGTTGTTACACACTCCTTAGCGGATTTCGACTTCCATGACCA<br>CCGTCCTGCTGTCTTAATCGACCAACACCCTTTGTGGGGTCTAGGTTAGCGCGCAGTTGGGCACCGTAACCCAGCTTCCG<br>GTTCATCCCGCATCGCCAGTTCTGCTTACCAAAAATGGCCCACTTGGAGCTCTCGATTCCATGGCATGGCTCAACAGAGC<br>AGCCACACCGTCCTACCTATTTAAAGTTTGAGAATAGGTCGAGGGCGTTGCGCCCCCGATGCCTCTAATCATTGGCTTTA<br>CCCGATAGAACTCGCCCTCGGGCTCCAGCTATCCTGAGGGAAACTTCGGAGGGAACCAGCTACTAGACGGTTCGATTAGT<br>CTTTCGCCCCTATACCCAAGTCAGACGAACGATTTGCACGTCAGTATCGCTGCGGGCCTCCACCAGAGTTTCCTCTGGCT<br>TCGCCCCGCTCAGGCATAGTTCACCATCTTTCGGGTCCCGACAGGTATGCTCTCACTCGAACCCTTCACAAAAGATCAGG<br>GTCGGTCGGCGGTGCAACCCACAAGAGGATCCCACCAATCAGCTTCCTTGCGCCTTACGGGTTTACTCGCCCGTTGACTC<br>GCACACATGTCAGACTCCTTGGTCCGTGTTTCAAGACGGGTCGAATGGGGAGCCCACAGGCCGACGCCAGGAGCACGCAA<br>GTGCCGAAGCACGCCGAATCGGCGCGTGCTGCCATCCACAATCGTGATGATGACGTCTCCGCGAGTATTTCAACAACCCA<br>GGCTTGGGCCACCATCACAATCCGCGTCGGTCAATGTCTCGAGTCGATTGGCGGACCGGCACAAACCGTTCCACATCCGA<br>CCGAGACACATCGCCGGCCCCCATCCGCTTCCCTCCCGACAATTTCAAGCACTCTTTGACTCTCTTTTCAAAGTCCTTTT | report |  | 0.9933024214322514 | 0.9715591670898933 | 0.8783249701820133 | 3882 | 185713 | 0.0 | 215 | 32.64%<br>45S_rDNA/25S_rDNA<br>23.96%<br>45S_rDNA/18S_rDNA<br>2.68%<br>45S_rDNA/5.8S_rDN<br>0.13%<br>organelle/mitochondr |

# Tandem Repeat Analyzer - TAREAN

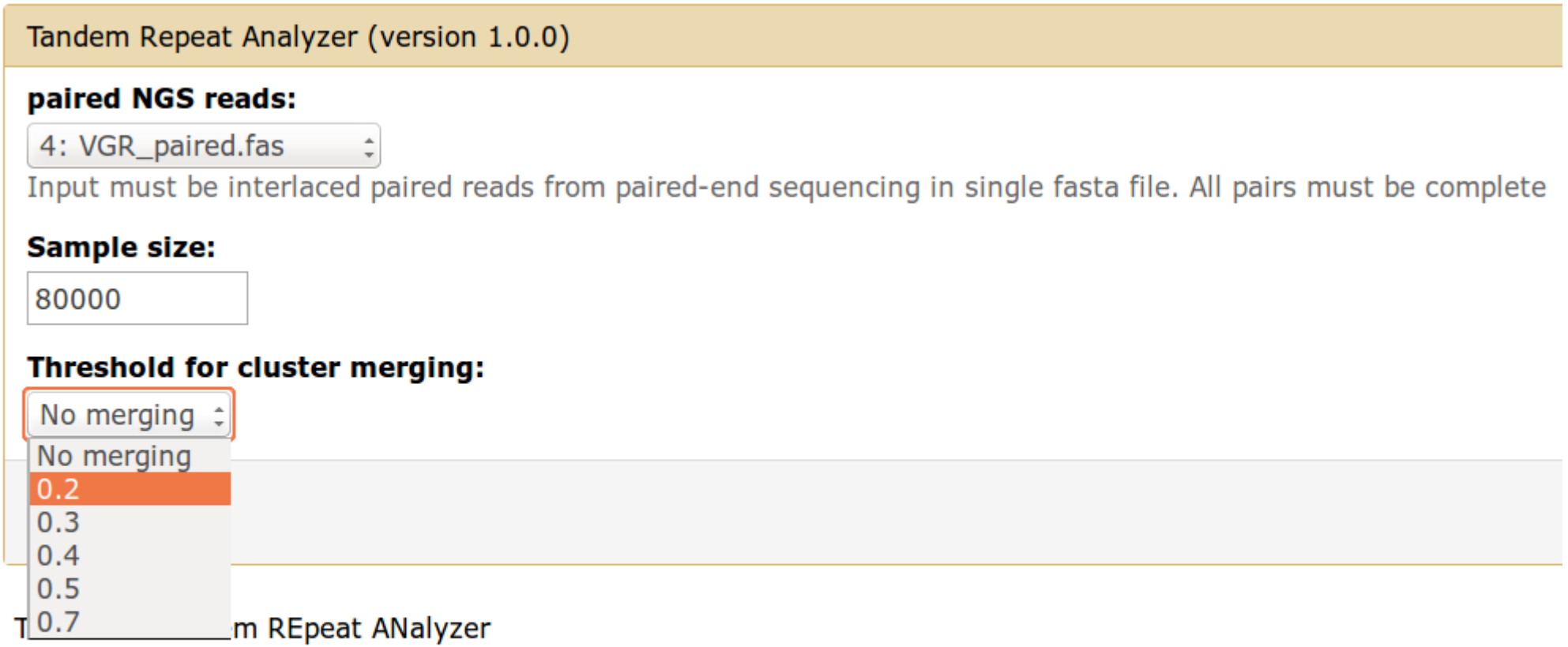
## Conclusions

- Simple repeats(microsatellites) are not necessary detected due to dust filtering in all-to-all megablast step ( e.g. telomeric)
- Clusters derived from some contaminants with circular DNA (plasmids, bacteriophages, viruses) look like tandem repeats → similarity based annotation
- Cluster fragmentation → merging
- Low coverage is sufficient for detection of major satellites
- Validated by FISH experiments/ agreement with previous results

# Tandem Repeat Analyzer - TAREAN

Galaxy implementation:

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



# Tandem Repeat Analyzer - TAREAN

Acknowledgement:

BIOLOGY CENTRE CAS

elixir CZECH REPUBLIC

cerit scientific cloud

Petr Novák
Jiří Macas

Martin Tomáš
Pavel Fibich