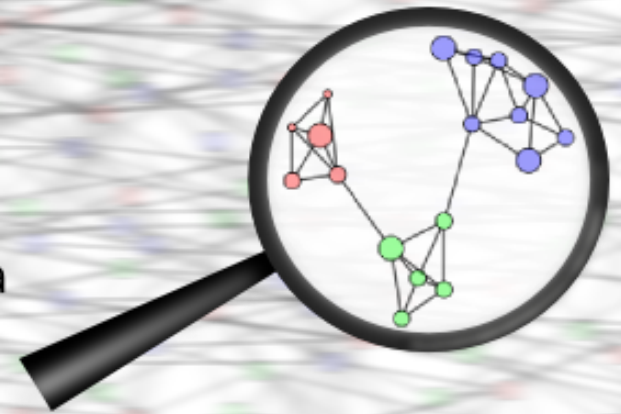


RepeatExplorer 2.0

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



Additional Tools Available on RepeatExplorer Server

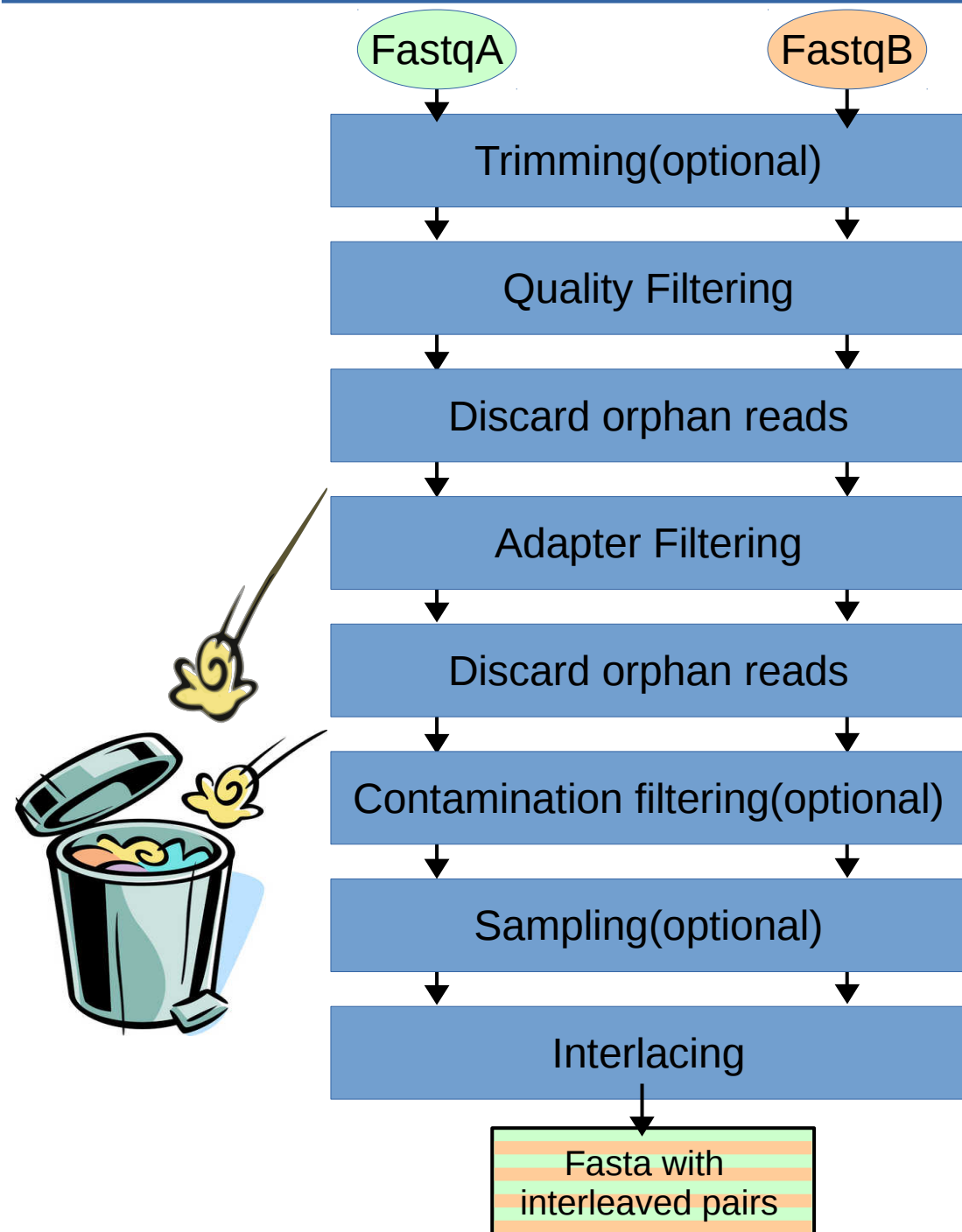
- Preprocessing and QC tools
 - Quality filtering
 - Sampling
 - Comparative analysis preparation
- Assembly annotation
 - DANTE
 - Profrep
- ChIP-Seq Mapper
- Nanopore Reads analysis...

Preprocessing of fastq paired-reads

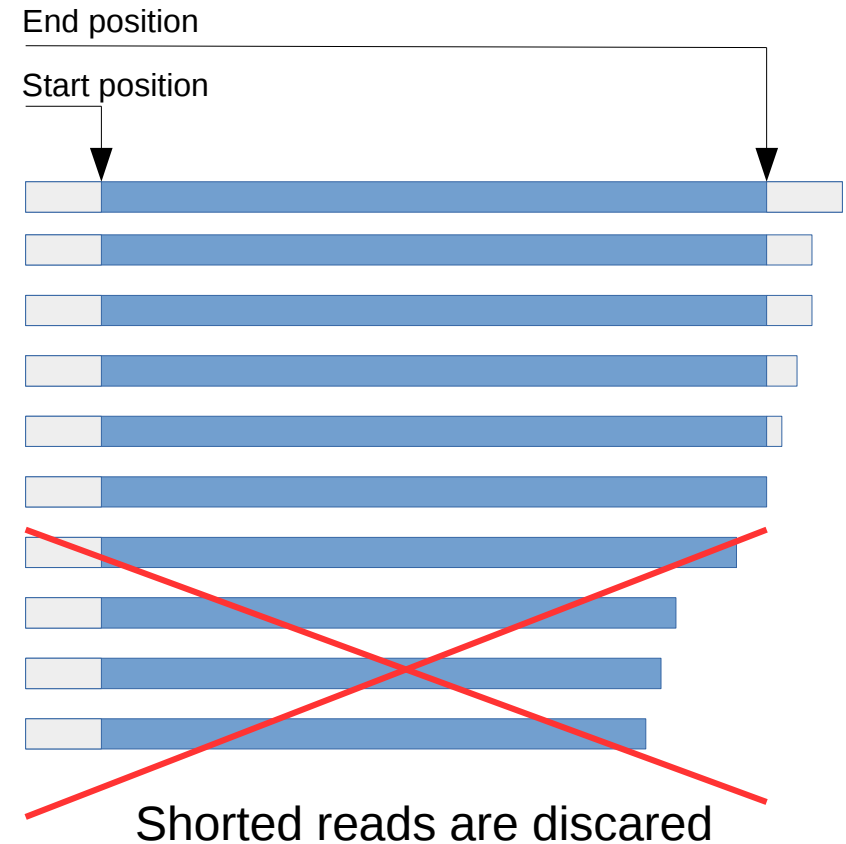
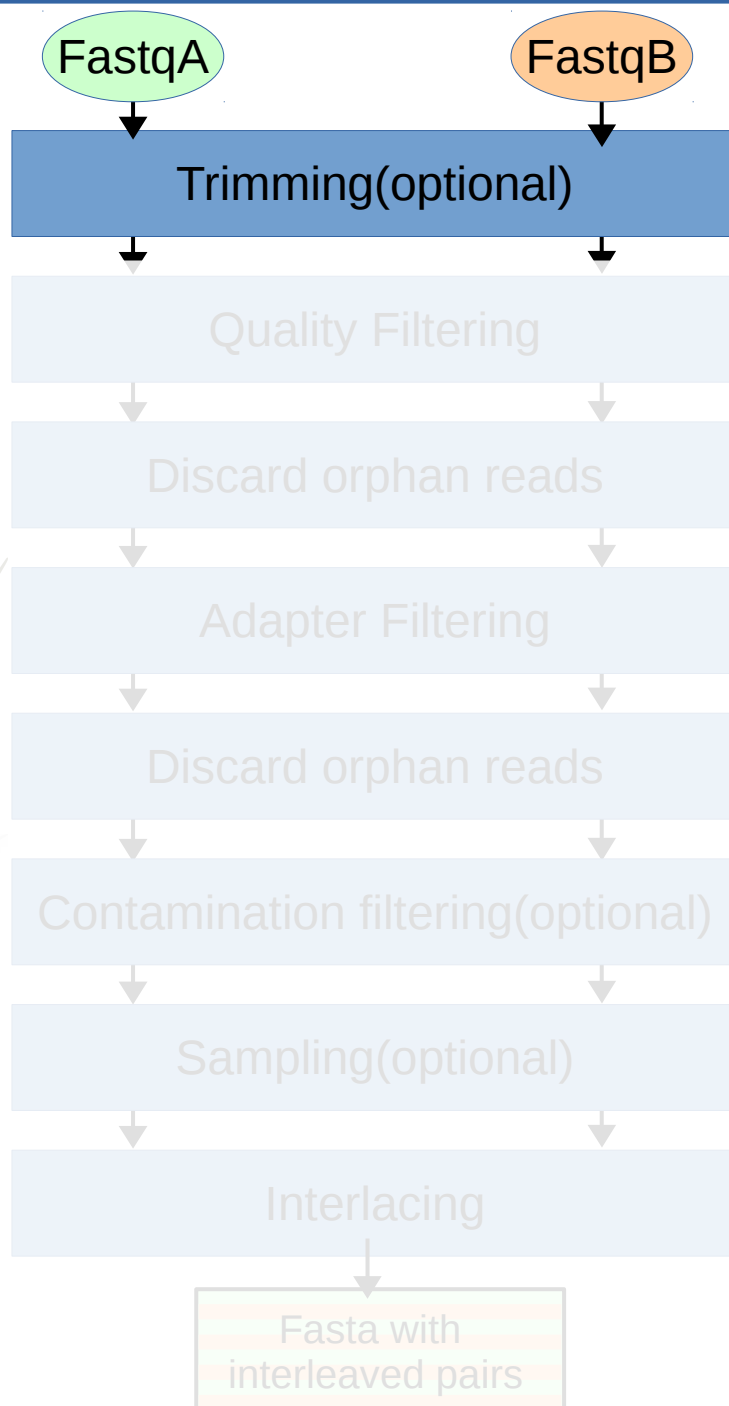
All in one tool for preparation of reads for RepeatExplorer clustering

- trimming
- quality filtering
- adapter removing
- sequence filtering
- sampling

Preprocessing of fastq paired-reads

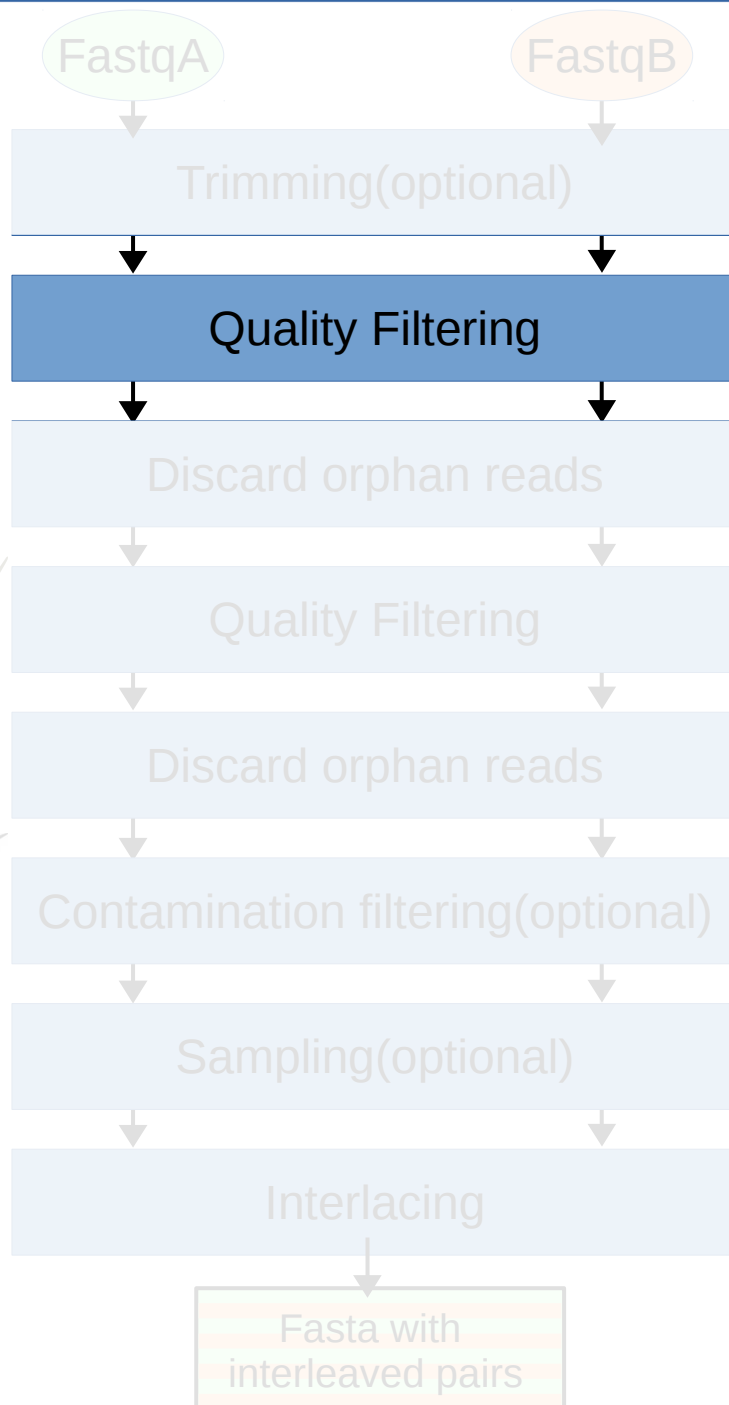


Preprocessing of fastq paired-reads



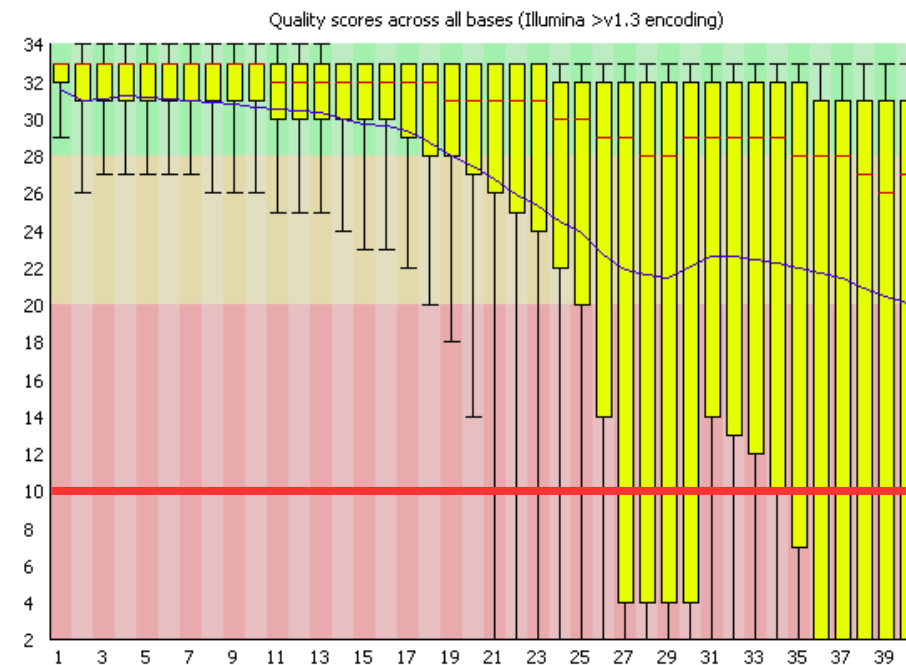
If your reads have variable length, it is recommended to use trimming!

Preprocessing of fastq paired-reads

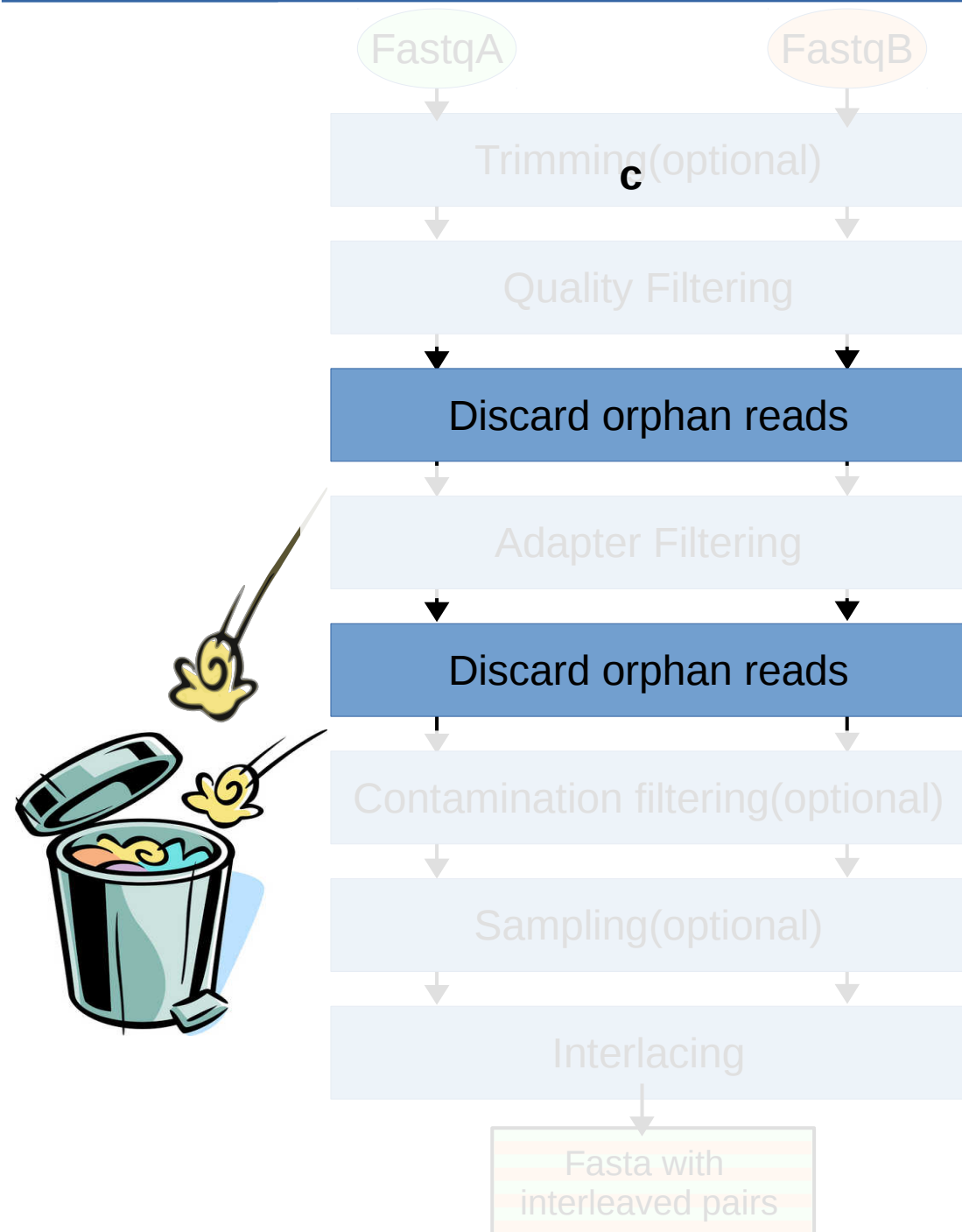


Default filtering:

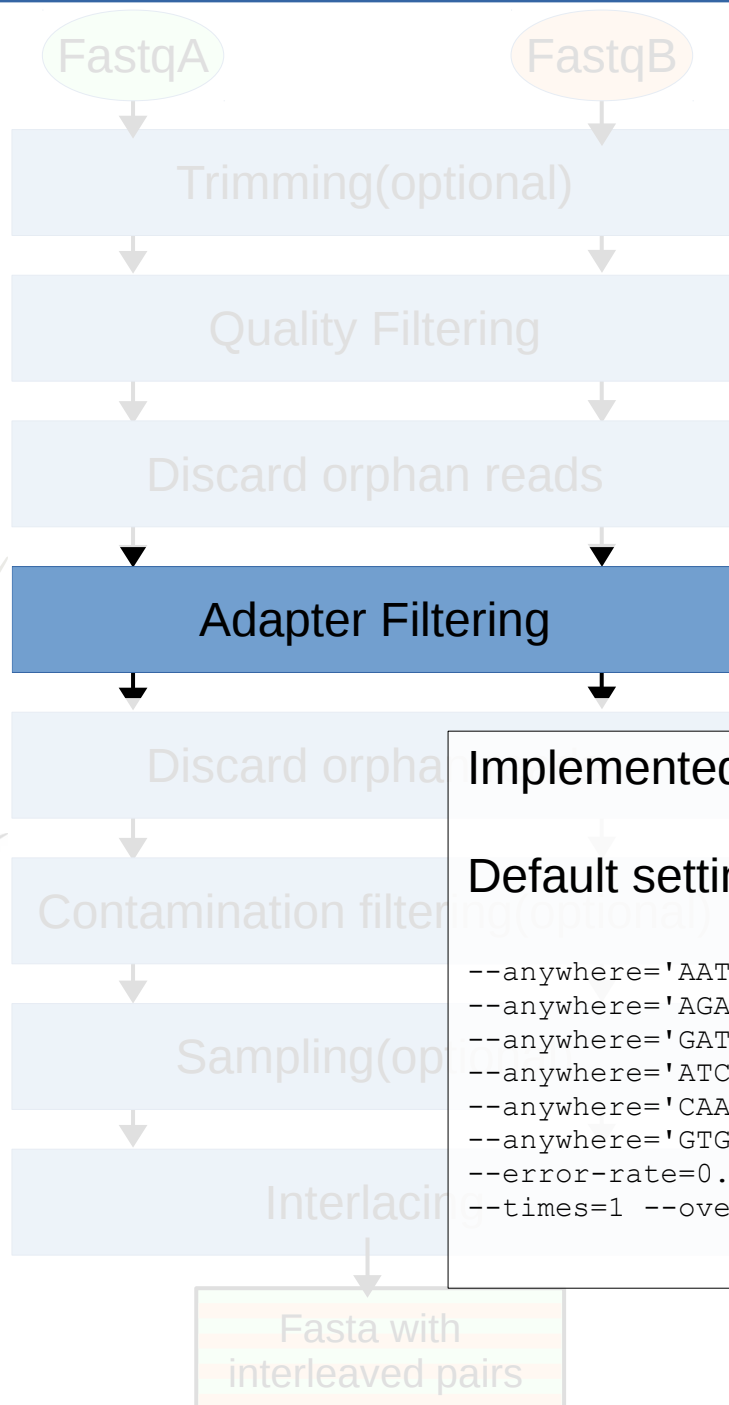
quality cutoff : 10
nucleotides above cutoff : 95%
Maximum Ns : 0



Preprocessing of fastq paired-reads



Preprocessing of fastq paired-reads

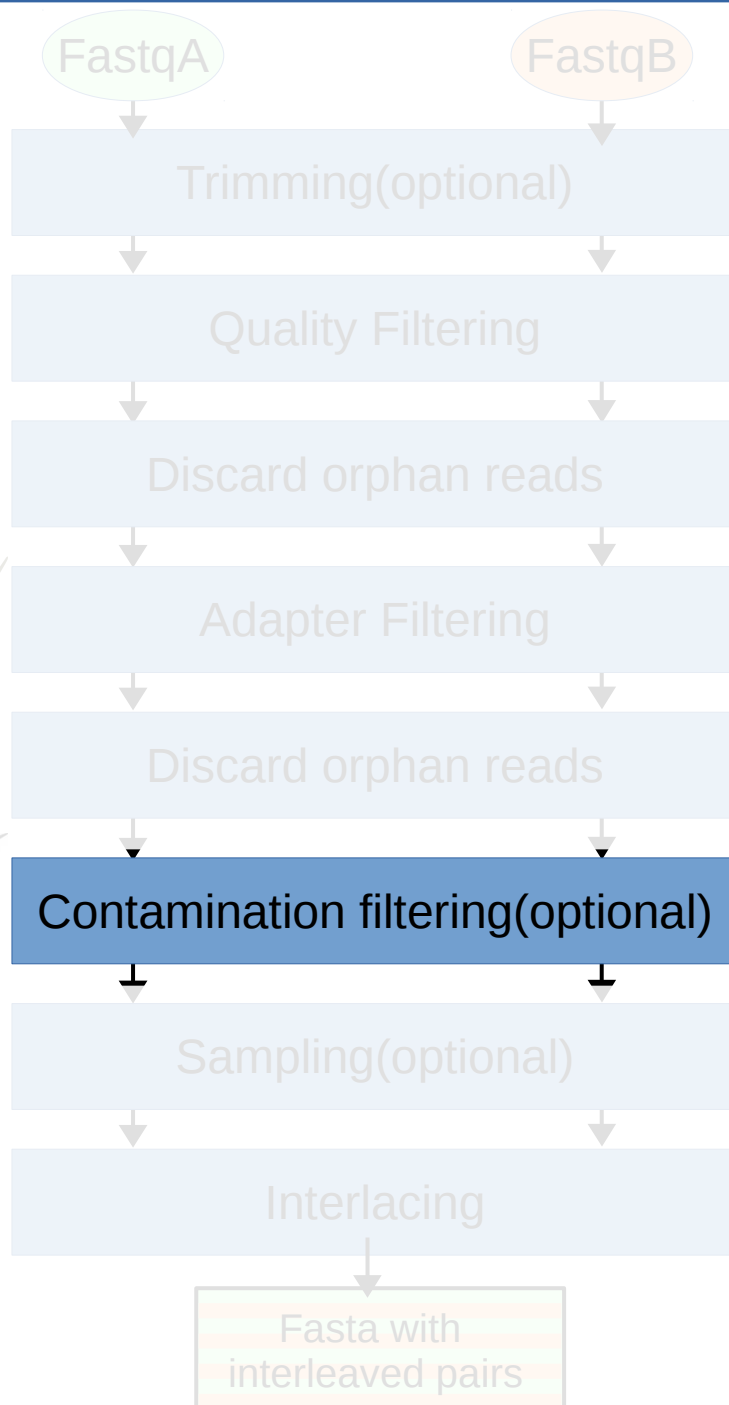


Implemented using **cutadapt** program

Default settings:

```
--anywhere='AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT'  
--anywhere='AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT'  
--anywhere='GATCGGAAGAGCACACGTCTGAACTCCAGTCAC'  
--anywhere='ATCTCGTATGCCGTCTTCTGCTTG'  
--anywhere='CAAGCAGAAGACGGCATACGAGAT'  
--anywhere='GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC'  
--error-rate=0.05  
--times=1 --overlap=15 --discard
```

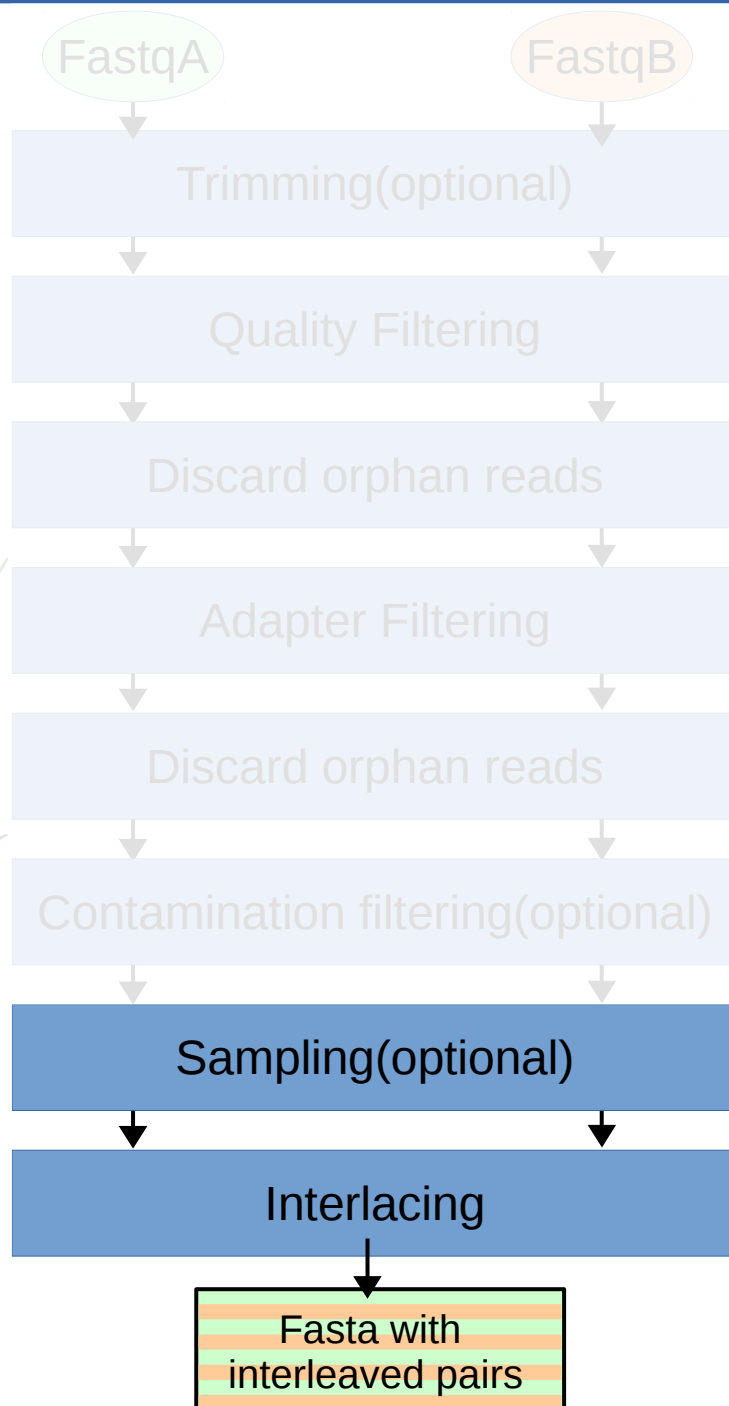
Preprocessing of fastq paired-reads



Provide fasta file (chloroplast, mitochondrial , bacteriophage phi-X174,...)

Sequence reads with identity above 90% over 90% of read length are removed.

Preprocessing of fastq paired-reads



Sample size = number of pairs!
For clustering 5M pairs is sufficient
most of the time

```
>SRR2040790.32 HWI-ST1146:96:C1J47ACXX:8:1101:3552:2172 1
ATTGATTCTCATTGGACAAACCTAAAAATCCTCCATAAGAGAGCATGAGATTGCTTTGAAGCTTCACGAAAGTGCTTCTCGGTGAATCT
CCGGTTATTTA
>SRR2040790.32 HWI-ST1146:96:C1J47ACXX:8:1101:3552:2172 2
TTTCTACTAAAGAGCCTTCGAGTCACCTTTCTTCTTCAAGAGAAACGATGTCAGATCTTCCACTAGATTGGTAGAAGAACTACTCTC
TCGTGTCCTCGG
>SRR2040790.86 HWI-ST1146:96:C1J47ACXX:8:1101:5893:2119 1
AGCTGAATTAAGGAAATTTTGTGAGCTTCAAAGAAATCAATCAATGGTATCTCTGAAAGCAATTCTTGCATATTTCAACAAATCAAAT
CTTGTAGACTA
>SRR2040790.86 HWI-ST1146:96:C1J47ACXX:8:1101:5893:2119 2
CGATTTAATAAGTACAATATATTTTGGTCTATTATATGGCAATTACAACATACTTTCTTAATCATTGTAATTTATAAACATGAAACT
TCAAAGGTTTG
>SRR2040790.96 HWI-ST1146:96:C1J47ACXX:8:1101:6039:2225 1
AGAAGAAGGTCTGCATATTTCTGAGGTTTGAGGTAGACATACATTGTCTTAATGGCAACAAATGACCCGTTATTGCGATACAAGTGATA
TCAATCGAGTG
>SRR2040790.96 HWI-ST1146:96:C1J47ACXX:8:1101:6039:2225 2
CTTCAAGTTCACTTTTAAGTTGAATCATTCTTGGCAAGGACCTAACAGCGGTAATATCCTTTCTTGCATTTATCACACTCGATTGATA
TCACTTGATC
>SRR2040790.104 HWI-ST1146:96:C1J47ACXX:8:1101:6592:2227 1
AGACAATAGAATCTAGGTGAGATTTTAATCGGCTAACCATACATTAGAAATAAATTTGTATAAAACATTACACACGCTATTGGTCGAT
AATCCGAGAAA
>SRR2040790.104 HWI-ST1146:96:C1J47ACXX:8:1101:6592:2227 2
TTACAACATCTACTATGAAGCCGCACATTAATCATACAAACCTTTGTCTACTTCCGAAGATTGAAAATCCTTCAACTTTCTCGGATTATC
GACCAATAGCG
>SRR2040790.133 HWI-ST1146:96:C1J47ACXX:8:1101:7820:2106 1
GAGCTACAAAGCAAGCTACAAACCTCGGAATAACTAGATAGGCGGATGACTGATTACCTGAGAGATATTAAGAAATGTTTGTGAGCAACTT
GCATCTATAAG
>SRR2040790.133 HWI-ST1146:96:C1J47ACXX:8:1101:7820:2106 2
TAATTCAGCAAAAATTTTCATCTTTCTAGTACATGACTACTTATAGATGCAAGTTGCTCACAAACATCTTAATATCTCTCAGGTAATC
AGTCATCGGCC
>SRR2040790.166 HWI-ST1146:96:C1J47ACXX:8:1101:9075:2203 1
GAAAAGTGGAGTATACATCAAGAAGCTCTCTGTAGTAGCAAGCAGGATTTTACGATTAGTTCTACTTCTAACTCCGTGAACATATGCG
AATAGCTCTGC
>SRR2040790.166 HWI-ST1146:96:C1J47ACXX:8:1101:9075:2203 2
CATTACCATGGAAATAGTATTTTGGAGCTGAACATAGGATAAGTACGCGAGGTGTTTATCCAGAAAGCGATTCCGGCACACCACTAC
GTCAGTATCAA
>SRR2040790.200 HWI-ST1146:96:C1J47ACXX:8:1101:10291:2231 1
CATCCGCTATTGAATCAGCTGCTATAGACTTAGCTAAAGAAATGCTTGCACGAGATTAATGTTATGTGCTTGTATGCGATCACATTTT
CTCGTAGTTGA
>SRR2040790.200 HWI-ST1146:96:C1J47ACXX:8:1101:10291:2231 2
ACAAGTACCAGGCCTAGACCGGGGAGTCGGAACTAATACAGCTTTTATAAATCCAAATCCAACCTTCCAACCTCCAAGTTGAACGCAG
GAAACTCTACC
>SRR2040790.266 HWI-ST1146:96:C1J47ACXX:8:1101:13442:2126 1
CTTGTGGTTTCGTTGATGATAGACTTATGGATGTTAGGACTACACCTCATCCTAATTTGTTCTTTGATGATTATGTCAAGCTCCTT
ATCTCTGTGA
>SRR2040790.266 HWI-ST1146:96:C1J47ACXX:8:1101:13442:2126 2
TCTTGAGTGATCTGCACATGGTCGACTTGGTGCTATTGCTGTTGTCTACACGGCTTCAATCCTAGAGTGCTTAAGTCATAAGTTCTTG
AGTTGGCTGGA
```

Annotation of repeats in assemblies

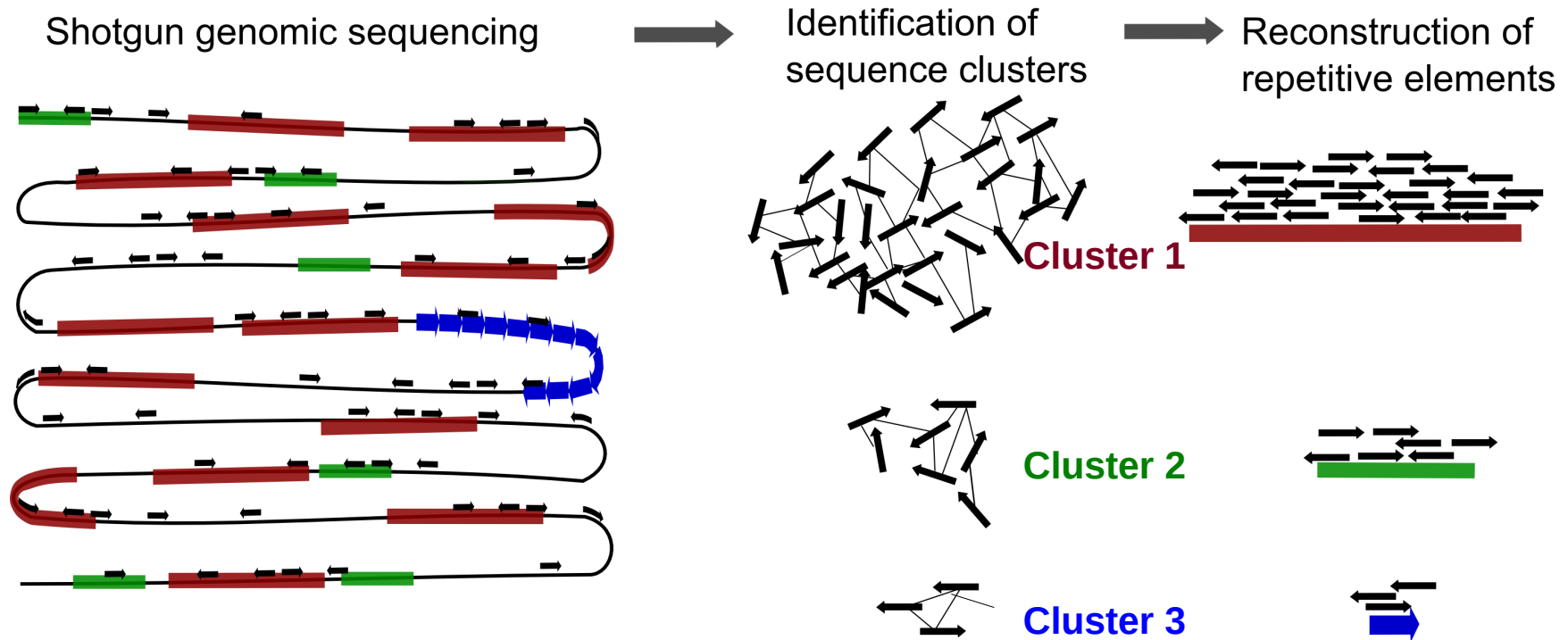
Profrep

annotation of assemblies using RepeatExplorer output

DANTE

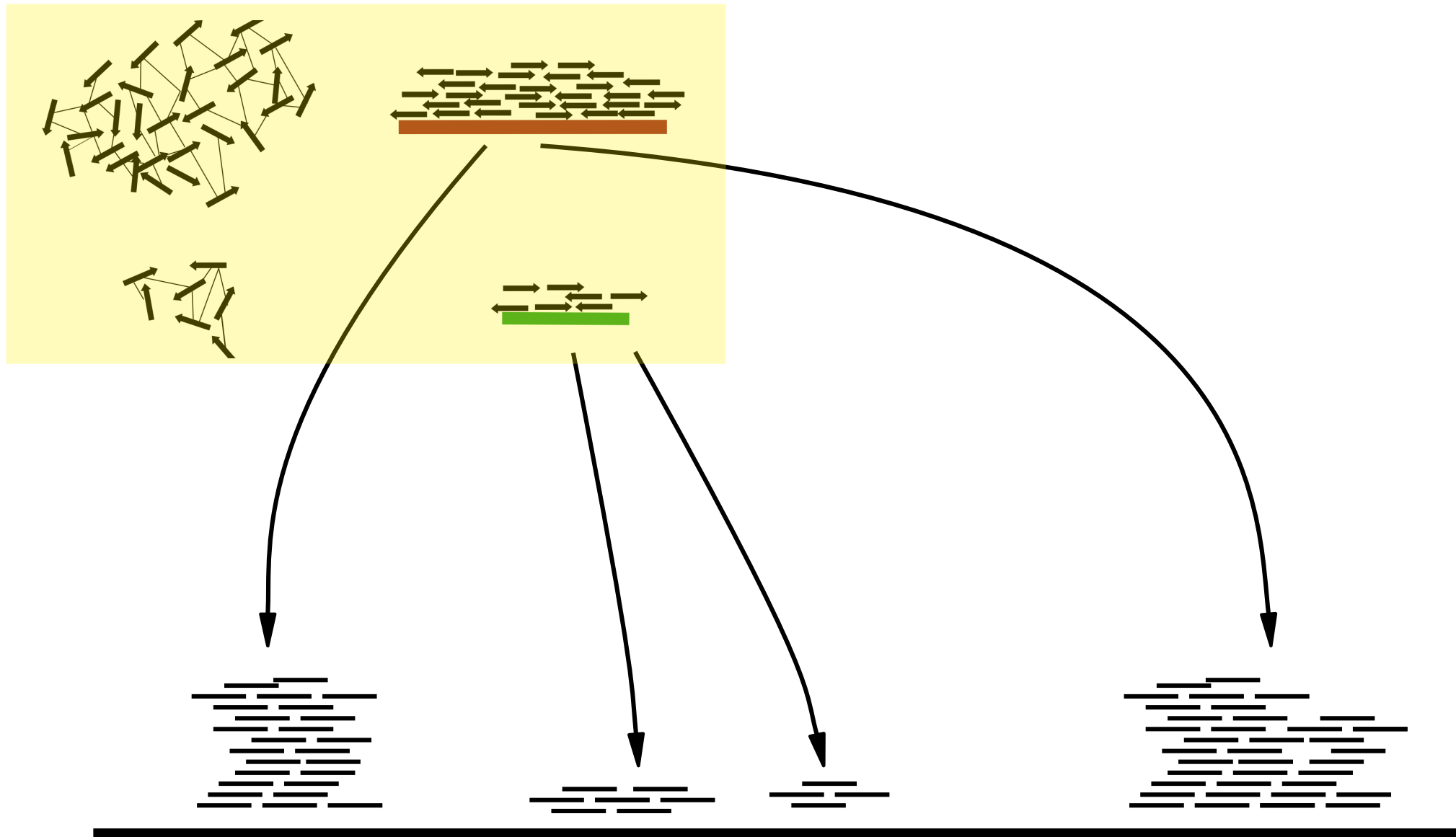
Annotation of assemblies using REXdb

Profrep

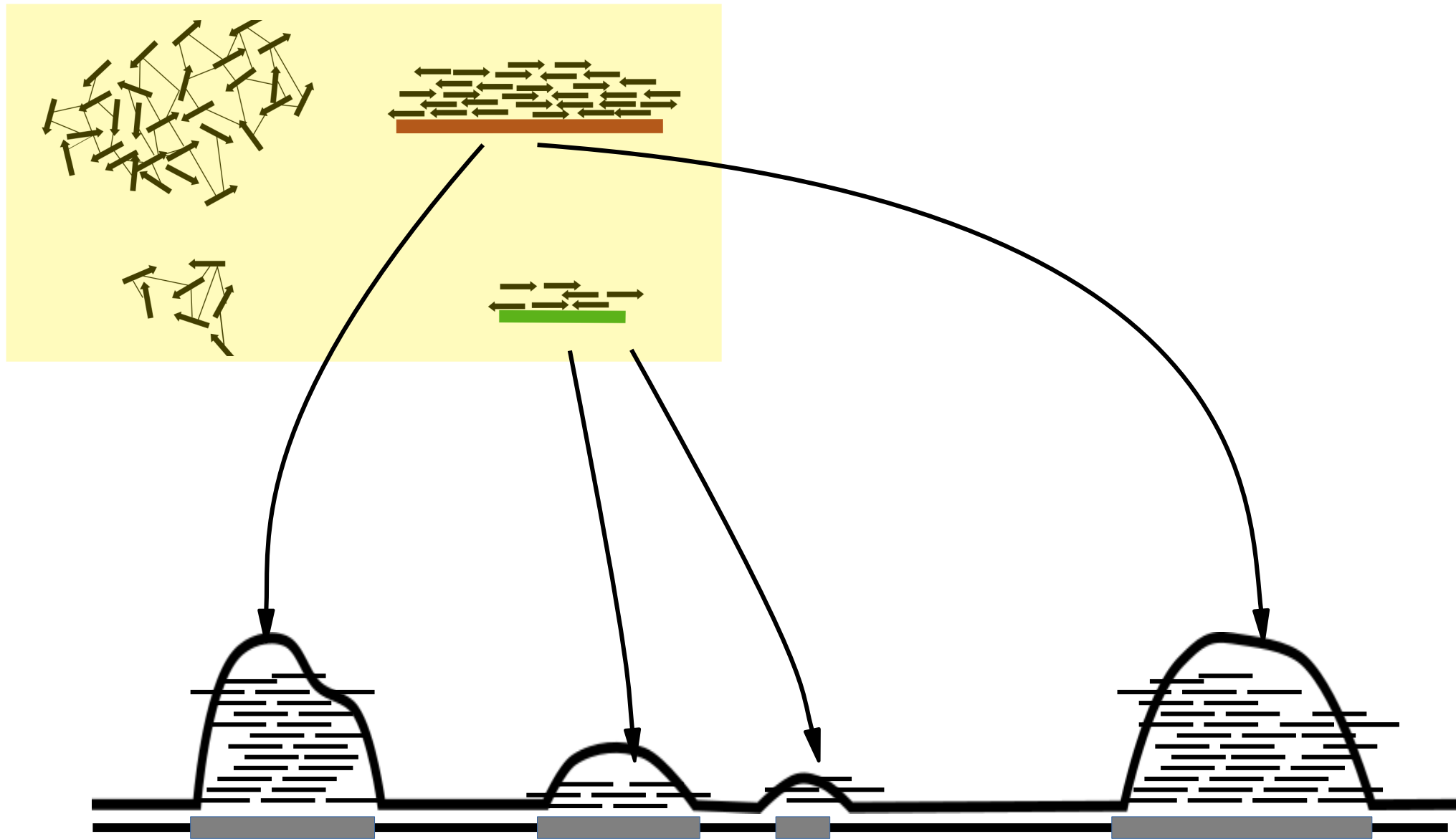


Profrep is mapping reads from clusters back to the genome

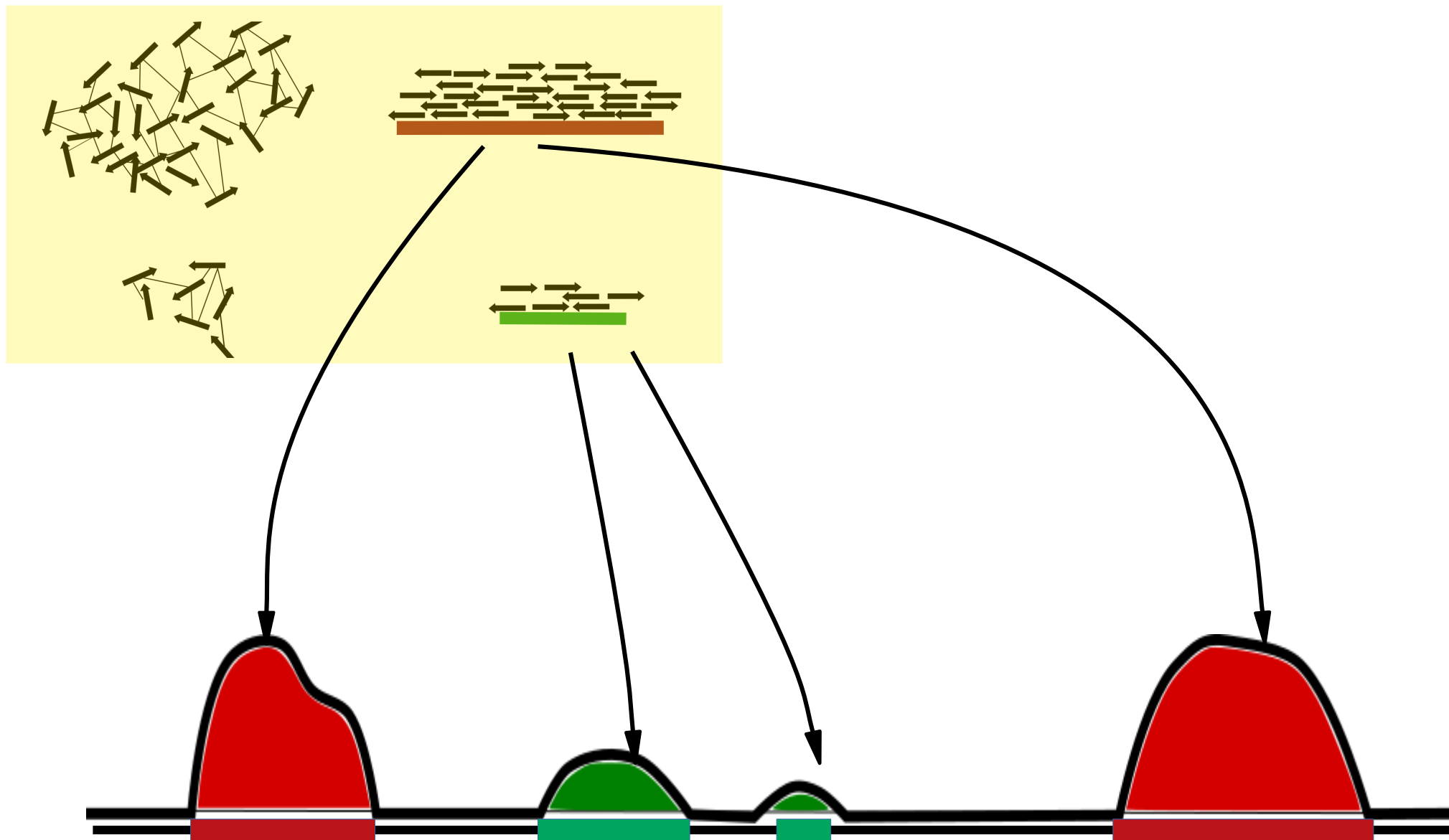
Profrep



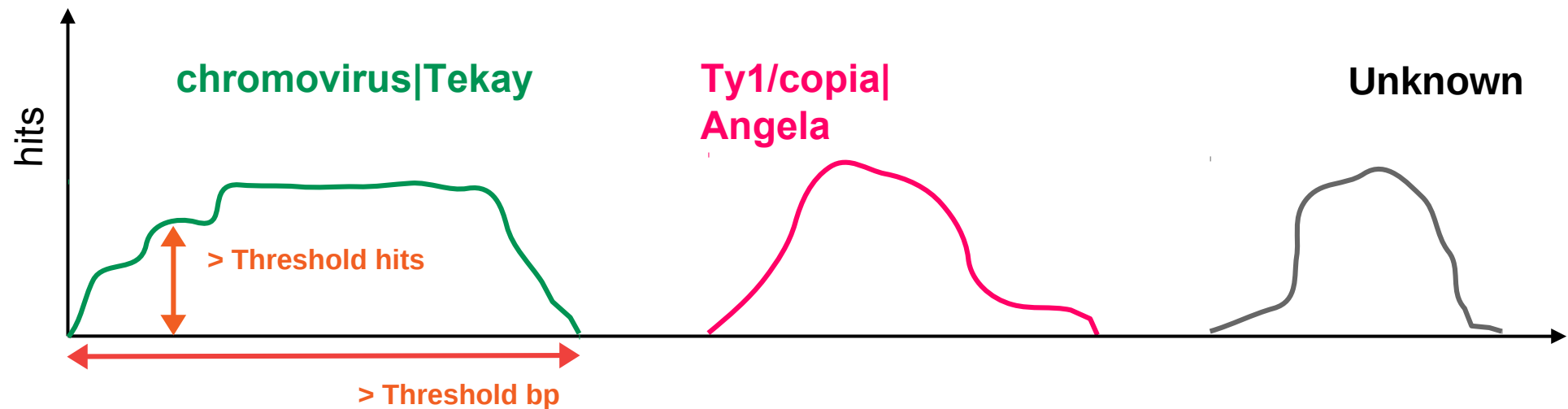
Profrep



Profrep



Profrep

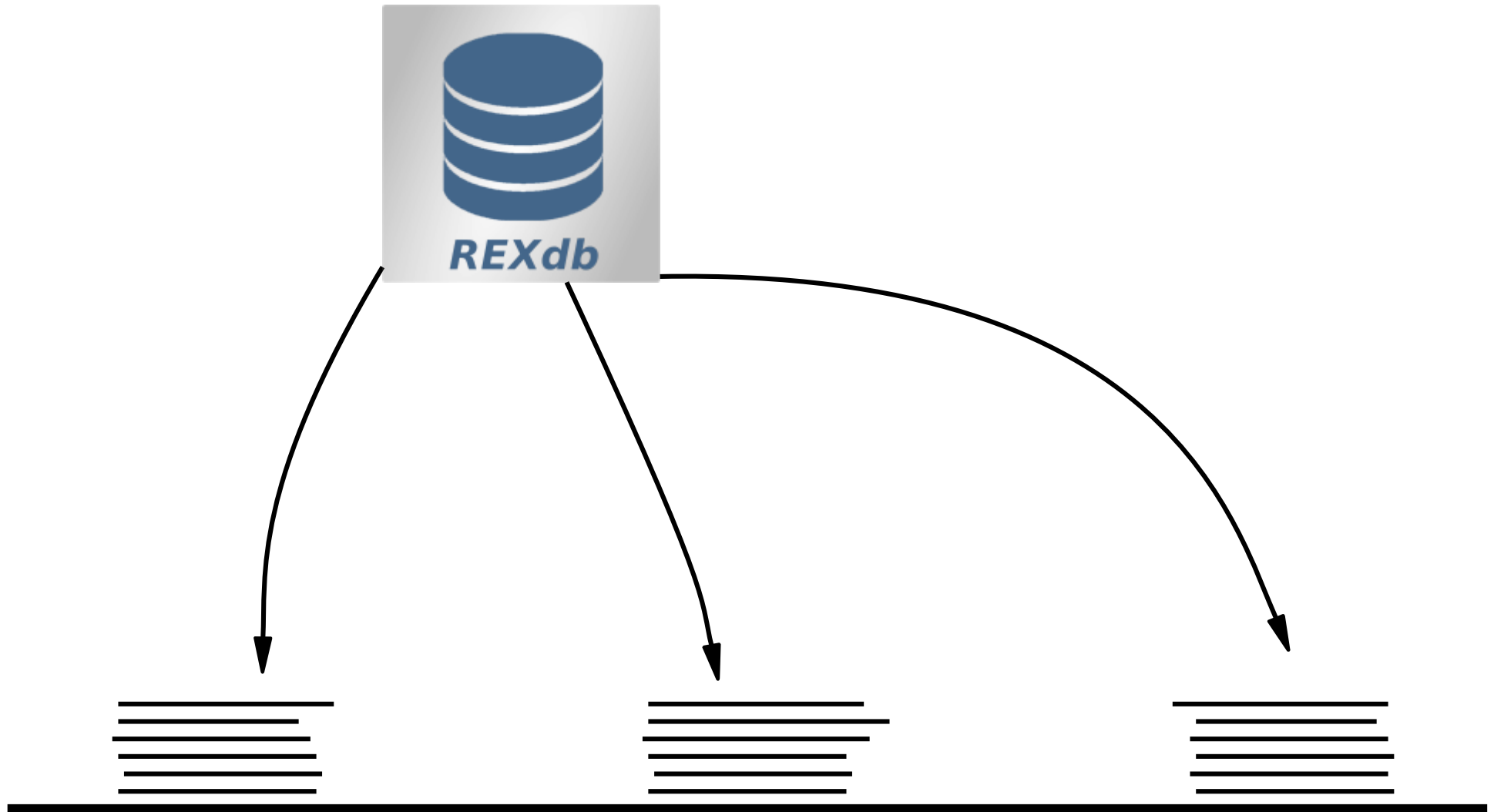


##gff-version 3

```
9 profrep repeat 53 1221 . . . Name=repeat|mobile_element|Class_I|LTR|Ty1/copia|SIRE;Average_PID=83
9 profrep repeat 7819 7921 . . . Name=repeat|mobile_element|Class_I|LTR|Ty3/gypsy|chromovirus|Tekay;Average_PID=77
9 profrep repeat 8365 8449 . . . Name=unknown_CL203;Average_PID=75
9 profrep repeat 9353 9434 . . . Name=unknown_CL209;Average_PID=83
```

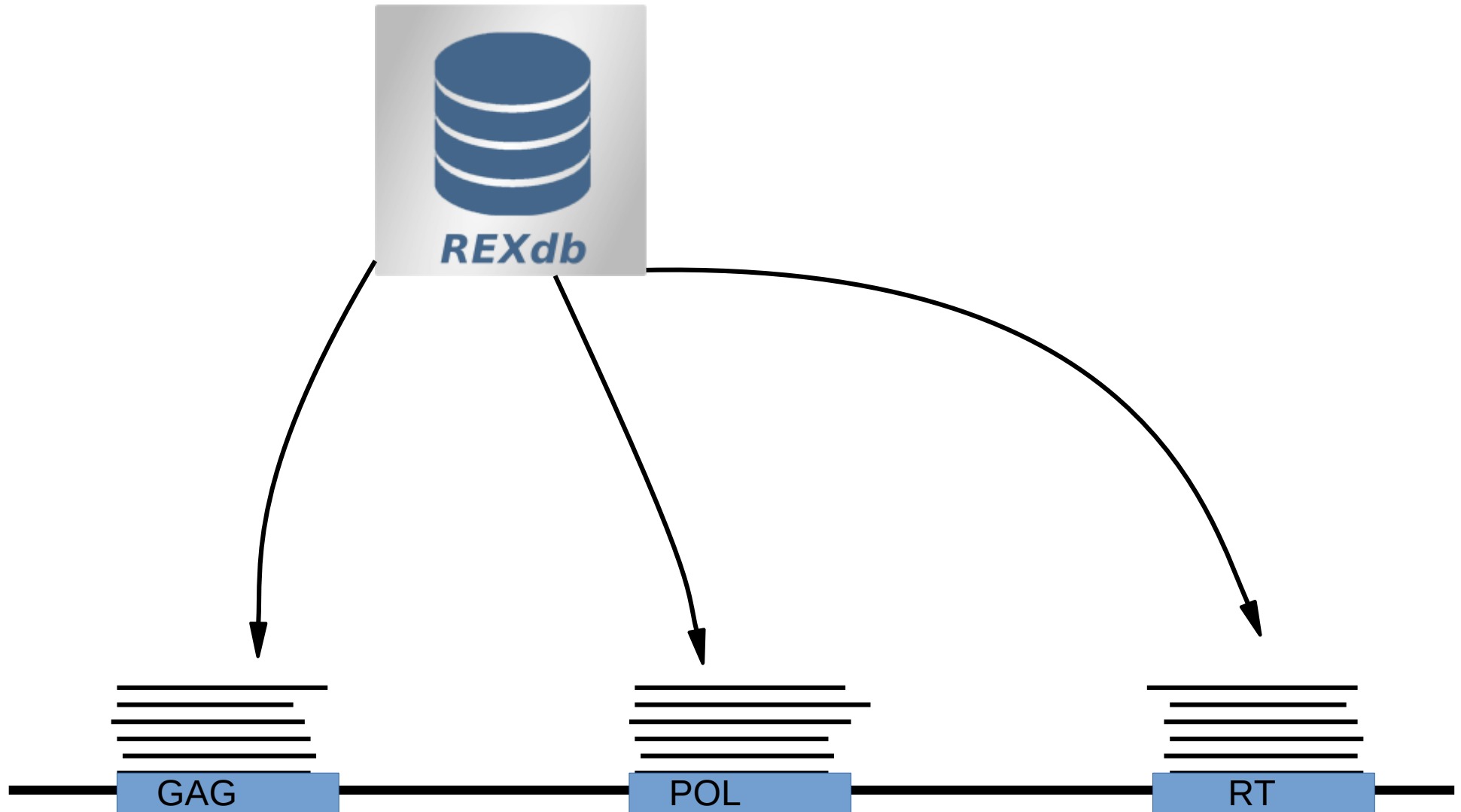
DANTE

Domain based annotation of transposable elements

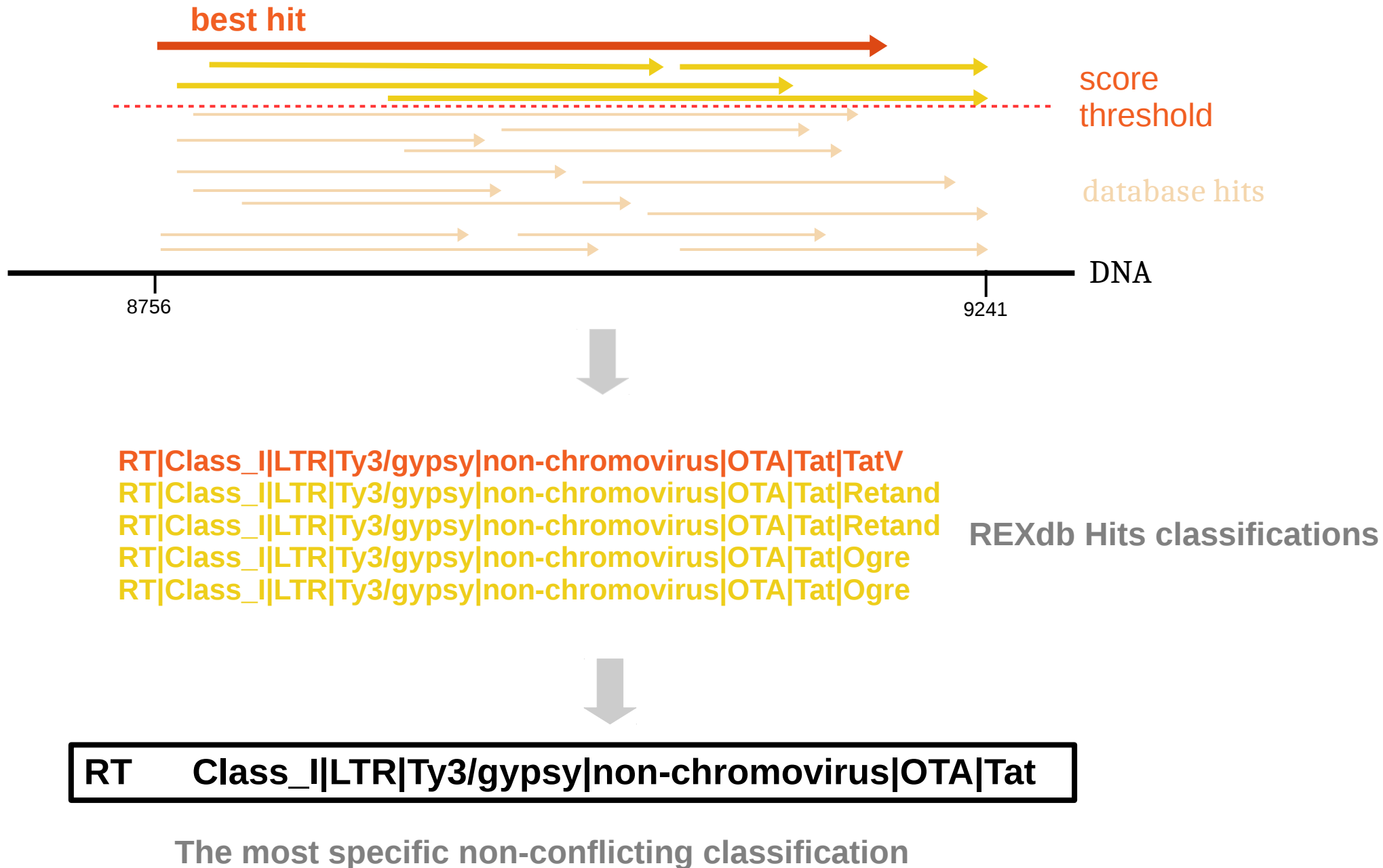


DANTE

Domain based annotation of transposable elements

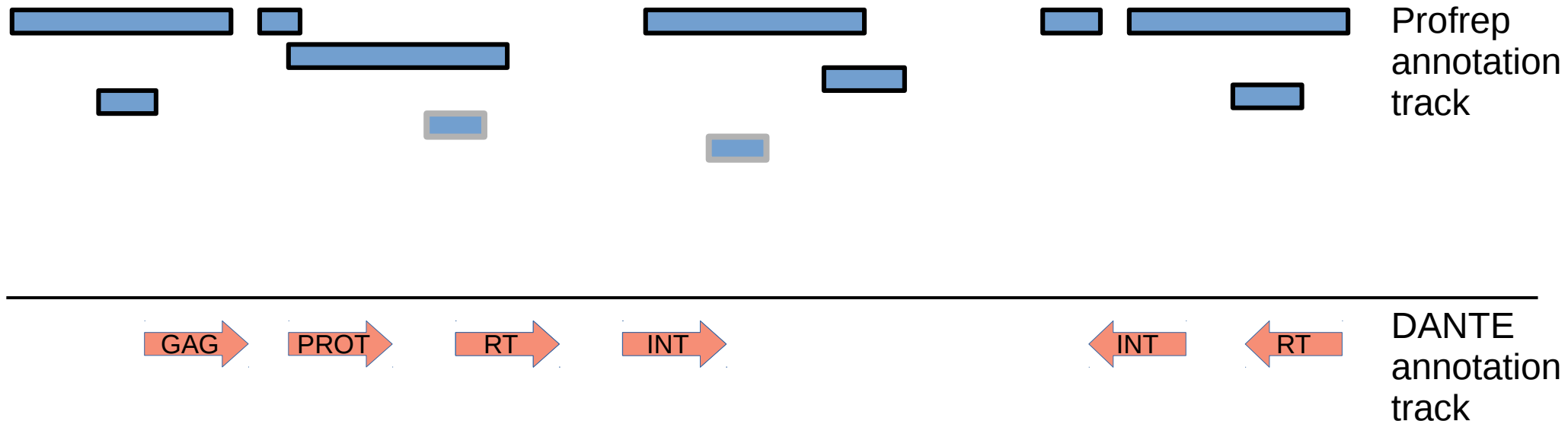


Domain based annotation of transposable elements



Combining Profrep and DANTE

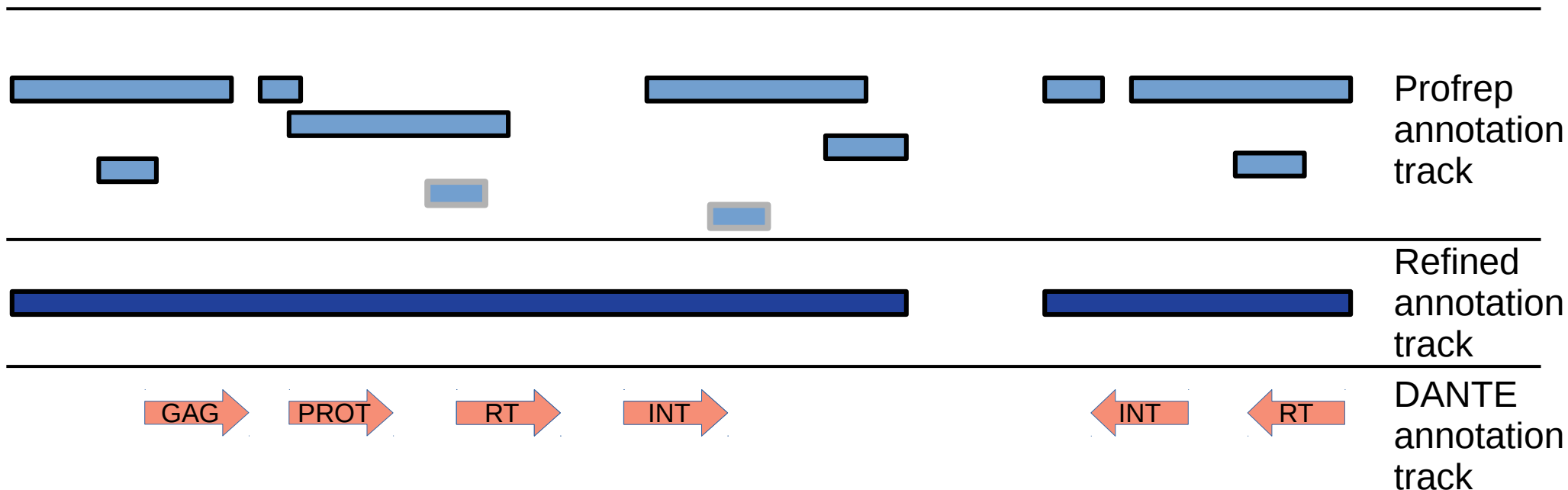
Profrep refiner tool



- Low quality hits are removed
- Overlapping or close annotated segments are merged providing the annotation is in agreement
- DANTE annotated domains are used as supporting information for segment merging (i.e. orientation and classification)

Combining Profrep and DANTE

Profrep refiner tool



- Low quality hits are removed
- Overlapping or segments in proximity are merged providing the annotation is in agreement
- DANTE annotated domains are used as supporting information for segment merging (i.e. orientation and classification)

Jbrowse - genome browser

- Web based
- Integrated into Galaxy environment
- Suitable for visualization of Profrep and DANTE results
- Portable



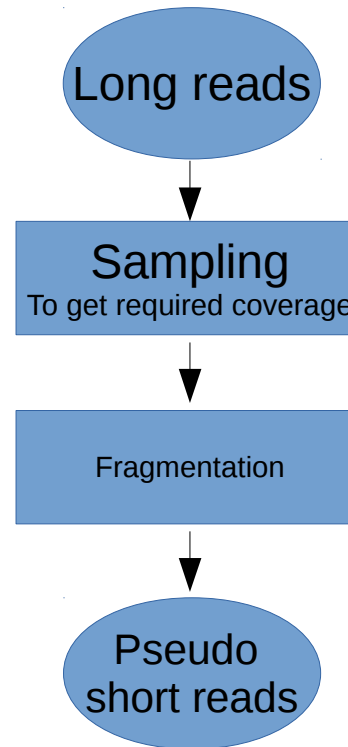
Using ultralong reads in RepeatExplorer clustering

- Ultralong reads – up to hundreds of kb
- Can we utilize such reads in RepeatExplorer pipeline?

Using ultralong reads in RepeatExplorer clustering

- Ultralong reads – up to hundreds of kb
- Can we utilize such reads in RepeatExplorer pipeline?

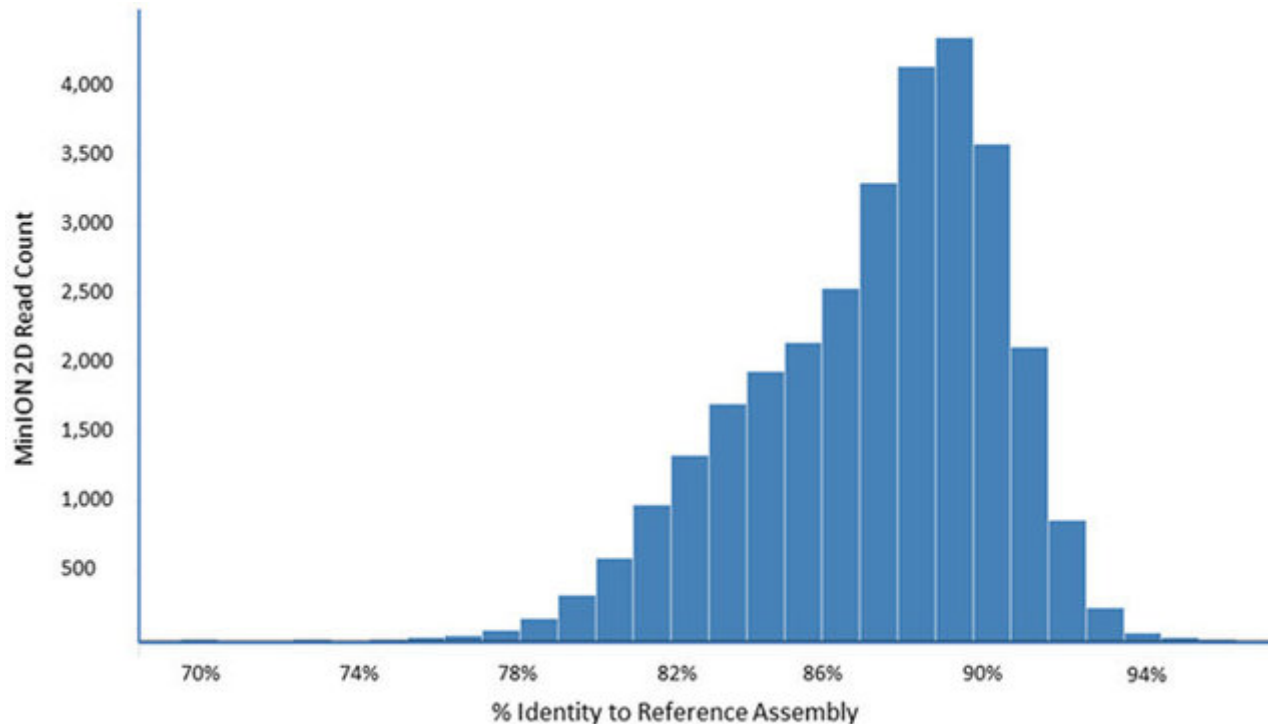
Reads preparation



Pseudo short paired end reads (defined by insert length, read length and coverage)
Original sequence position is coded in sequence name

Using ultralong reads in RepeatExplorer clustering

High error rate!



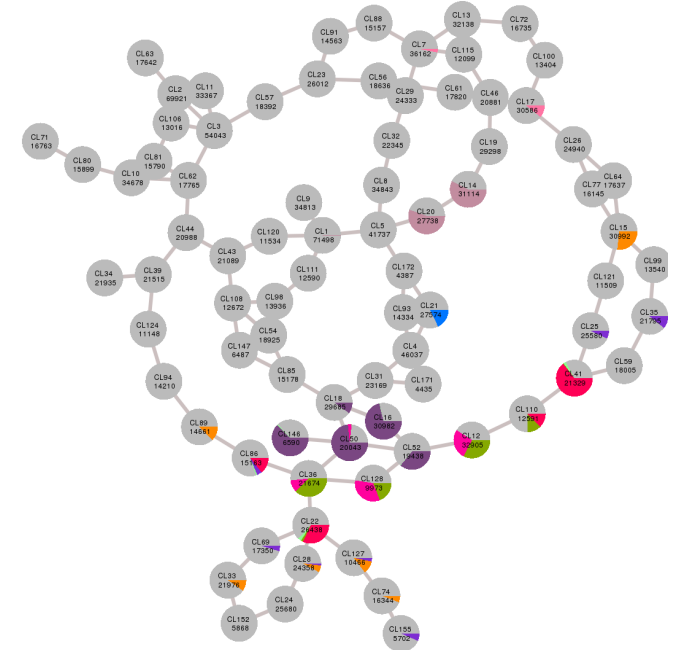
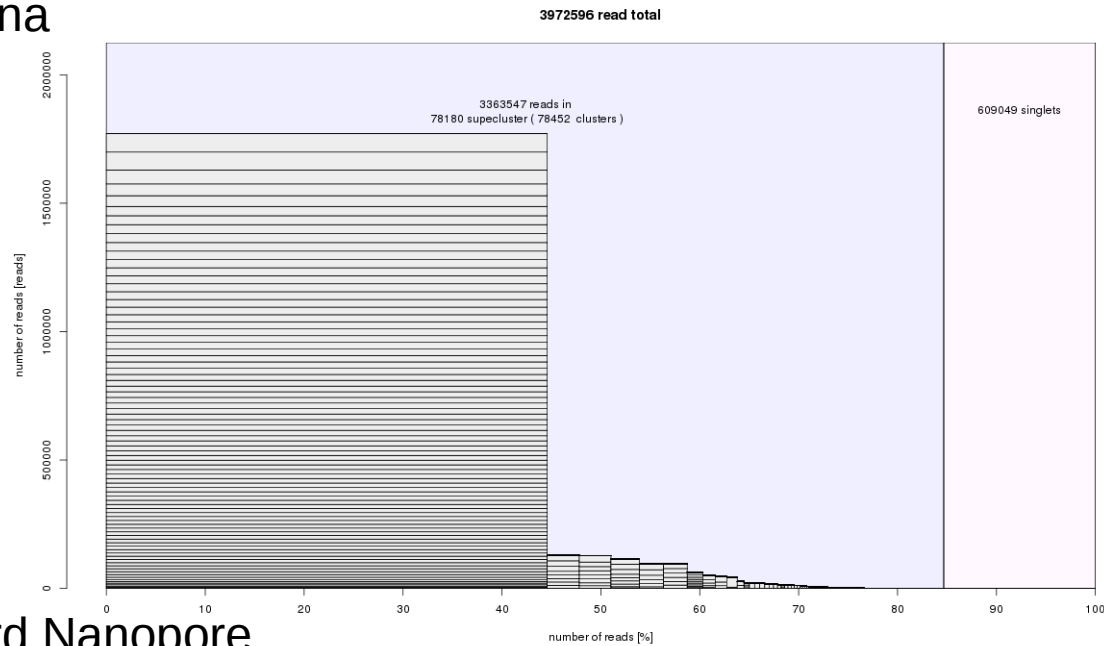
Default clustering threshold – 90% identity – is too strict for such high error rate.

Solution :

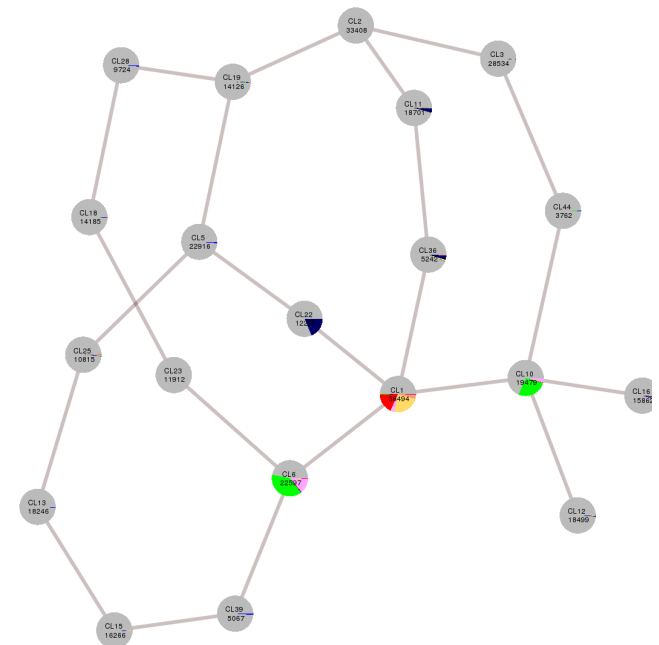
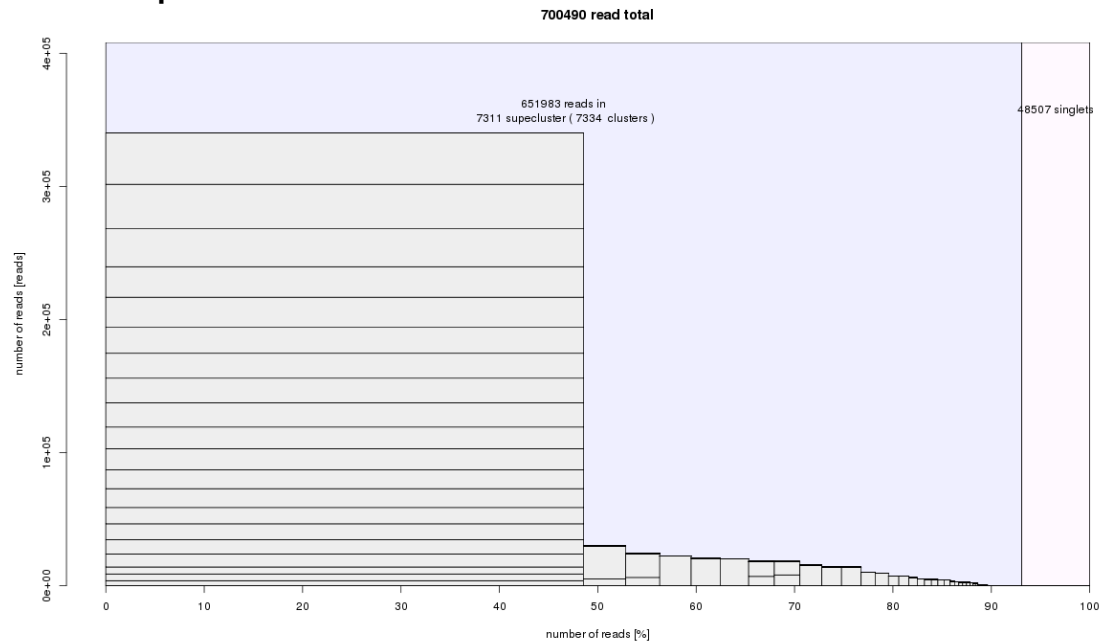
- clustering is performed on “longer” read (800nt)
- similarity search is performed using **lastal** program trained on ON reads
- similarity search must be more “indel” tolerant

Using ultralong reads in RepeatExplorer clustering

Illumina

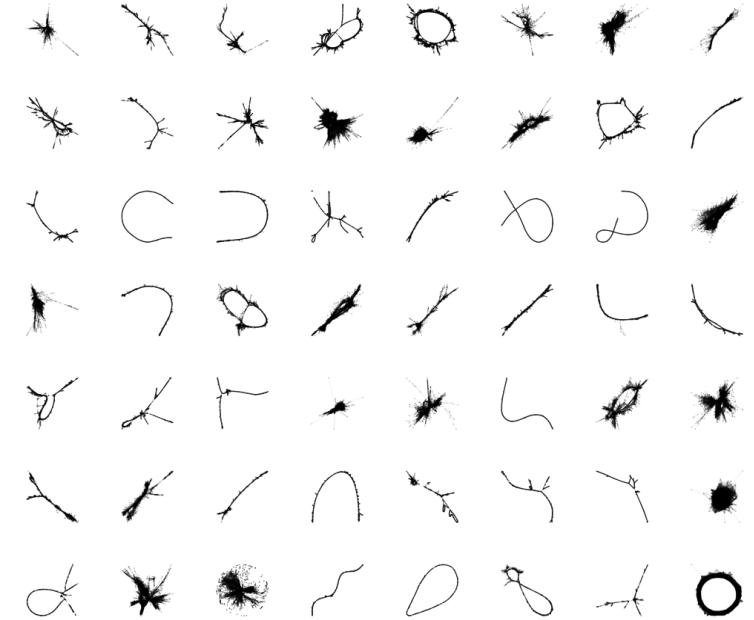
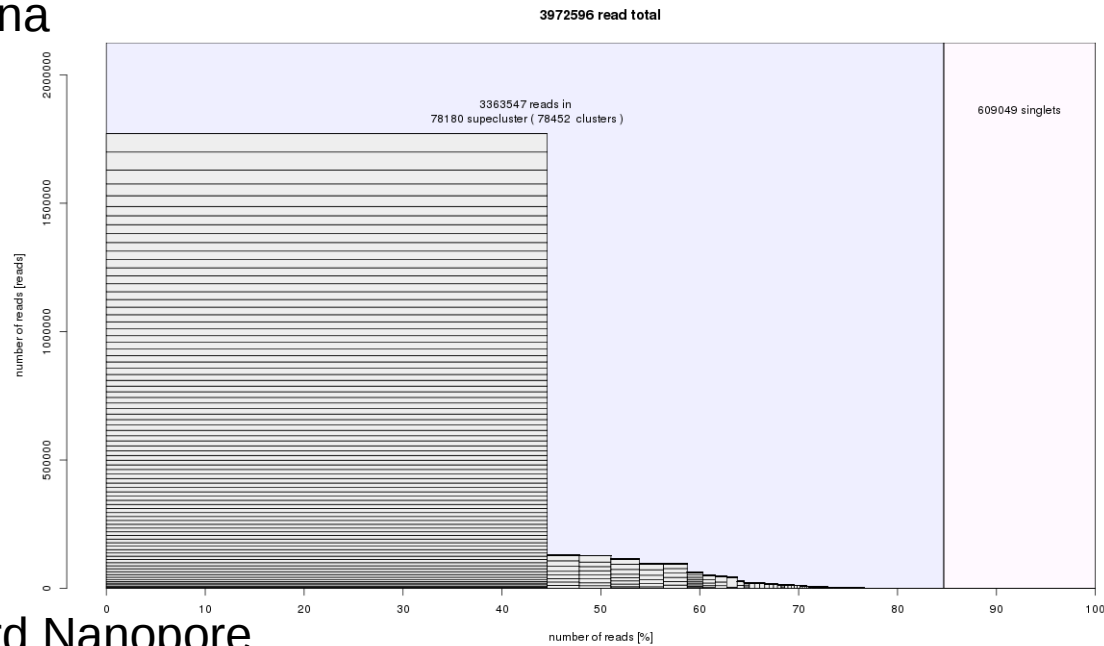


Oxford Nanopore

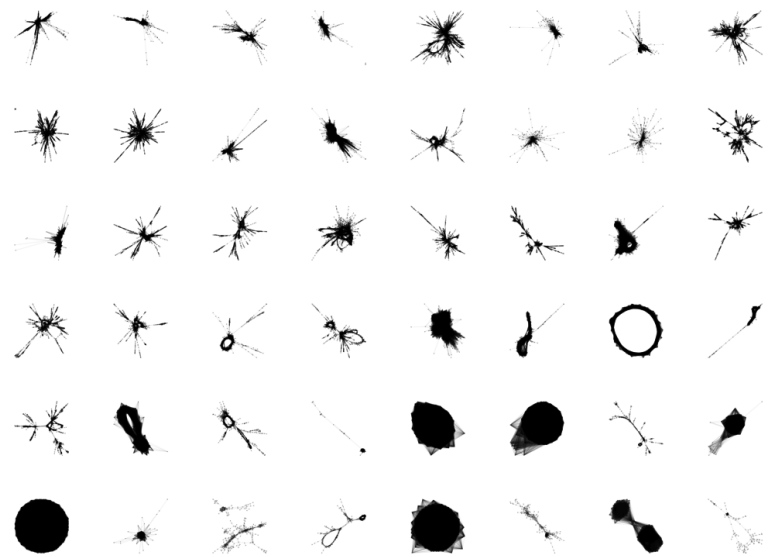
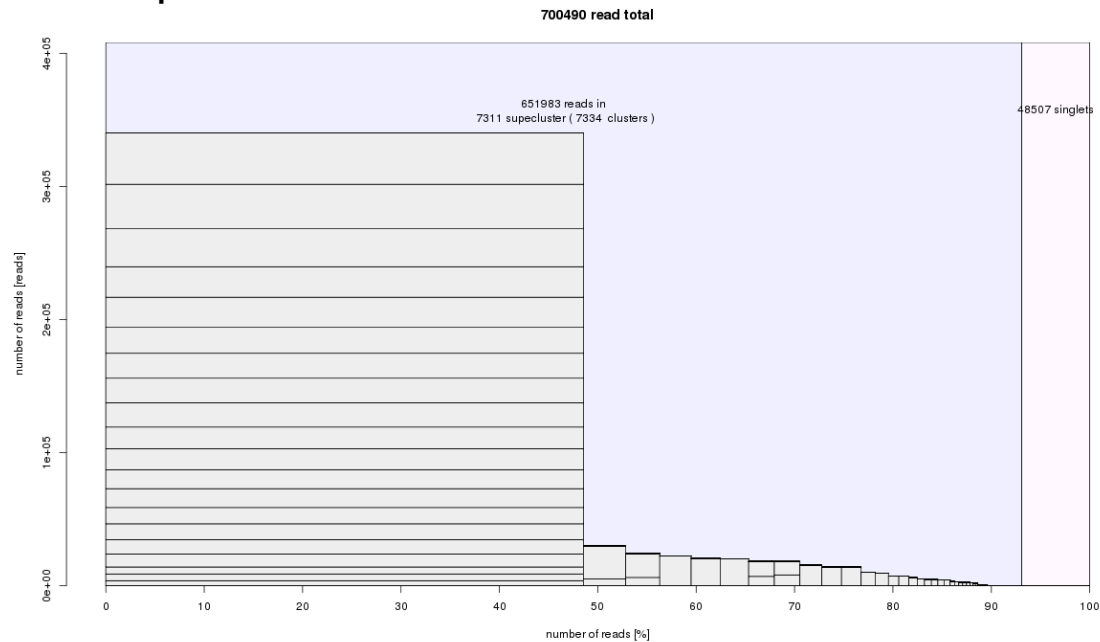


Using ultralong reads in RepeatExplorer clustering

Illumina

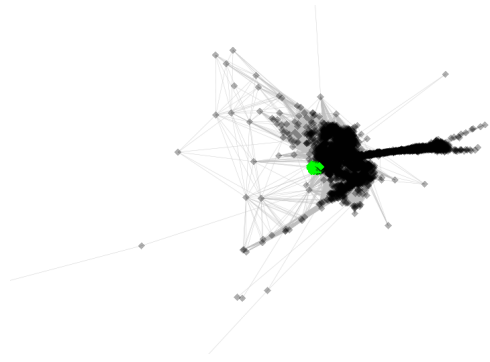


Oxford Nanopore

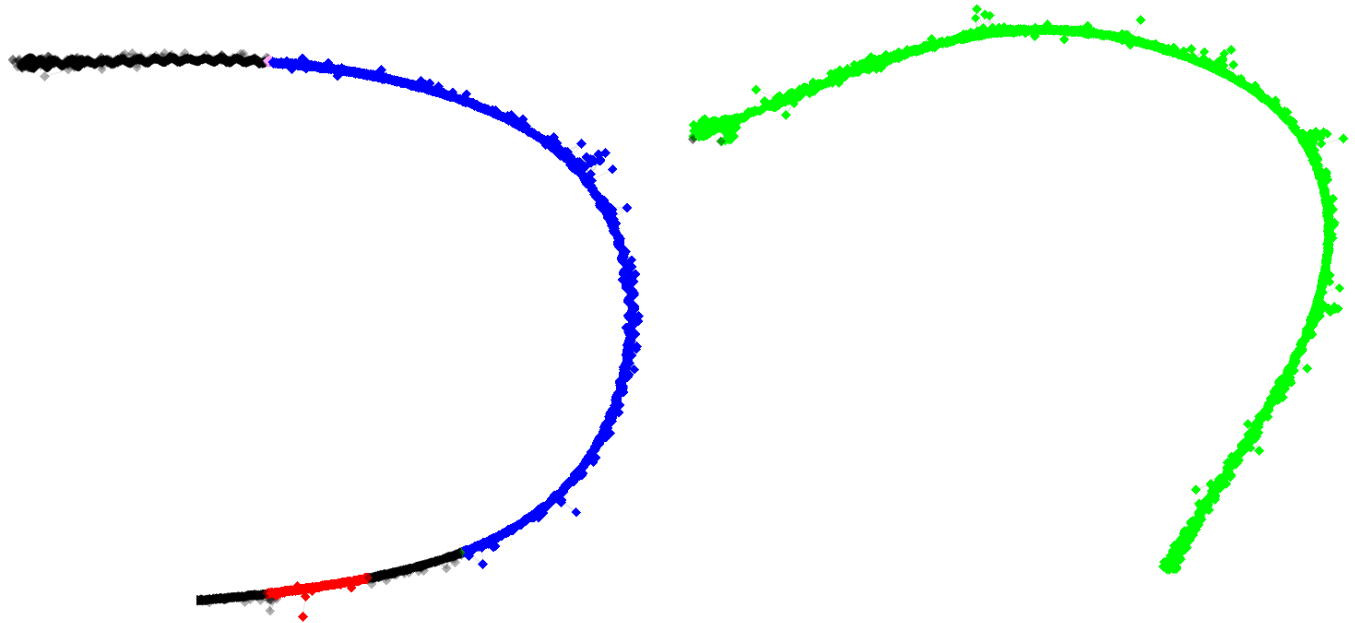


Using ultralong reads in RepeatExplorer clustering

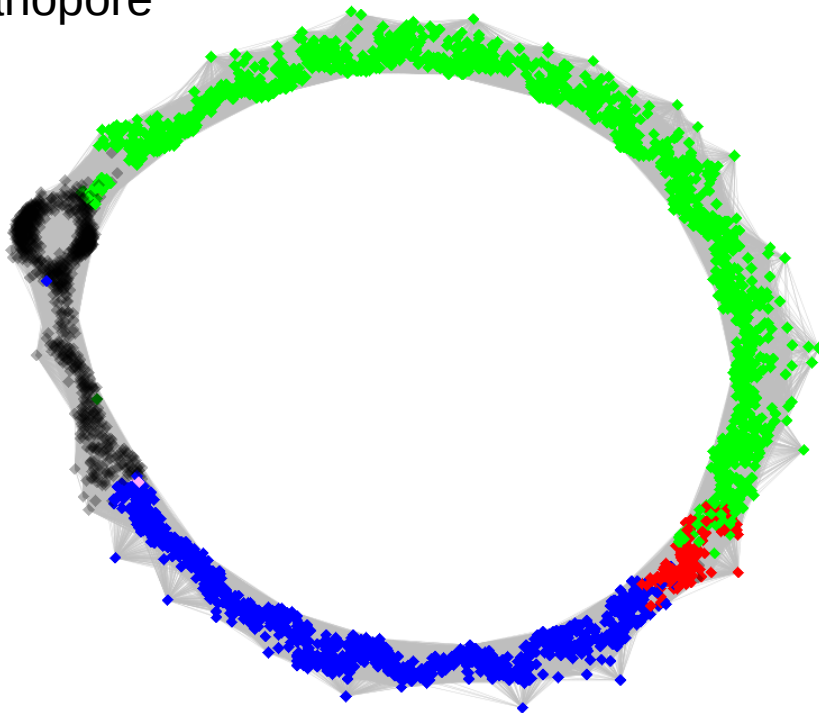
Illumina



0.55%



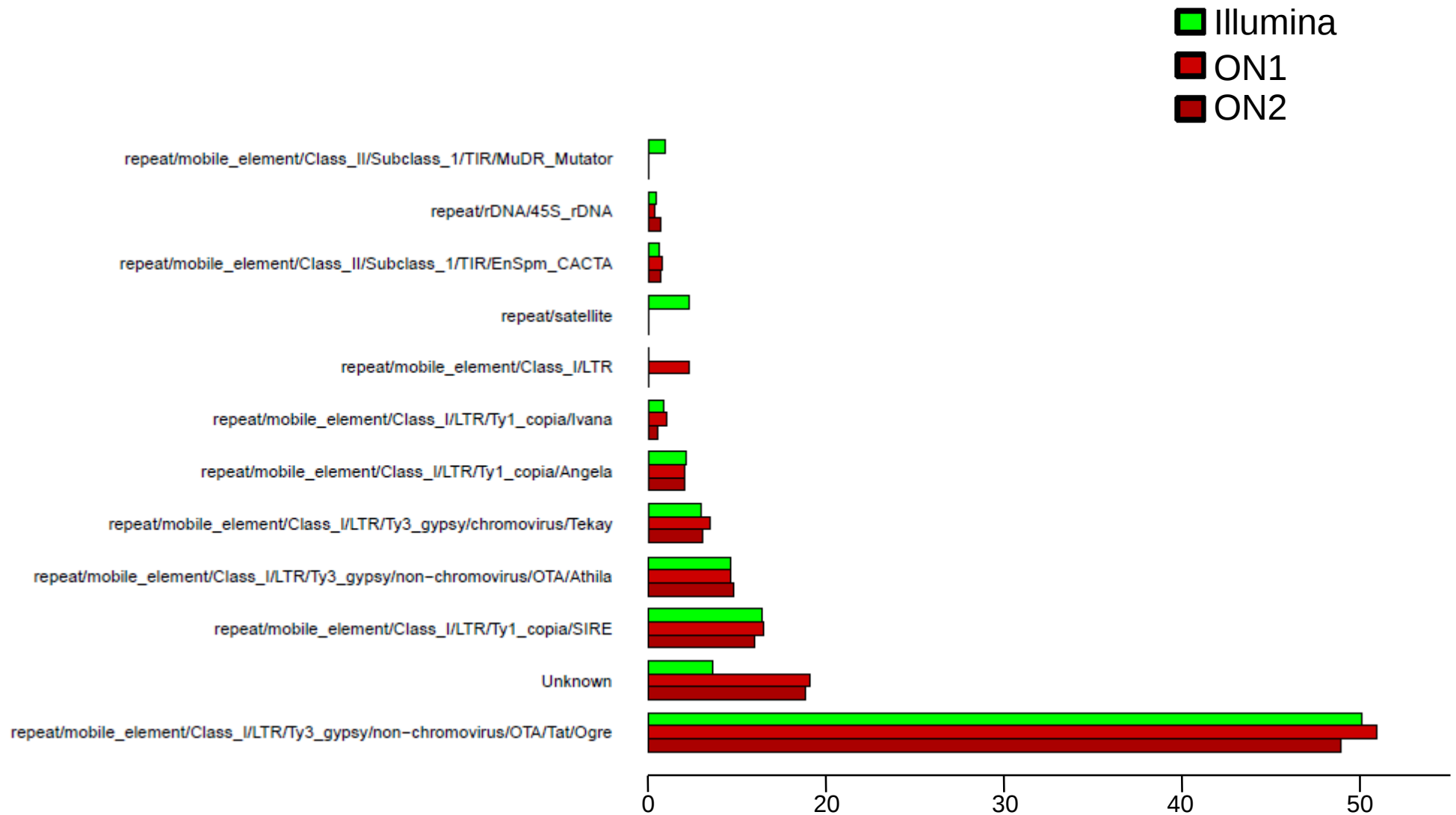
Oxford Nanopore



0.43%

- 45S_rDNA/25S_rDNA
- 45S_rDNA/18S_rDNA
- 45S_rDNA/5.8S_rDNA

Using ultralong reads in RepeatExplorer clustering



Using ultralong reads in RepeatExplorer clustering

Limitations:

- Currently less sensitive
- Depends on base-calling algorithm
- Problematic detection of satellites using TAREAN module (kmer based approach does not work well on erroneous reads)

Potential advantages:

Identification of repetitive element variants

Identification of long range associations