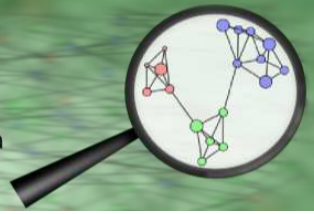
RepeatExplorer

Discover repeats in your next generation sequencing data



RepeatExplorer pipeline – algorithms and implementation

5th Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis in Plants

May 24-26, 2016

Petr Novak

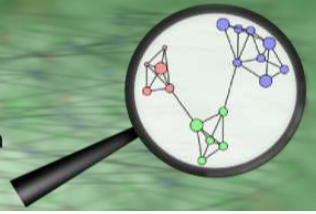
Laboratory of Molecular Cytogenetics, Biology Centre CAS, České Budějovice, Czech Republic

elixir CZECH REPUBLIC

BIOLOGY CENTRE CAS

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)*

Available Tools:

NGS data preprocessing

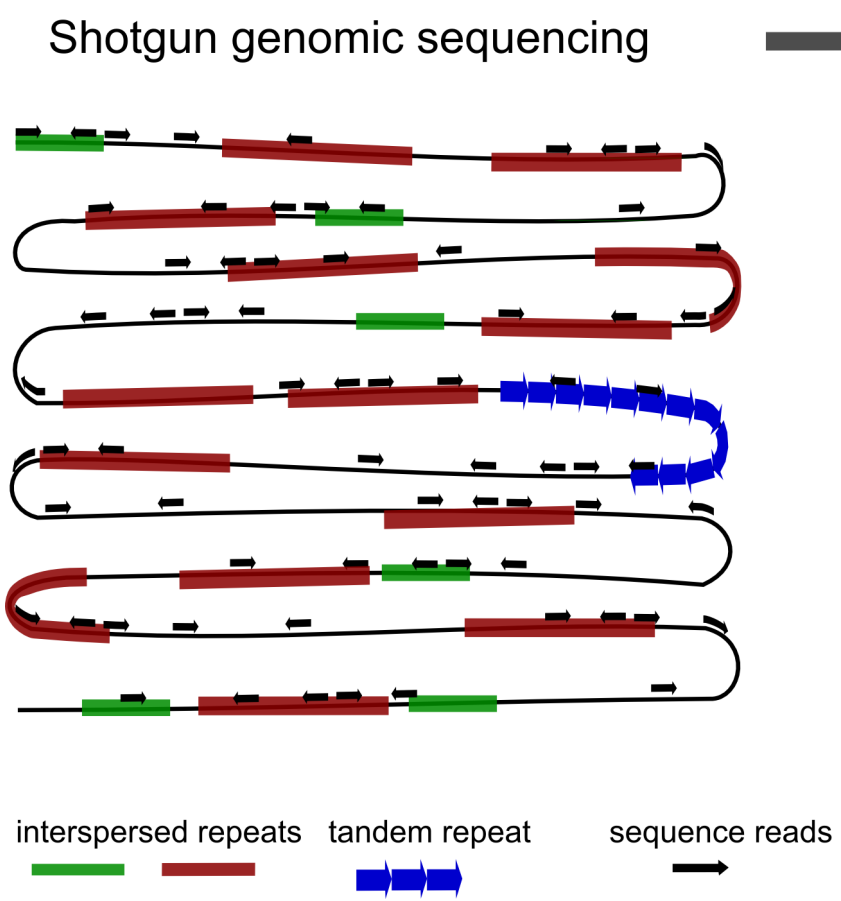
- Graph-based clustering
- Characterization of repeats:
    - Identification
    - Annotation
    - Quantification
    - ChIP Seq analysis
    - Satellite identification

Contributors:

Jiri Macas
Pavel Neumann
Jaroslav Steinhaisel
Jiri Pech
Karsten Klein
Georg Hermanutz
Petr Novak

RepeatExplorer

Discover repeats in your next generation sequencing data



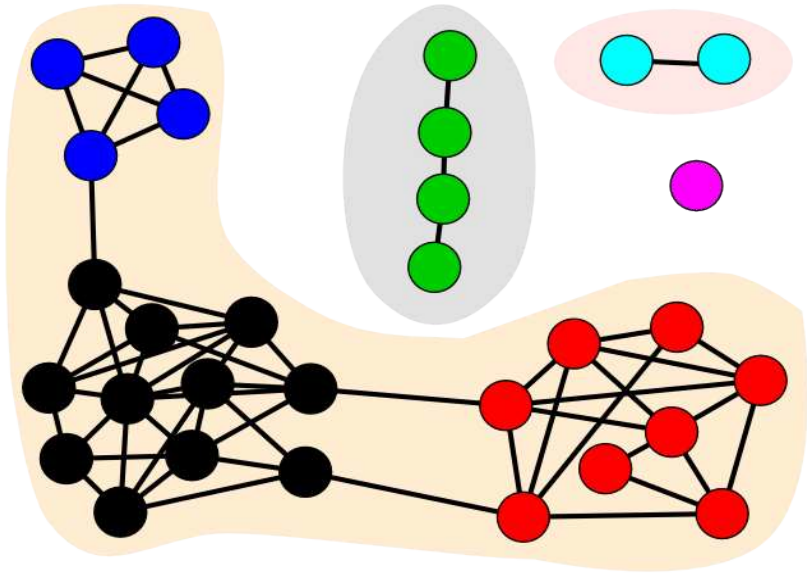
Low coverage sequencing

CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

Why to use **graph representations** :

Available algorithms for graph structure analysis

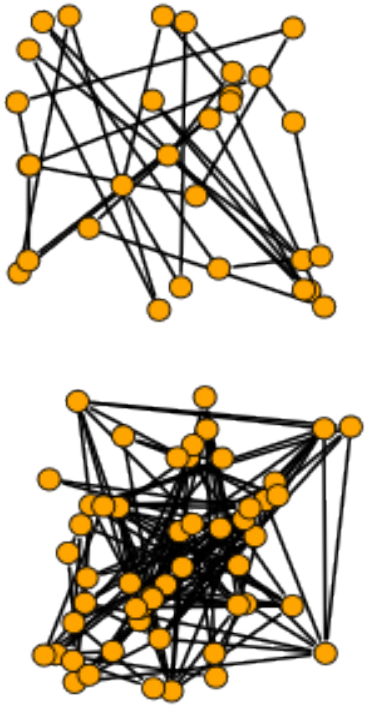
Robust partitioning/classification of reads based on mutual similarities



Why to use **graph representations** :

Alternative visualization of DNA sequence clusters

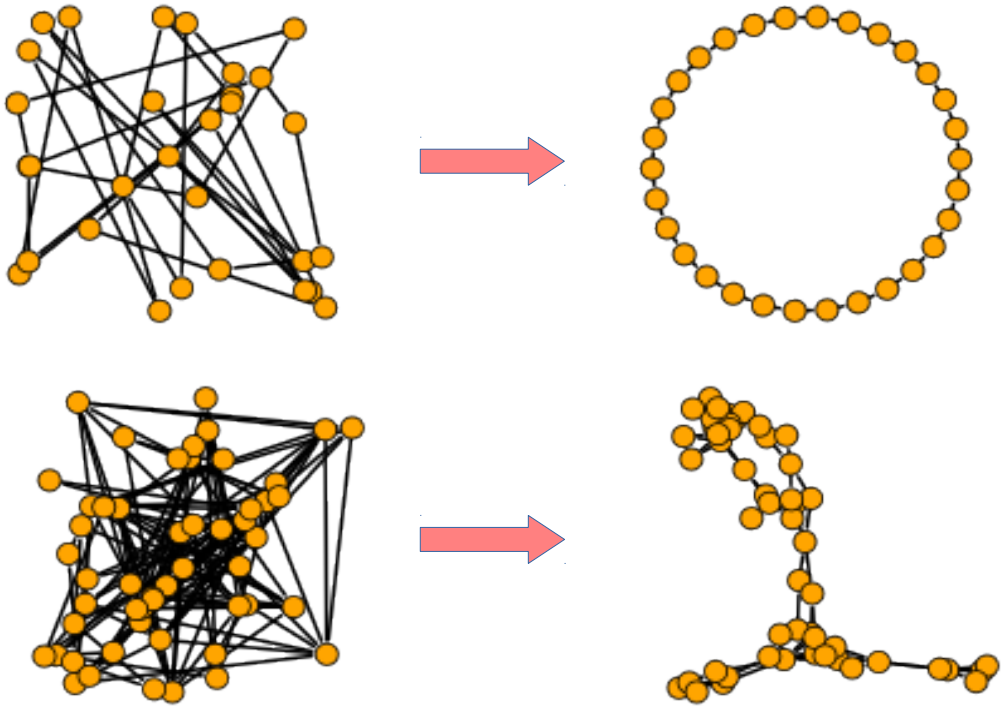
Suitable layout to enable intuitive analysis of clusters



Why to use **graph representations** :

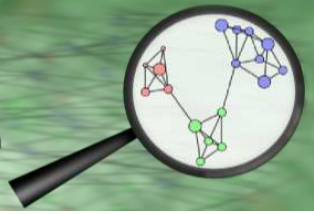
Alternative visualization of DNA sequence clusters

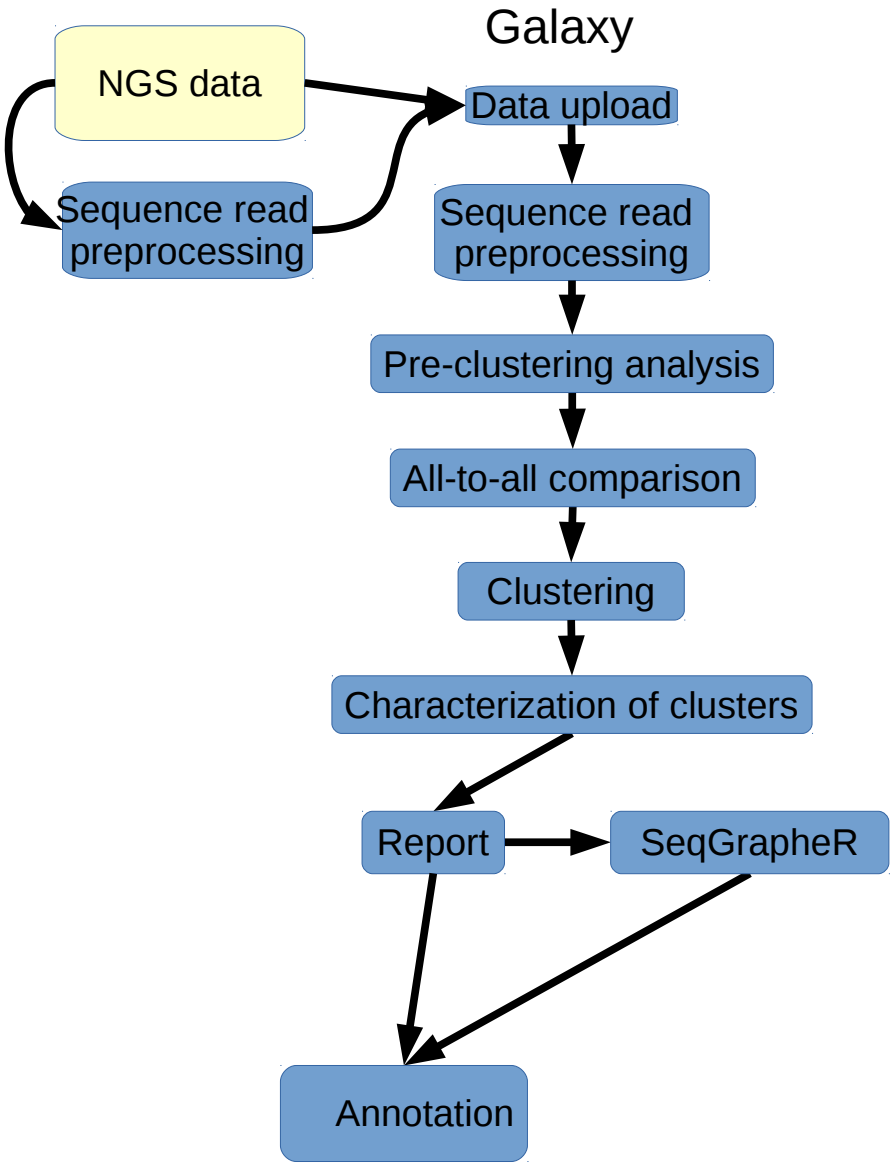
Suitable layout to enable intuitive analysis of clusters



RepeatExplorer

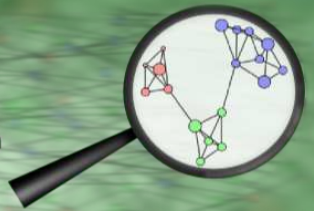
Discover repeats in your next generation sequencing data

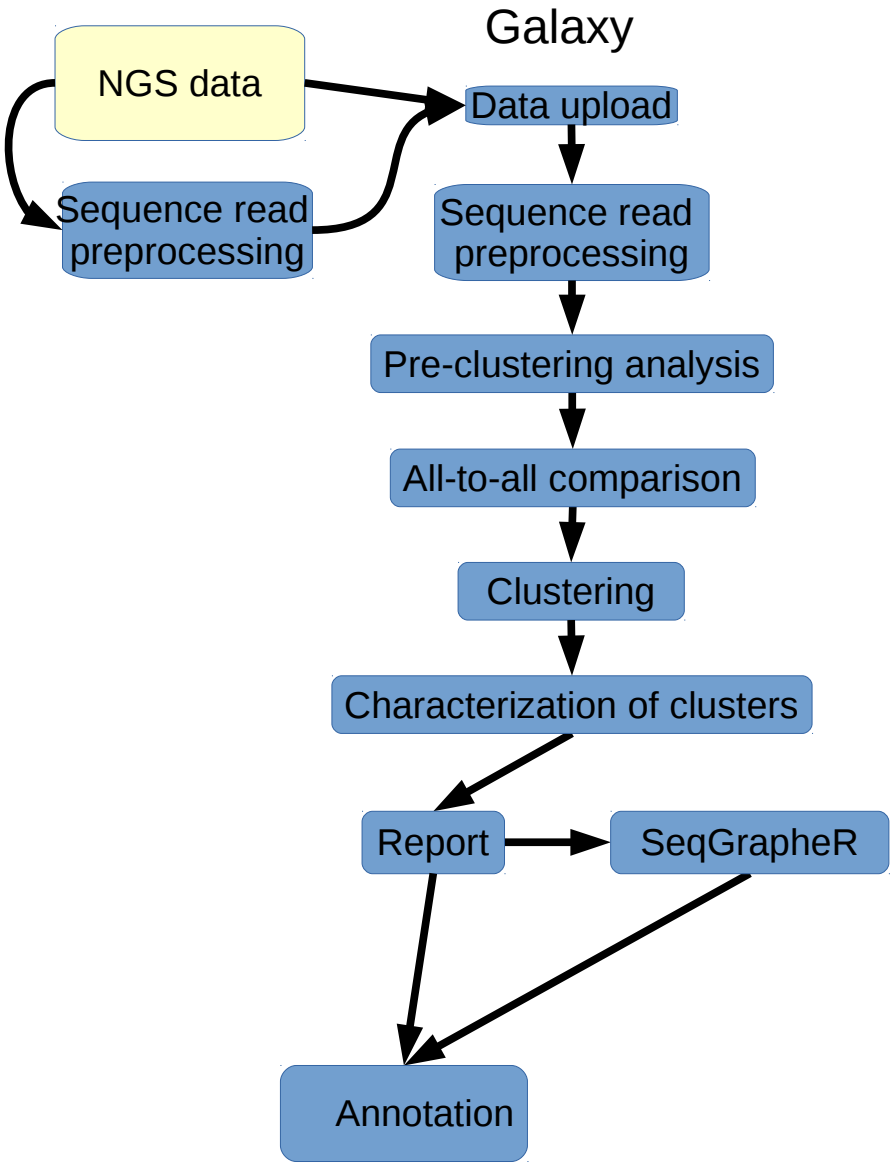




RepeatExplorer

Discover repeats in your next generation sequencing data





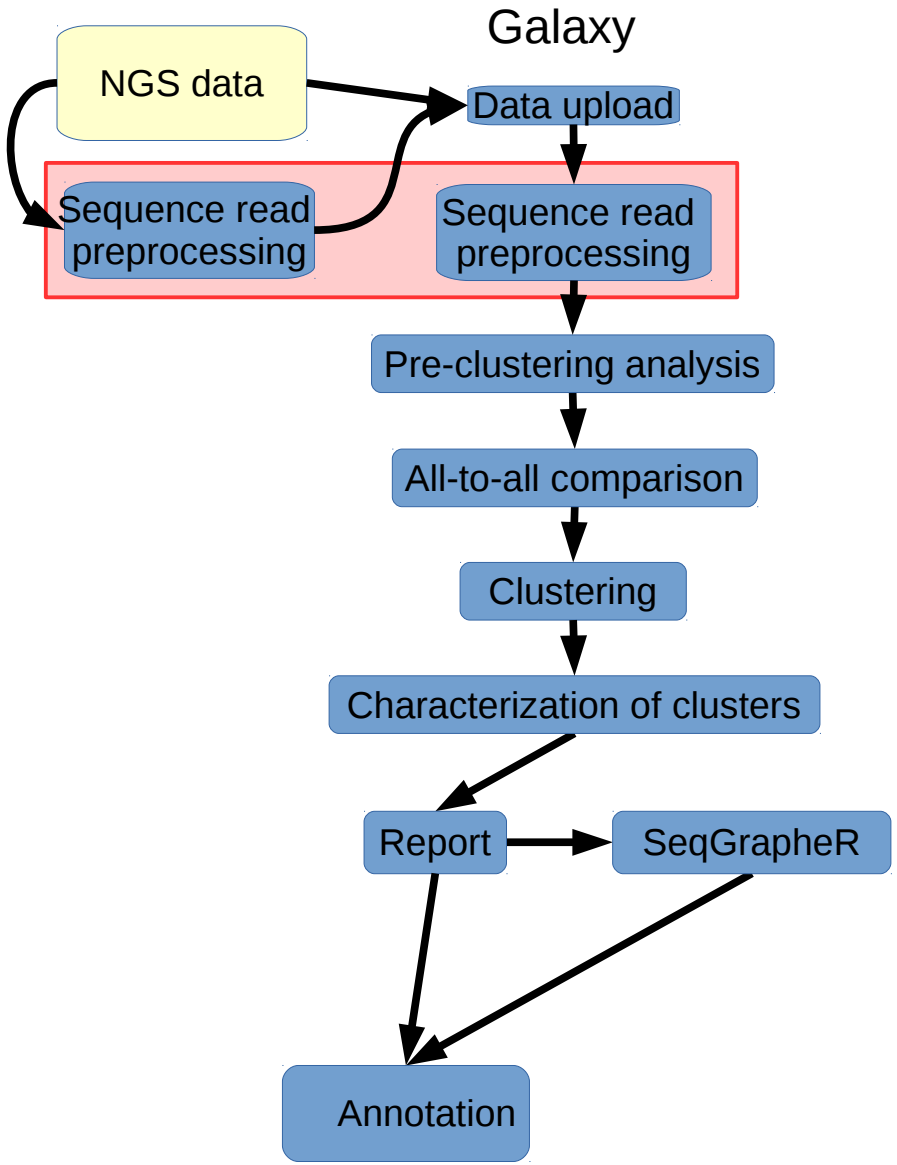
Input data requirements:

Sequence read:

- 100 nt or more
- Paired (optional)
- Uniform length

RepeatExplorer

Discover repeats in your next generation sequencing data



Preprocessing:

The most important step!

Local x Remote (Galaxy)

Amount of data:

10^5 – 10^7 reads

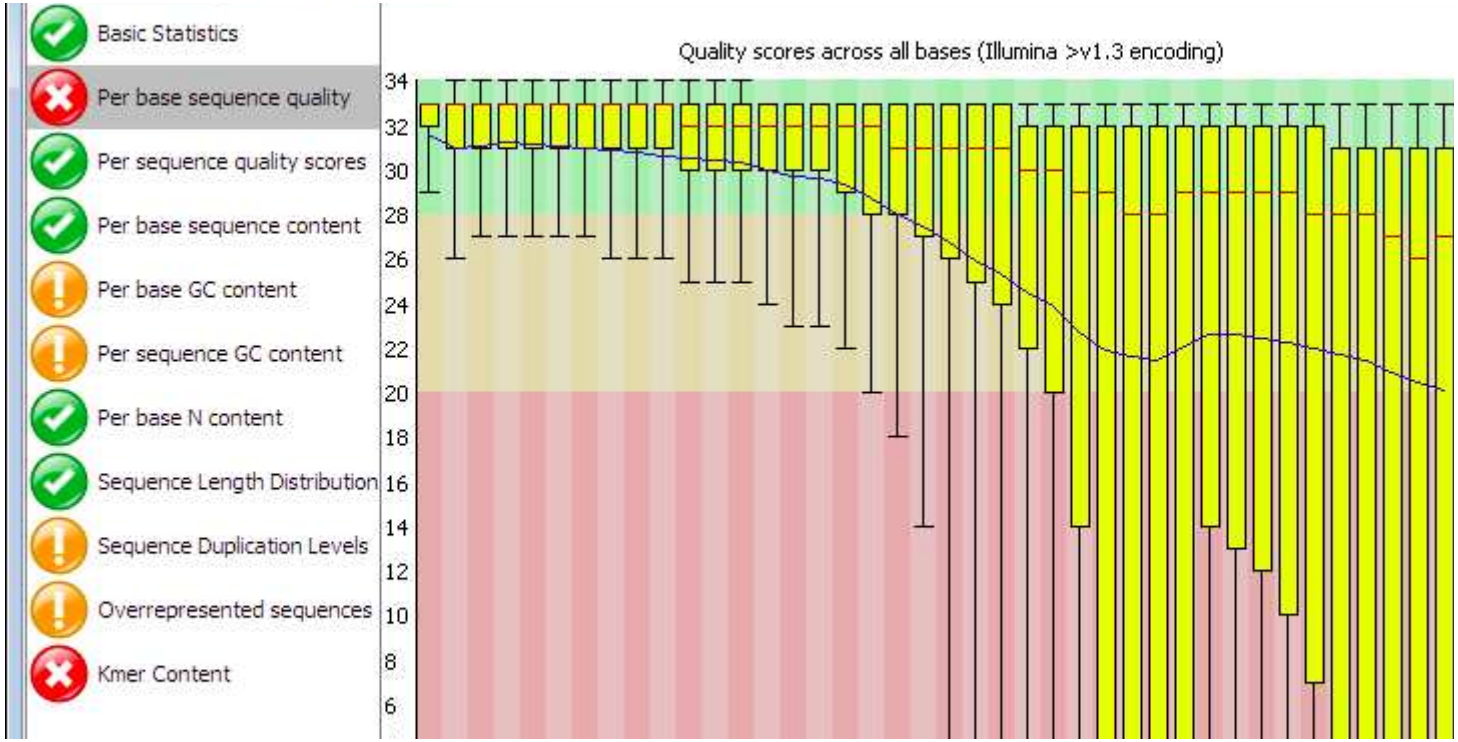
Test runs on small sample – validation of your dataset

Preprocessing

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

FastQC

- 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

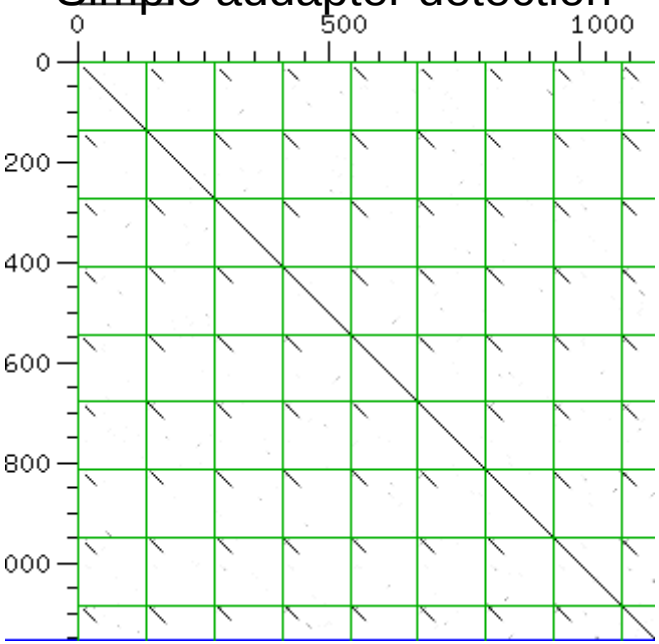
Back to summary

Preprocessing

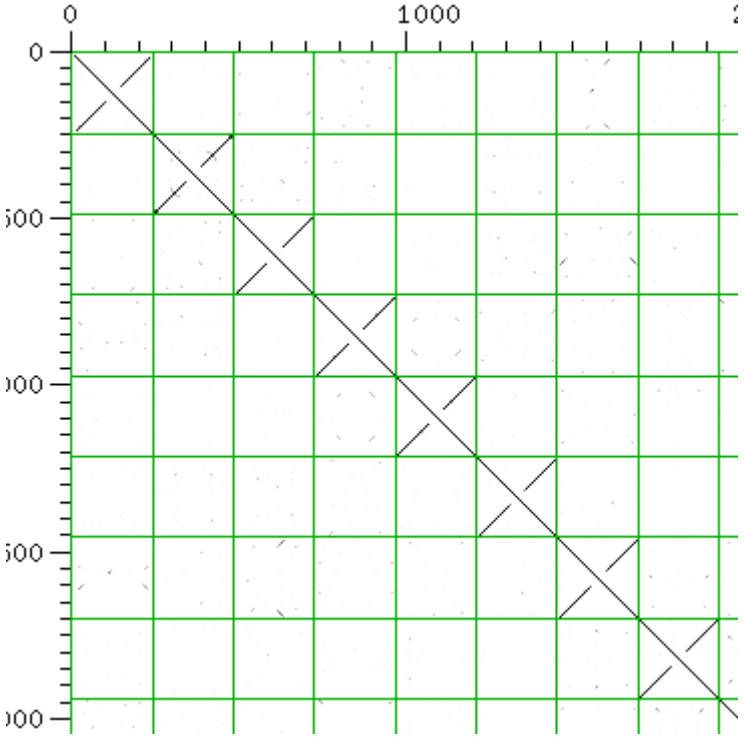
- **Quality control**
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identificators

Dotter - graphical dotplot program for detailed comparison of two sequences

Simple addapter detection



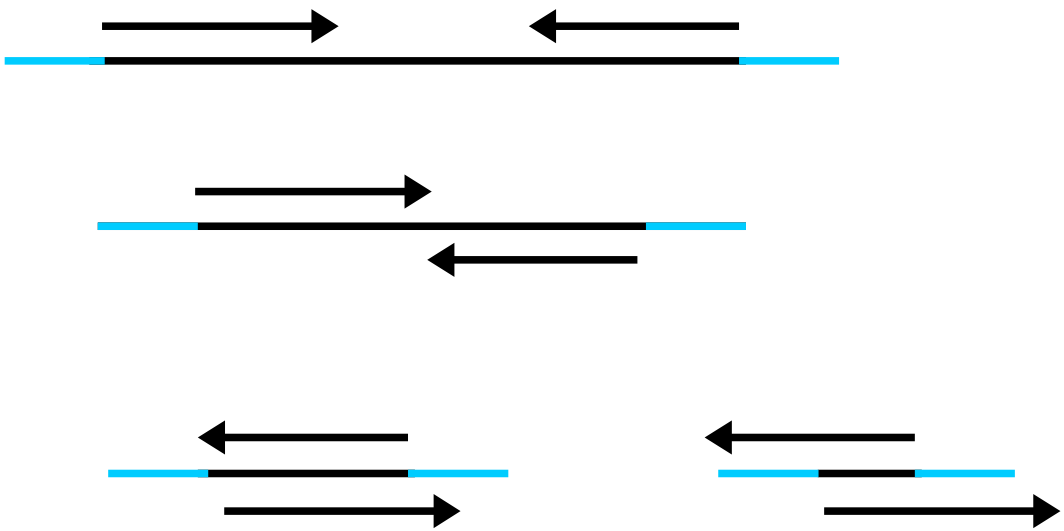
Concatenated reads!



Preprocessing

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

Effect of paired-end reads quality, overlap



Preprocessing

- 100 nt or more
- Paired
- Uniform length
- Quality control
- **Trimming, filtering, adapter removing**
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identificators

cutadapt - https://cutadapt.readthedocs.io

A tool that removes adapter sequences from high througput sequence reads

- command line
- available in galaxy

- Sequence trimming
  or
- Complete removal

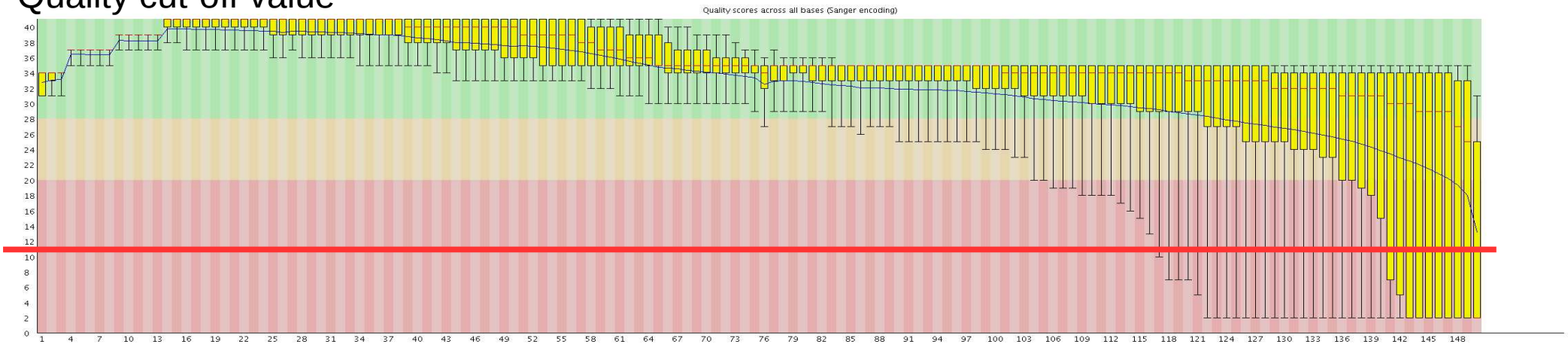
Presence of adapter greatly affect all-to-all step duration

Preprocessing

- 100 nt or more
- Paired
- Uniform length
- Quality control
- **Trimming, filtering, adapter removing**
- Convert fastq to fasta
- Interlacing, sampling
- Modification of sequence identificators

Quality filtering criteria – FASTX-toolkit

Quality cut-off value



Percent of bases in sequence that must have quality equal to / higher than cut-off value

Proportion of Ns

We can afford stringent filtering - most of the time we have more data than needed

Preprocessing

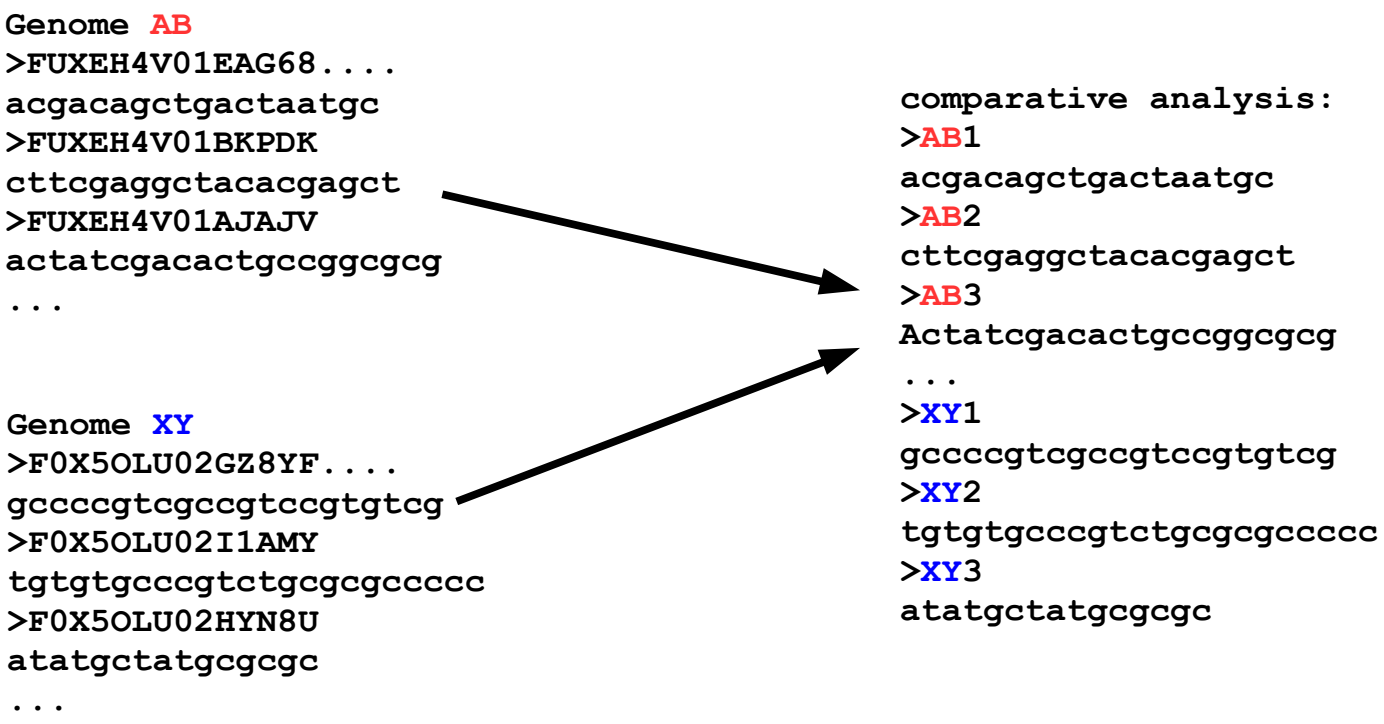
- 100 nt or more
- Paired
- Uniform length
- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- **Interlacing, sampling**
- **Modification of sequence identificators**

```
>HWI-ST863:57:D037GACXX:7:1101:1358:2087 1:N:0:GGCTAC
acgacagctgactaatgc
>HWI-ST863:57:D037GACXX:7:1101:1372:2162 1:N:0:GGCTAC
cttcgaggctacacgagct
>HWI-ST863:57:D037GACXX:7:1101:1484:2168 1:N:0:GGCTAC
actatcgacactgccggcgcg
```

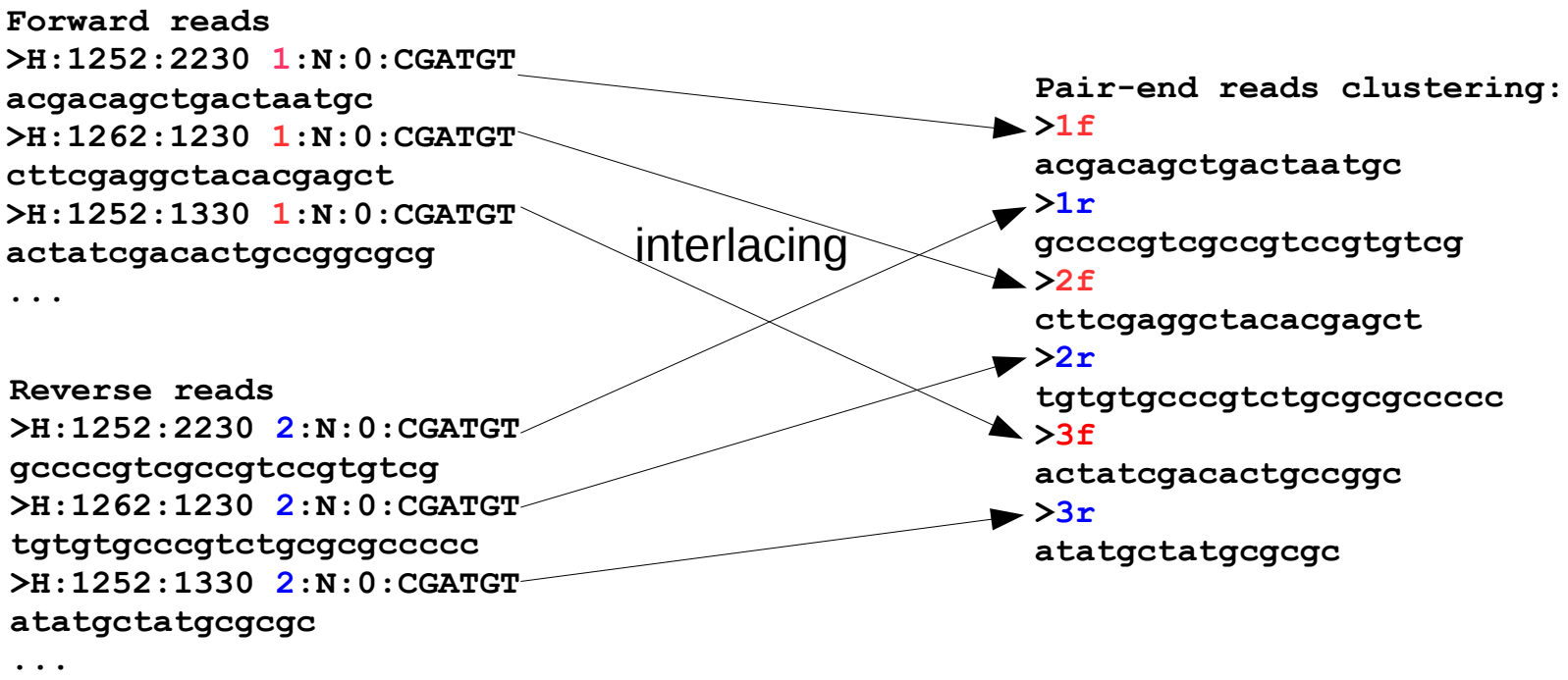
```
>1
acgacagctgactaatgc
>2
cttcgaggctacacgagct
>3
actatcgacactgccggcgcg
```

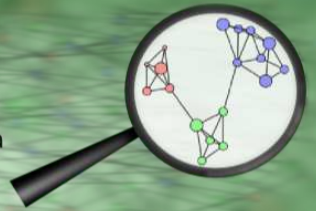
Simple names → smaller files!

Comparative analysis of multiple genomes



Pair-end reads





Preprocessing

- 100 nt or more
- Paired
- Uniform length
- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- **Interlacing, sampling**
- **Modification of sequence identificators**

Fastq interlacer – Galaxy build tool

- Does not expect ordered sequences
- Memory demanding

Fasta interlacer – RepeatExplorer tool

- Assumes ordered sequences
- Low memory requirements
- Fast

Original Illumina reads are always ordered!

RepeatExplorer

Discover repeats in your next generation sequencing data

Preprocessing

- 100 nt or more
- Paired
- Uniform length
- Quality control
- Trimming, filtering, adapter removing
- Convert fastq to fasta
- **Interlacing, sampling**
- **Modification of sequence identificators**

Sampling

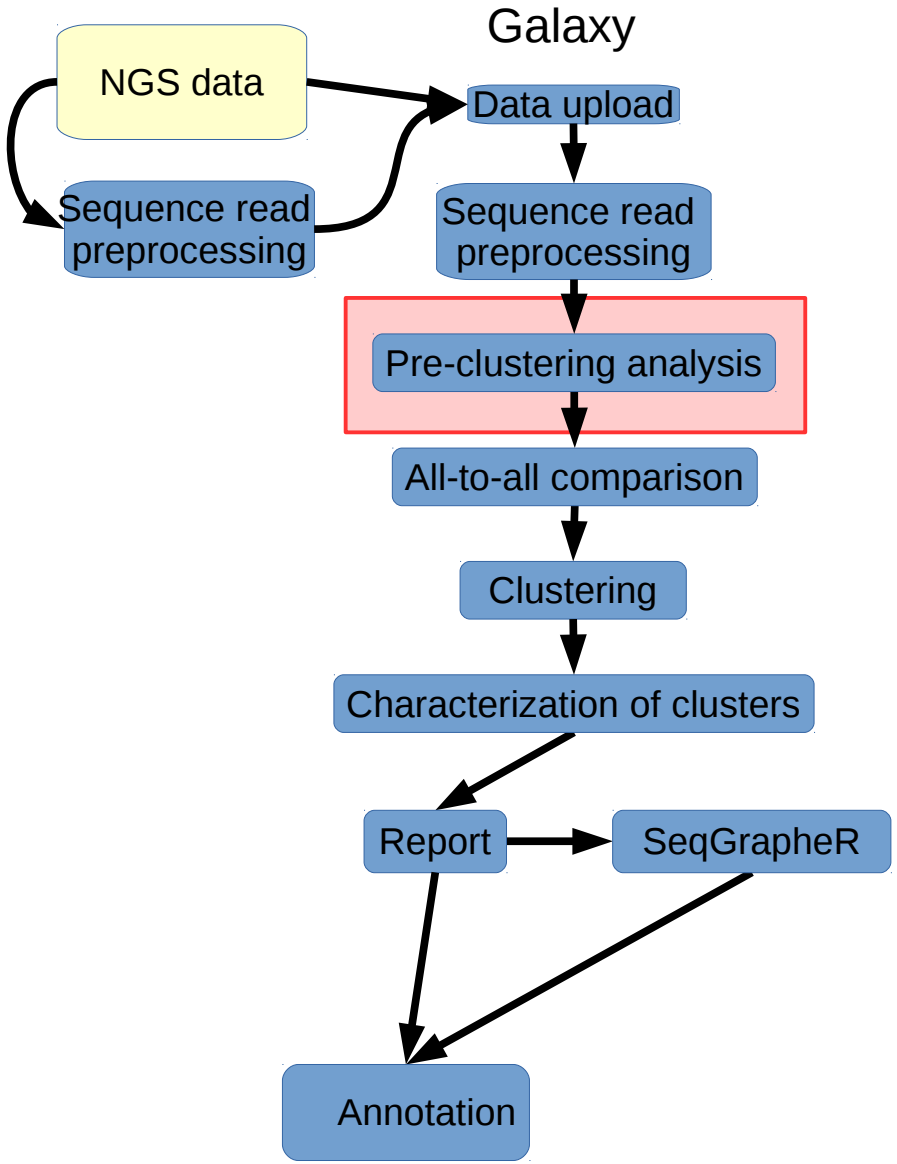
- Reproducible pseudo-random samples
- Single/paired reads
- Avoid using huge datasets
  50GB space limit on public RE server!

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

RepeatExplorer

Discover repeats in your next generation sequencing data



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

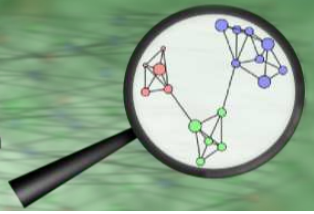
k_g genome specific coefficient - **graph density** depends on repetitive content and genome size

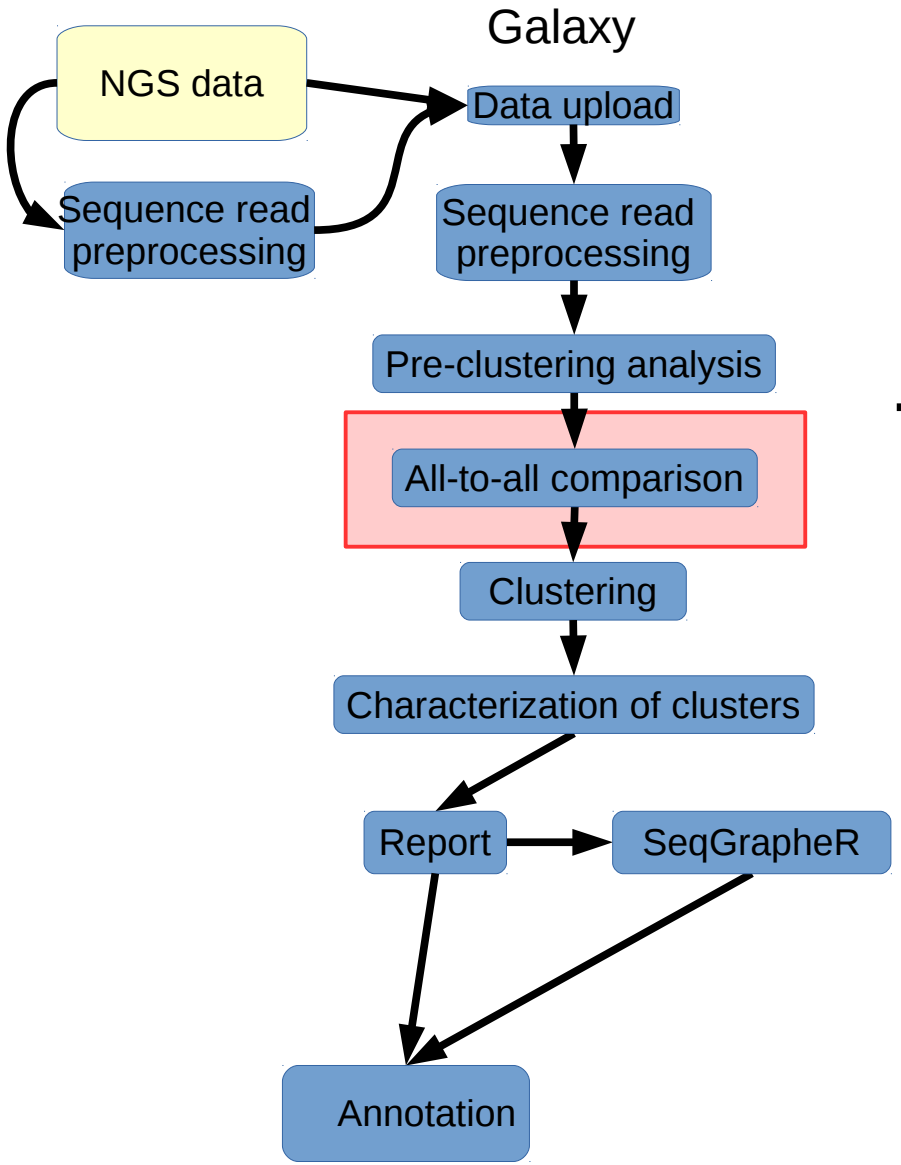
Density corresponds to probability that two randomly taken sequences from genome will be similar

k $_g$ is used to estimate maximum number of reads providing that **we can process ~ *340 ·10*^{6} of similarity hits on machine with 16GB of RAM**

RepeatExplorer

Discover repeats in your next generation sequencing data





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%

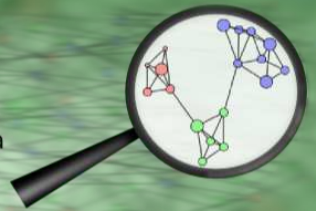
Stringency affects maximum number of reads!

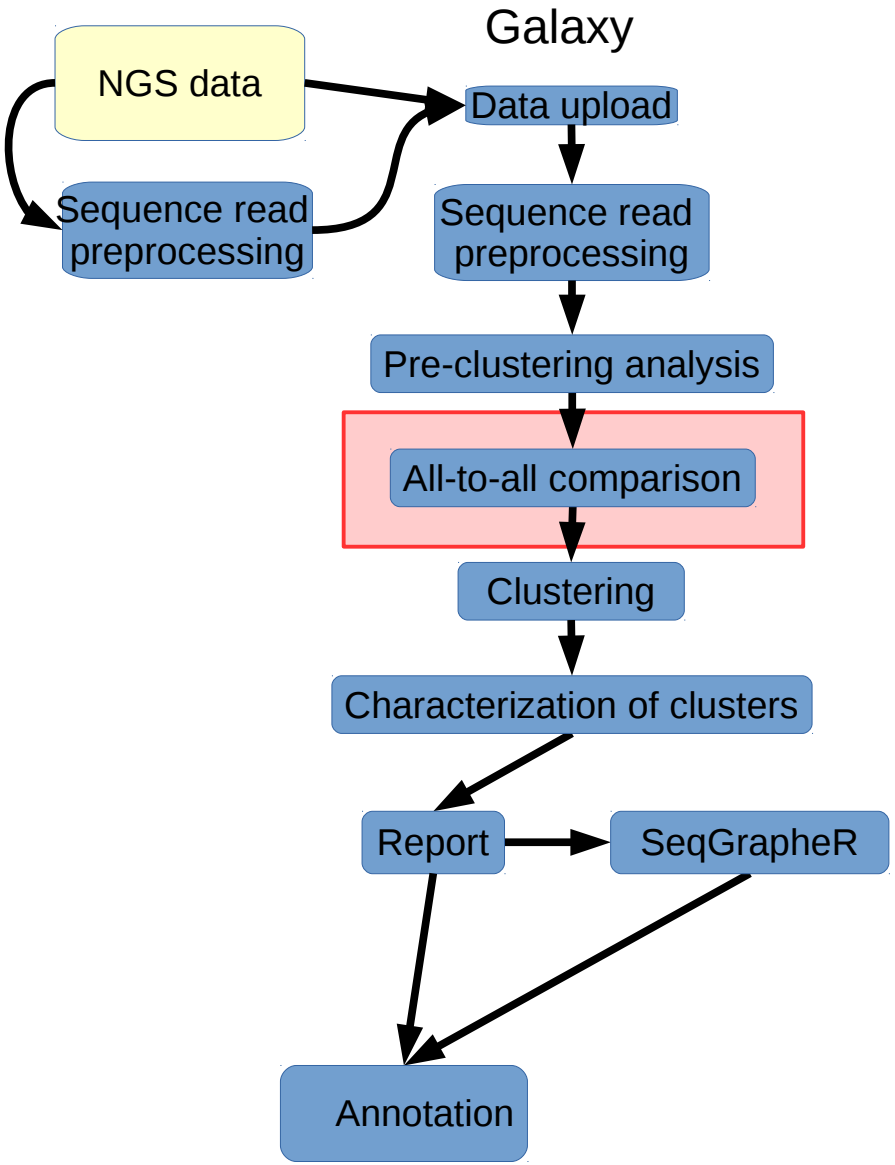
Length of overlap must be adjusted for reads shorter then 100 nt

55% of length and 90% similarity is hardcoded
Cannot be changed!

RepeatExplorer

Discover repeats in your next generation sequencing data

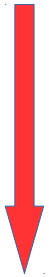




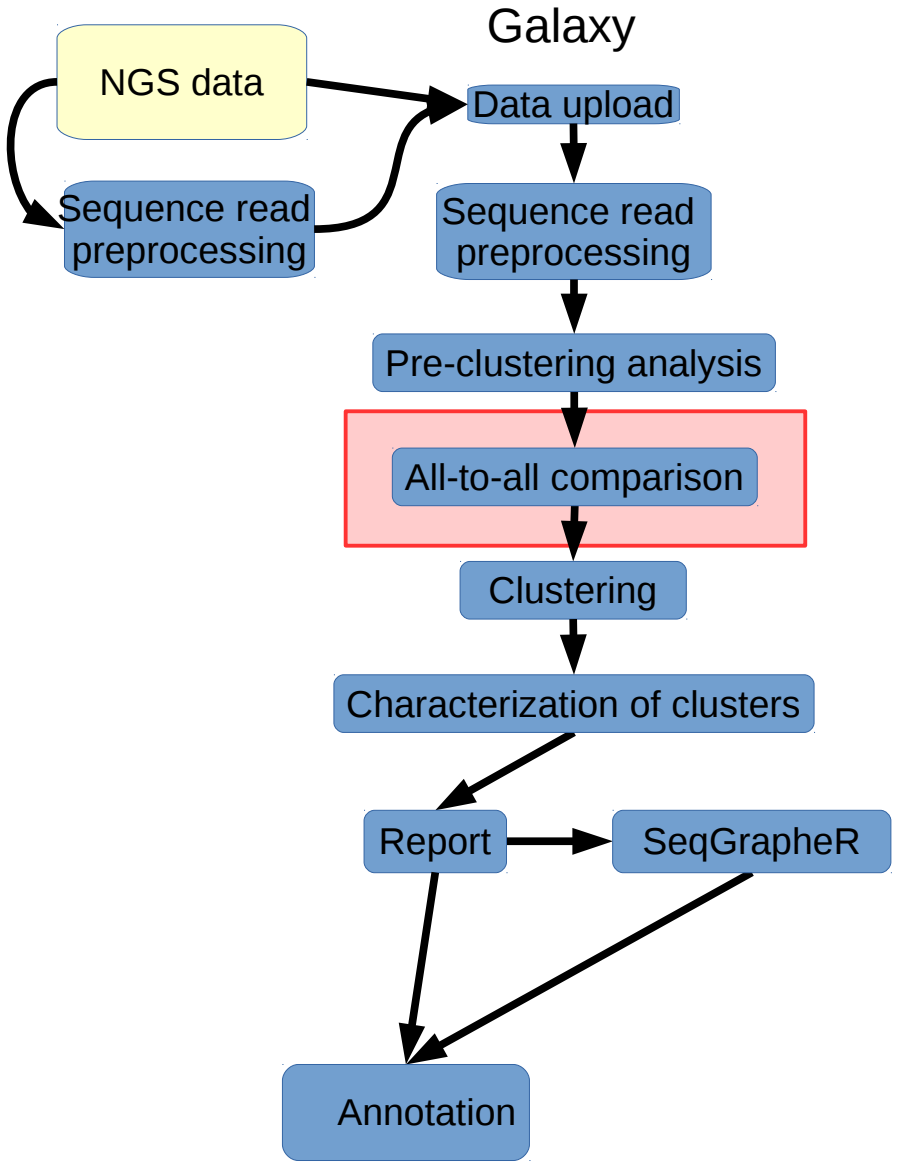
All-to-all comparison

All pairs of reads with similarity above threshold are found using megablast:

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



All-to-all comparison extremely slow!



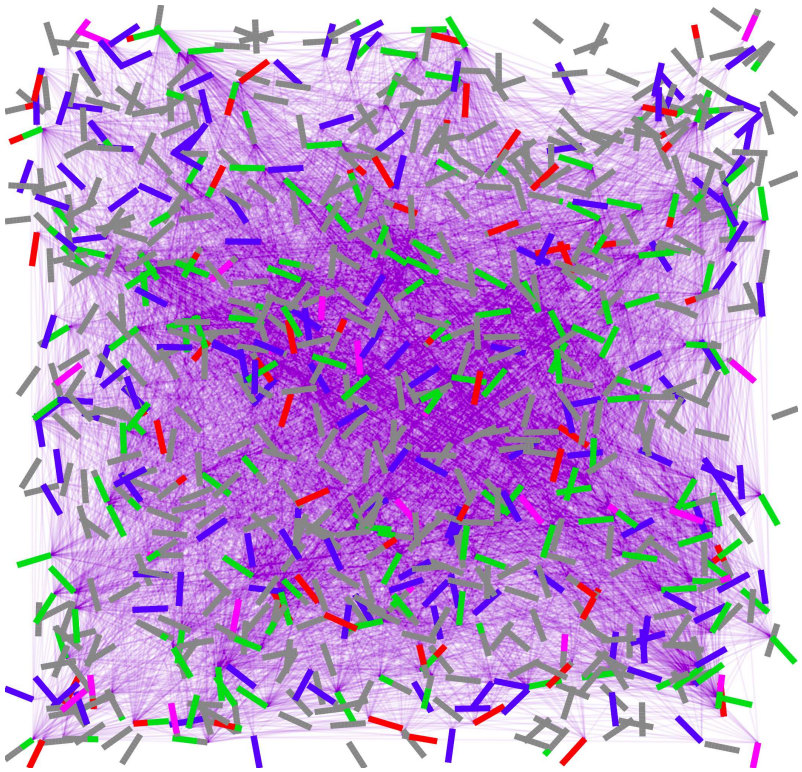
All-to-all comparison

Low complexity repeat – DustMasker

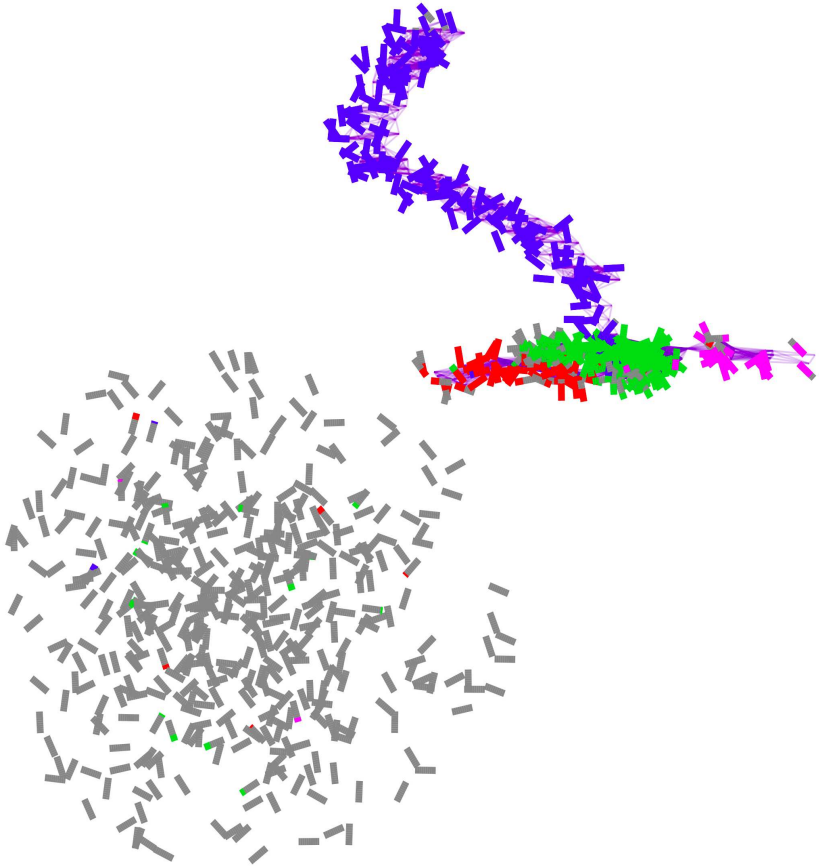
Simple repeats underestimated:
- Telomeric motifs
- microsatellites
…

RepeatExplorer

Discover repeats in your next generation sequencing data

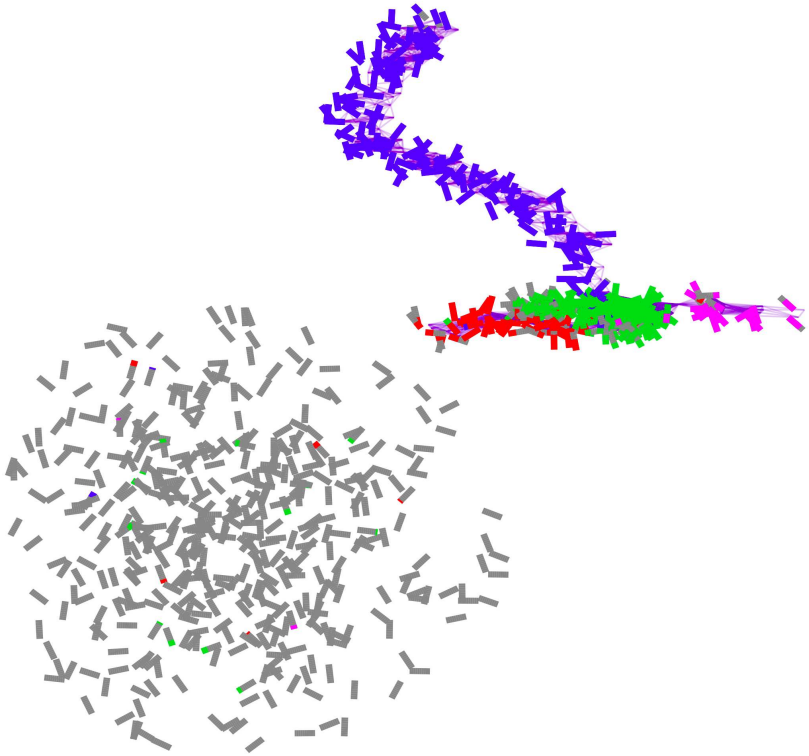


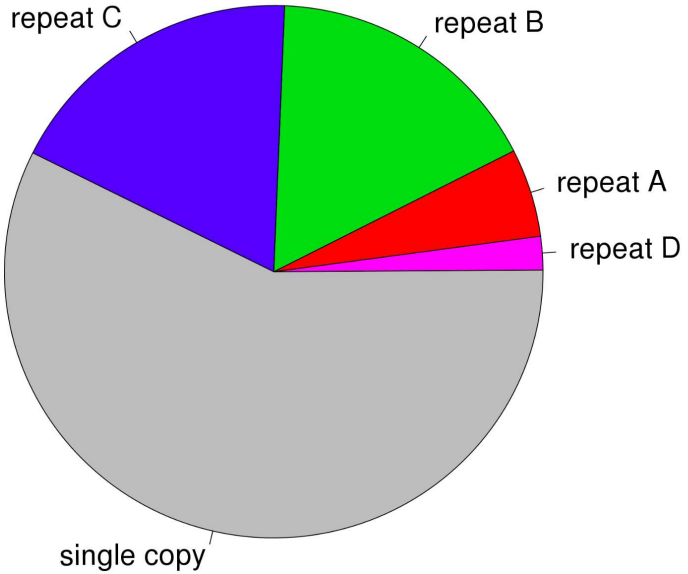
Clustering



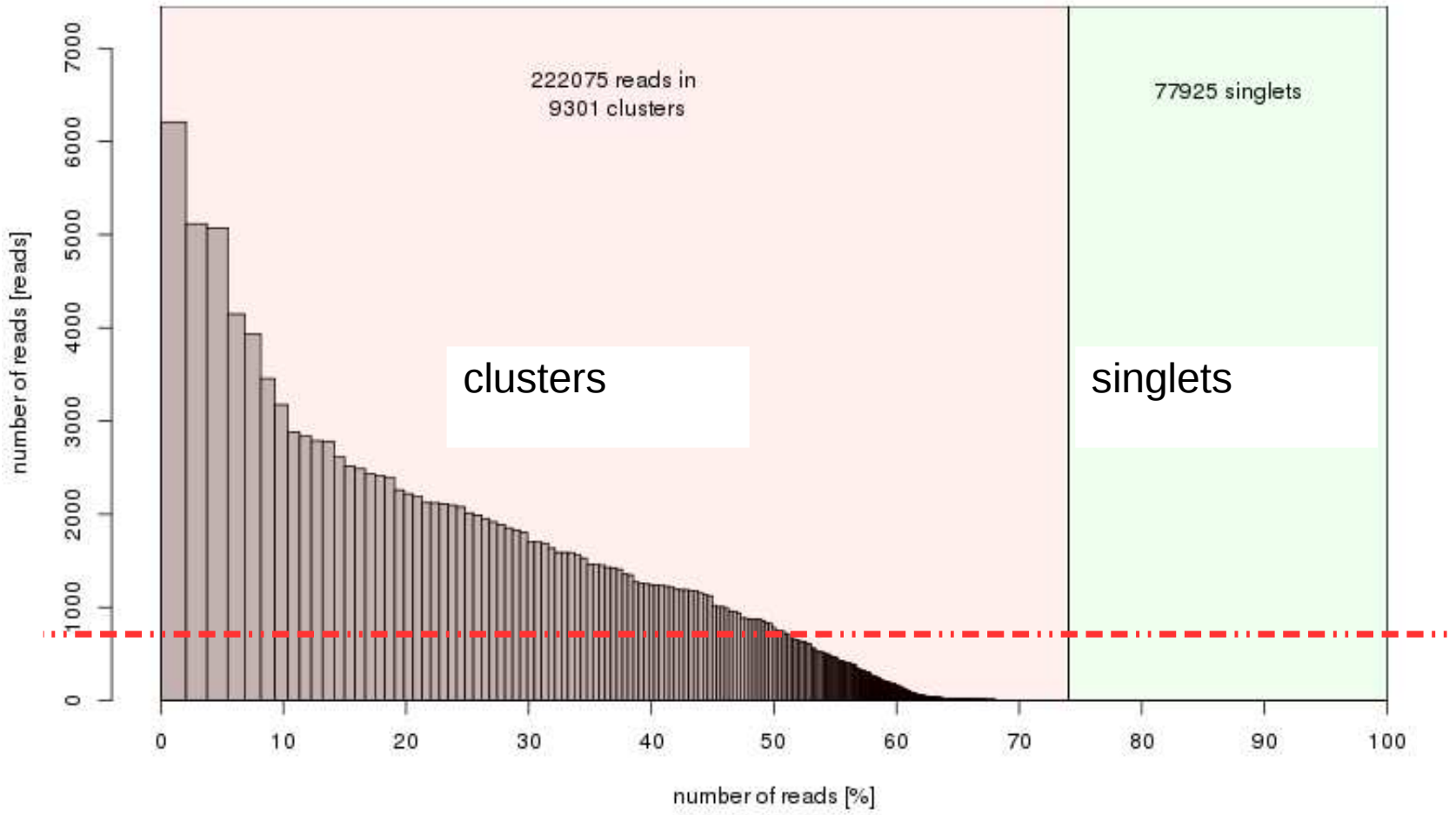
RepeatExplorer

Discover repeats in your next generation sequencing data





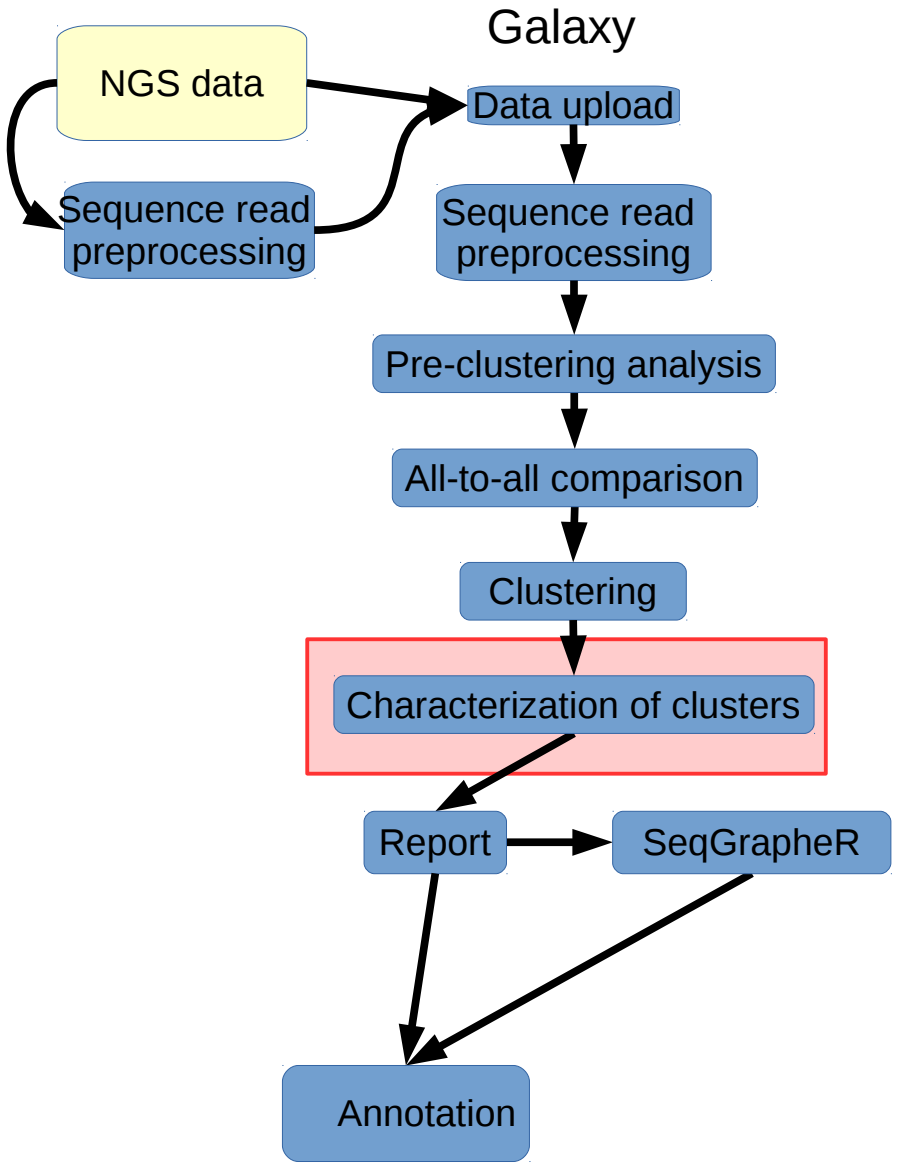
Size of clusters correspond to proportion of repeats in the genome



Typical cluster size distribution

RepeatExplorer

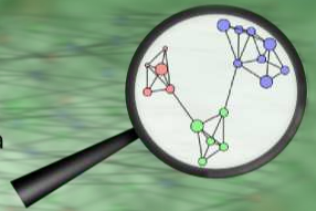
Discover repeats in your next generation sequencing data

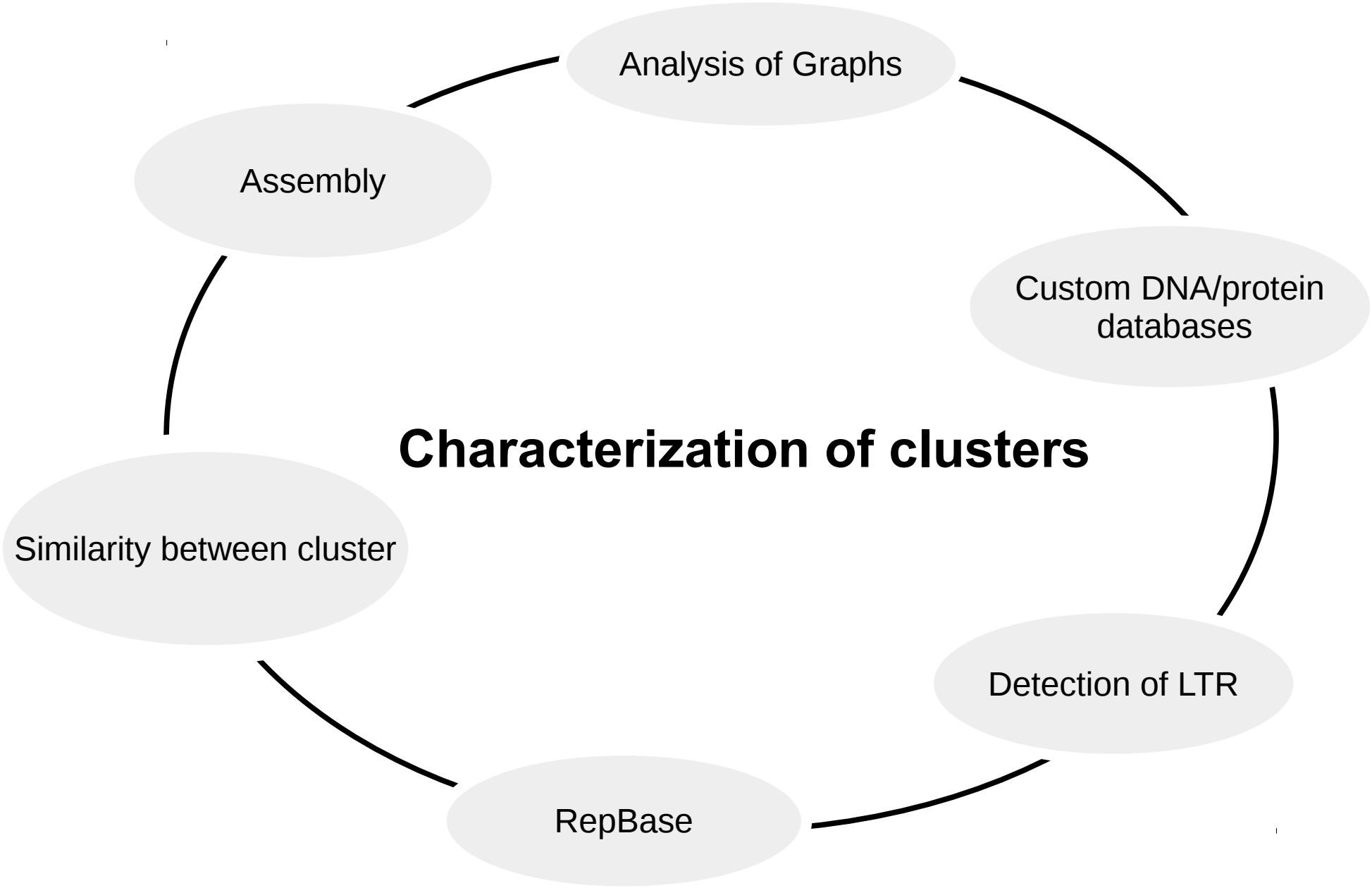


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

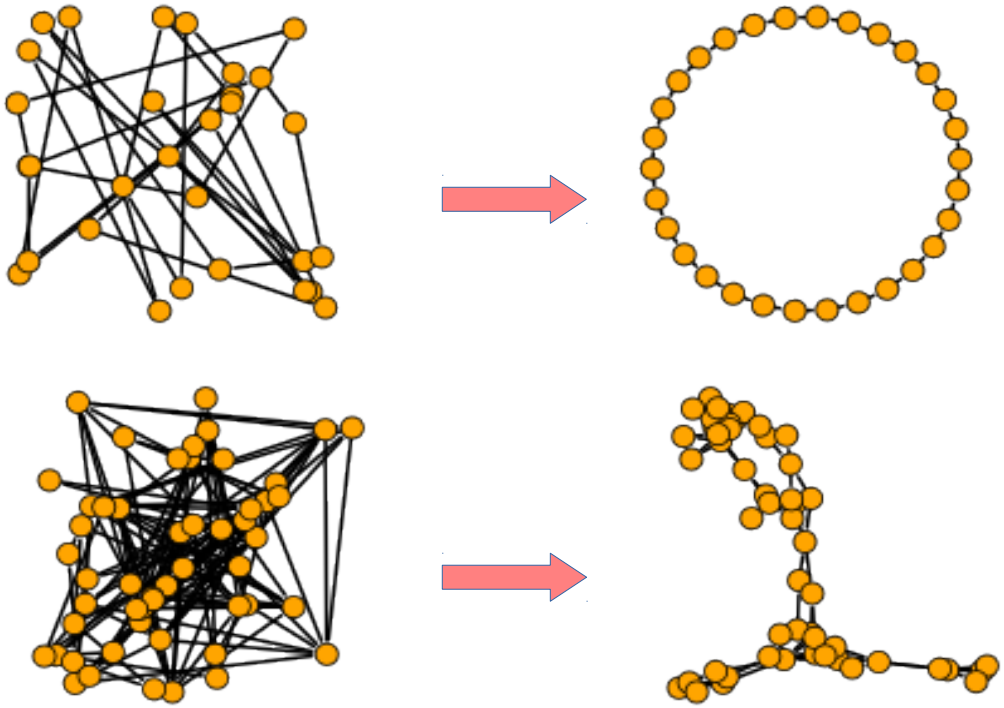
RepeatExplorer

Discover repeats in your next generation sequencing data

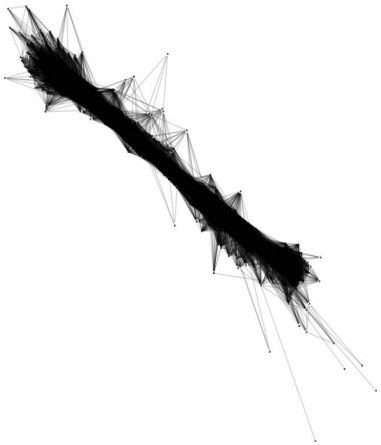


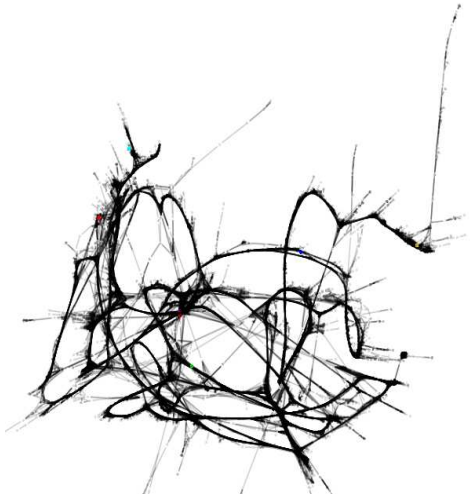


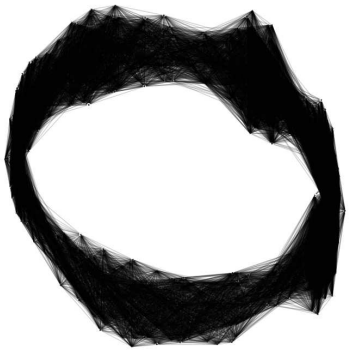
Analysis of graphs - layouts

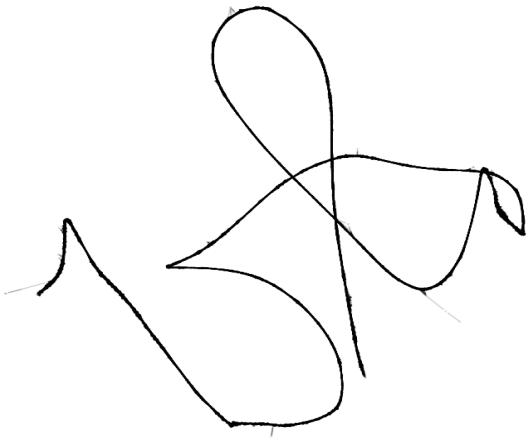


Analysis of graphs - layouts



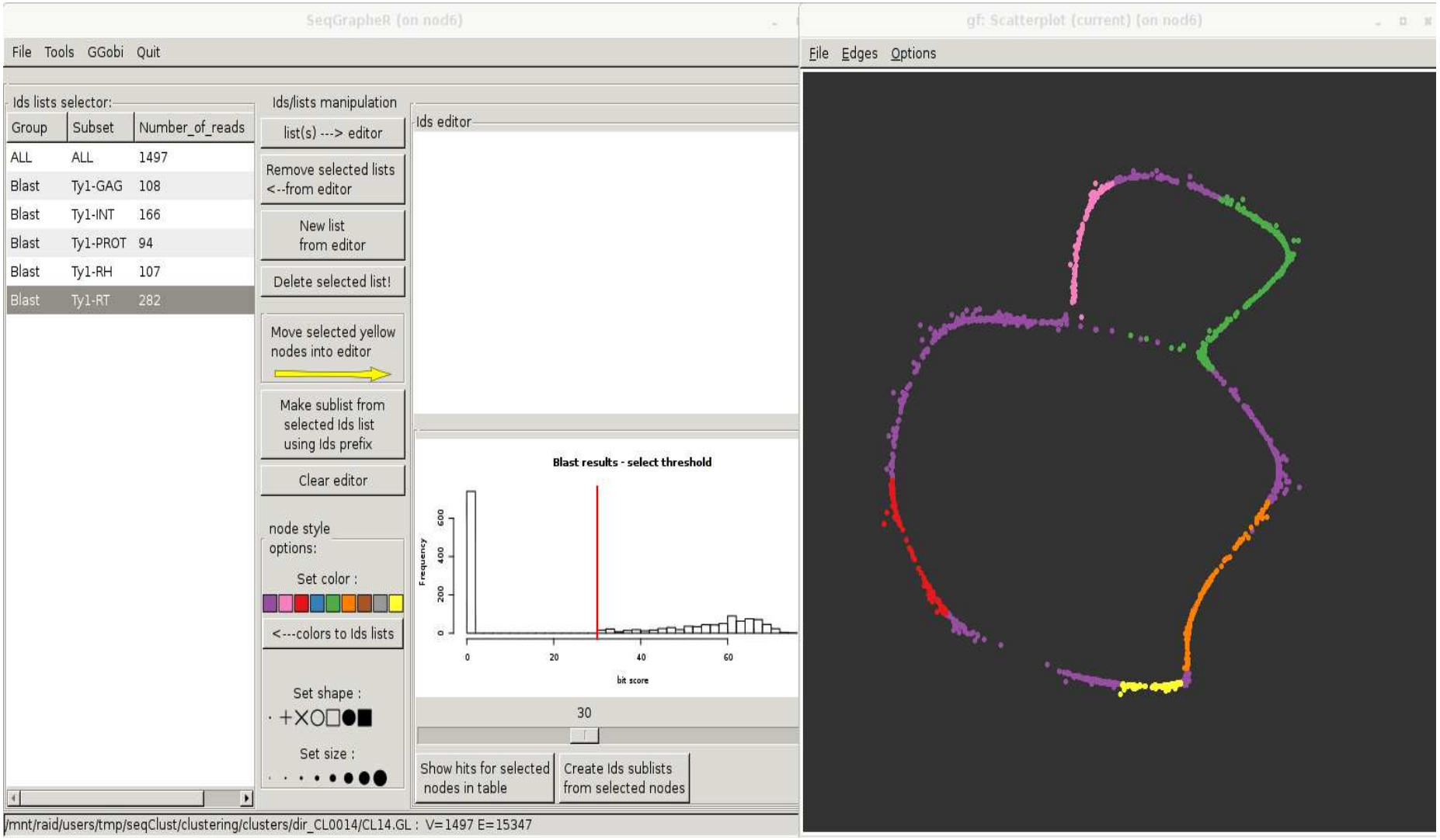






Analysis of graphs - layouts

SeqGrapheR – interactive exploration of graphs



RepeatMasker search:

Repeat Databases:

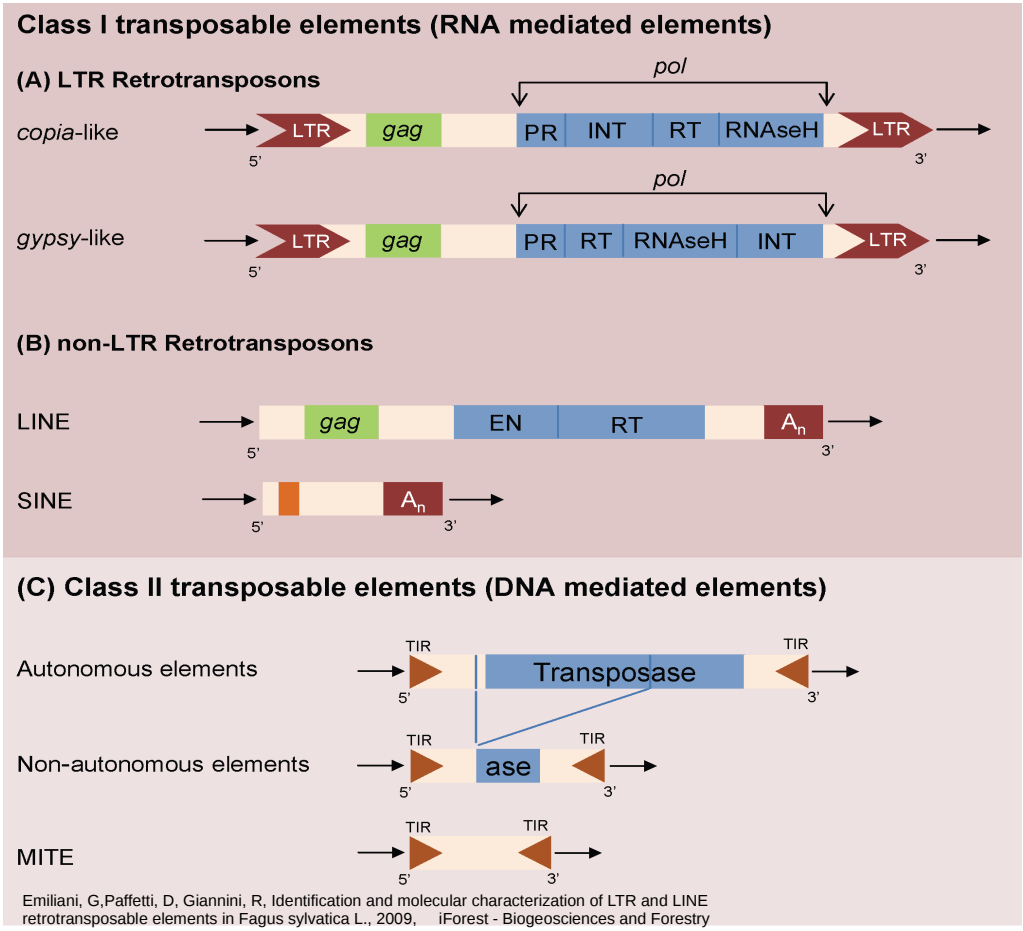
- RepBase (Default)

- Custom database (Optional)
Use specified format!

```
>reapeatname#class/subclass
acgatcgatcghagsctcgacagtcgatca
>reapeatname#class/subclass
acgatcgatcghagsctcgacagtcgatca
```

class.fammily	class	family	hits	hits[%]	content	Mean_Score
LTR/Copia	LTR	Copia	236	94.8	85.04	536
Low_complexity	Low_complexity	Low_complexity	3	1.2	0.25	22

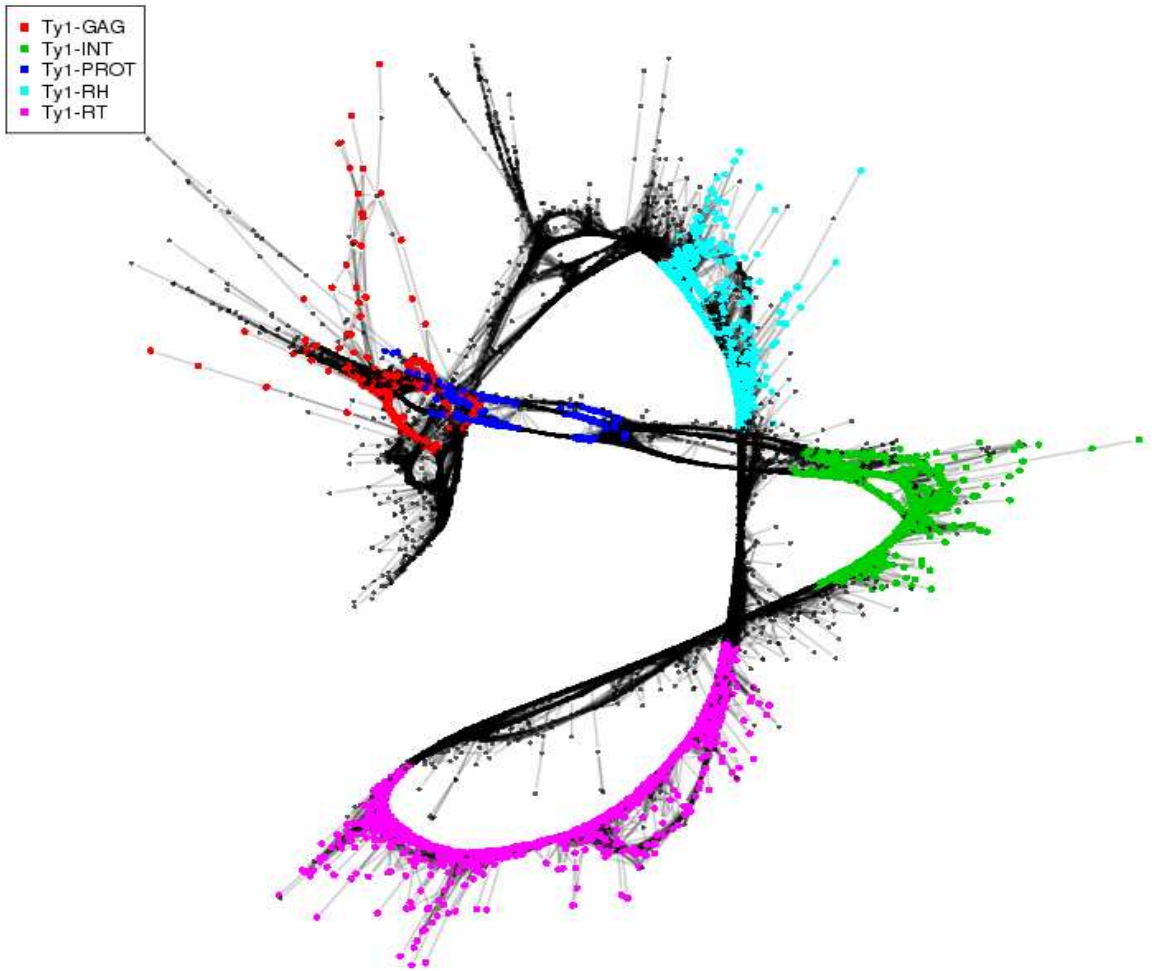
- Similarity search against **custom protein domain database** with blastx



Emiliani, G,Paffetti, D, Giannini, R, Identification and molecular characterization of LTR and LINE retrotransposable elements in Fagus sylvatica L., 2009, iForest - Biogeosciences and Forestry

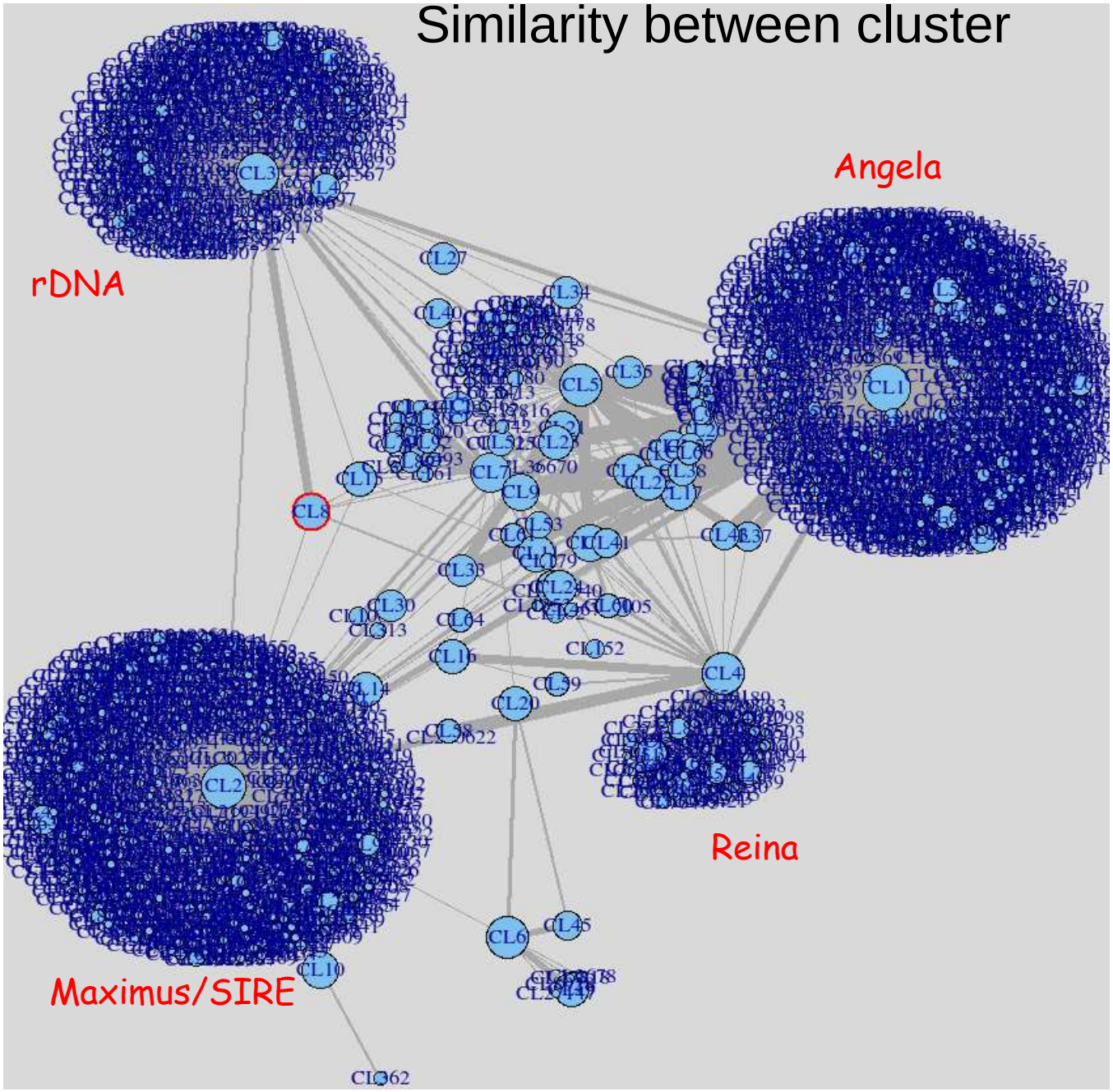
- Similarity search against **custom DNA database** (rDNA, plastid, mitochondria, contamination)

Protein domain database + Graph layouts

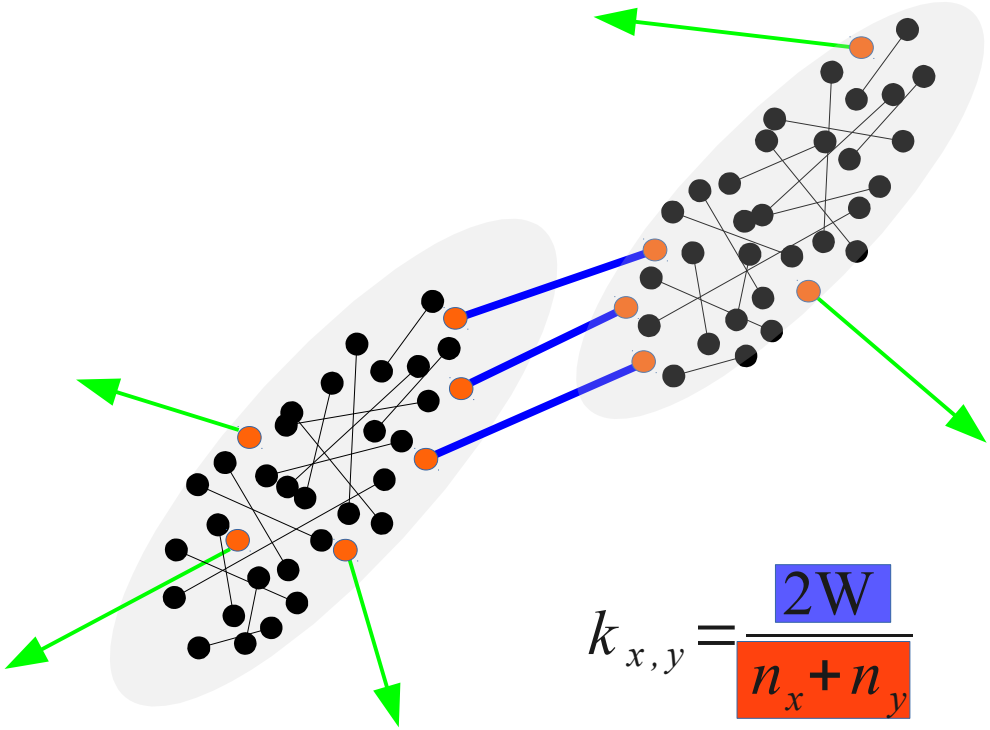


RepeatExplorer

Discover repeats in your next generation sequencing data



Similarity between cluster - presence of paired reads



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

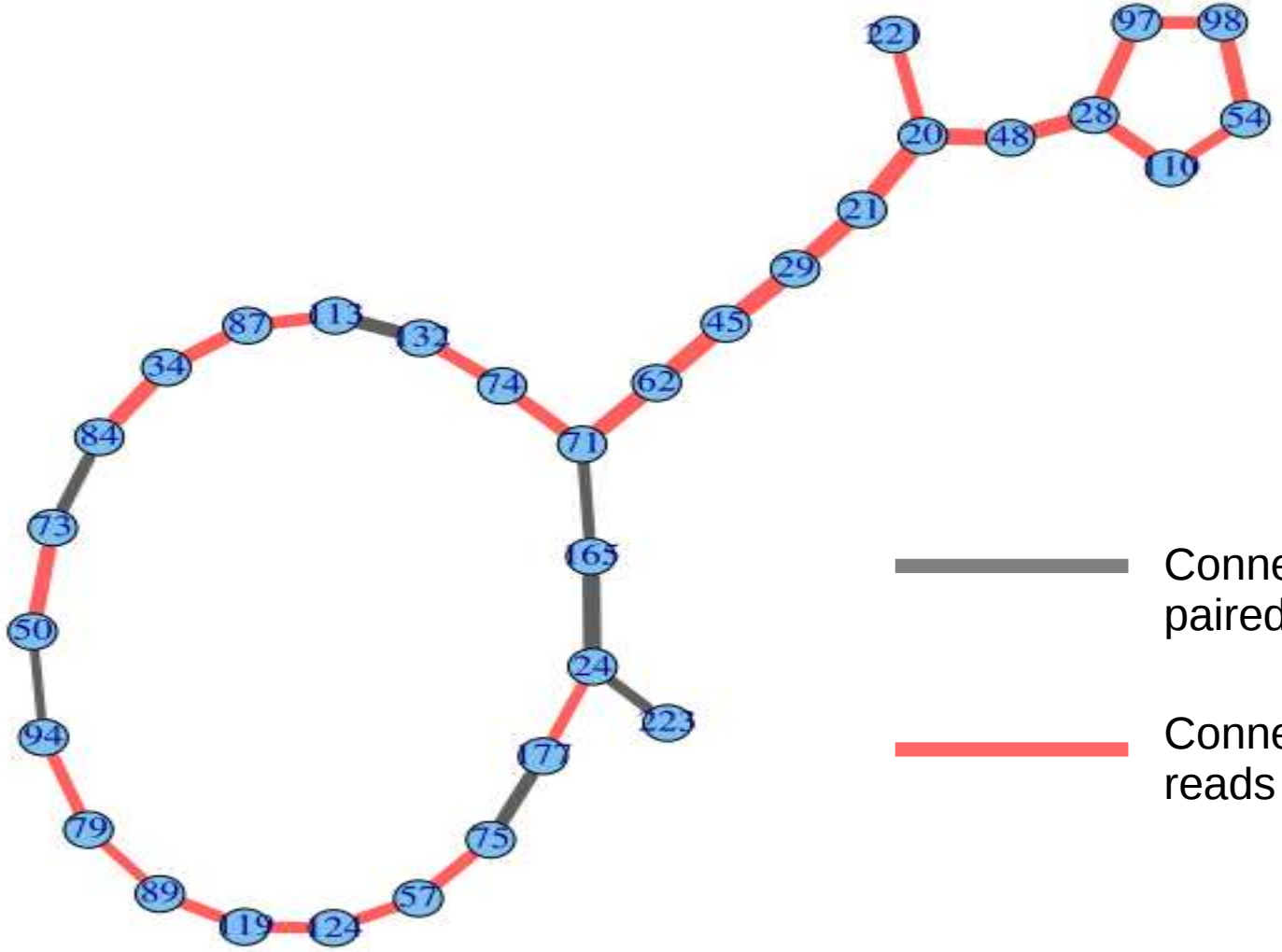
W number of reads pairs shared between clusters x and y
n_x and n_y is number of reads in cluster *x* and cluster *y* with absent read mate within the same cluster respectively

Suitable $k_{x,y}$ cutoff 0.05 – 0.2

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

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

Similarity between cluster - presence of paired reads

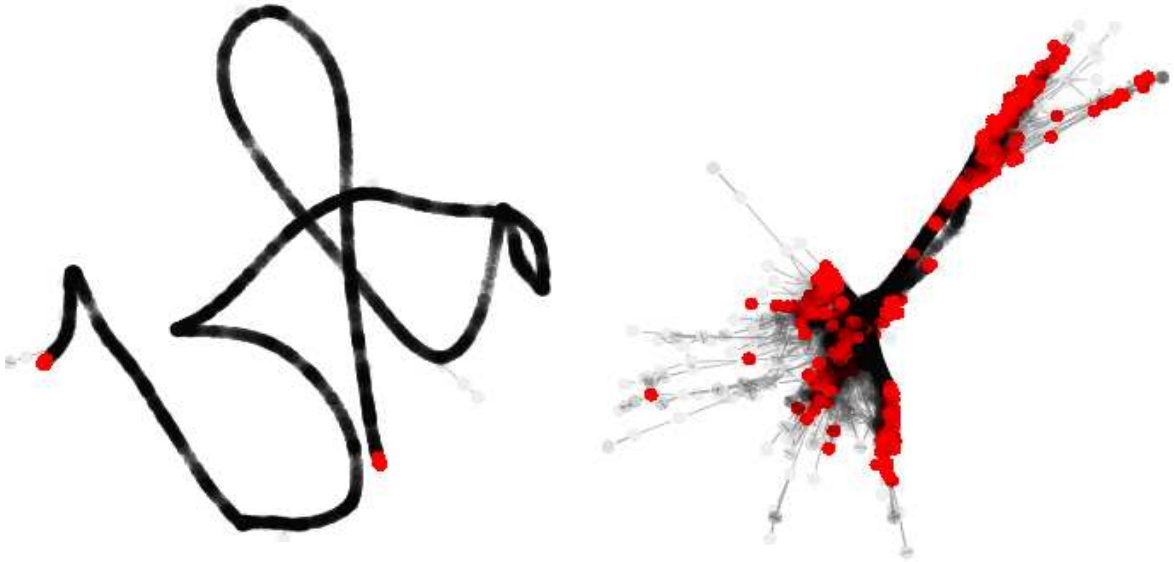


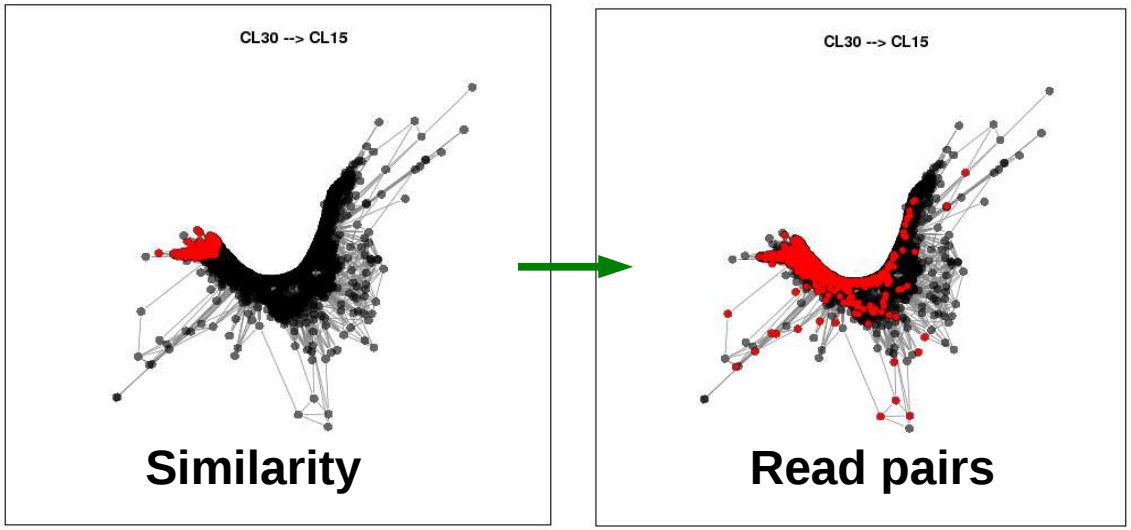
Connection through paired reads only

Connection through paired reads and similarity

Similarity between cluster

Distribution of similarity hits to other clusters

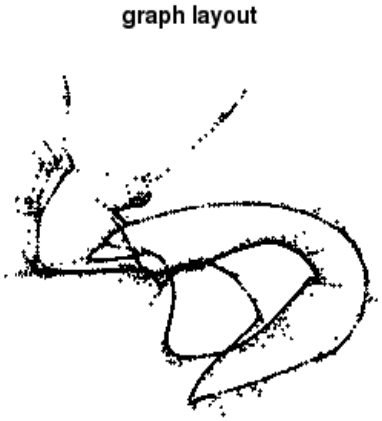


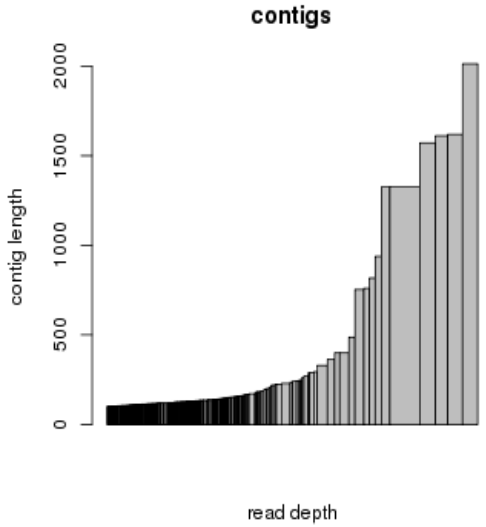


Similarity vs. read pairs

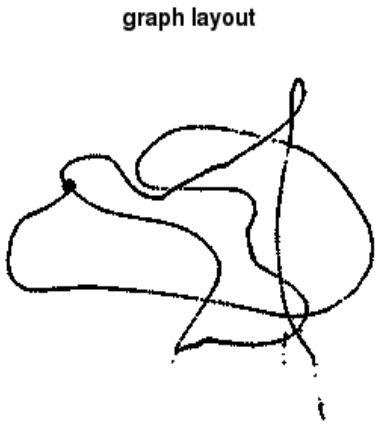
Pairs → physical connection

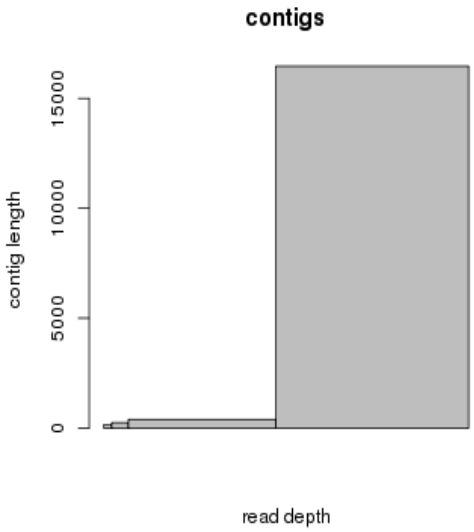
Sequence assembly – each cluster is assembled separately





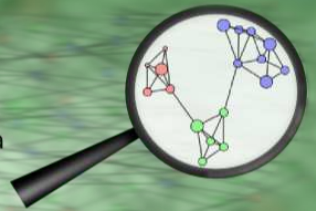
- Number of contigs
- Contig length
- Variability

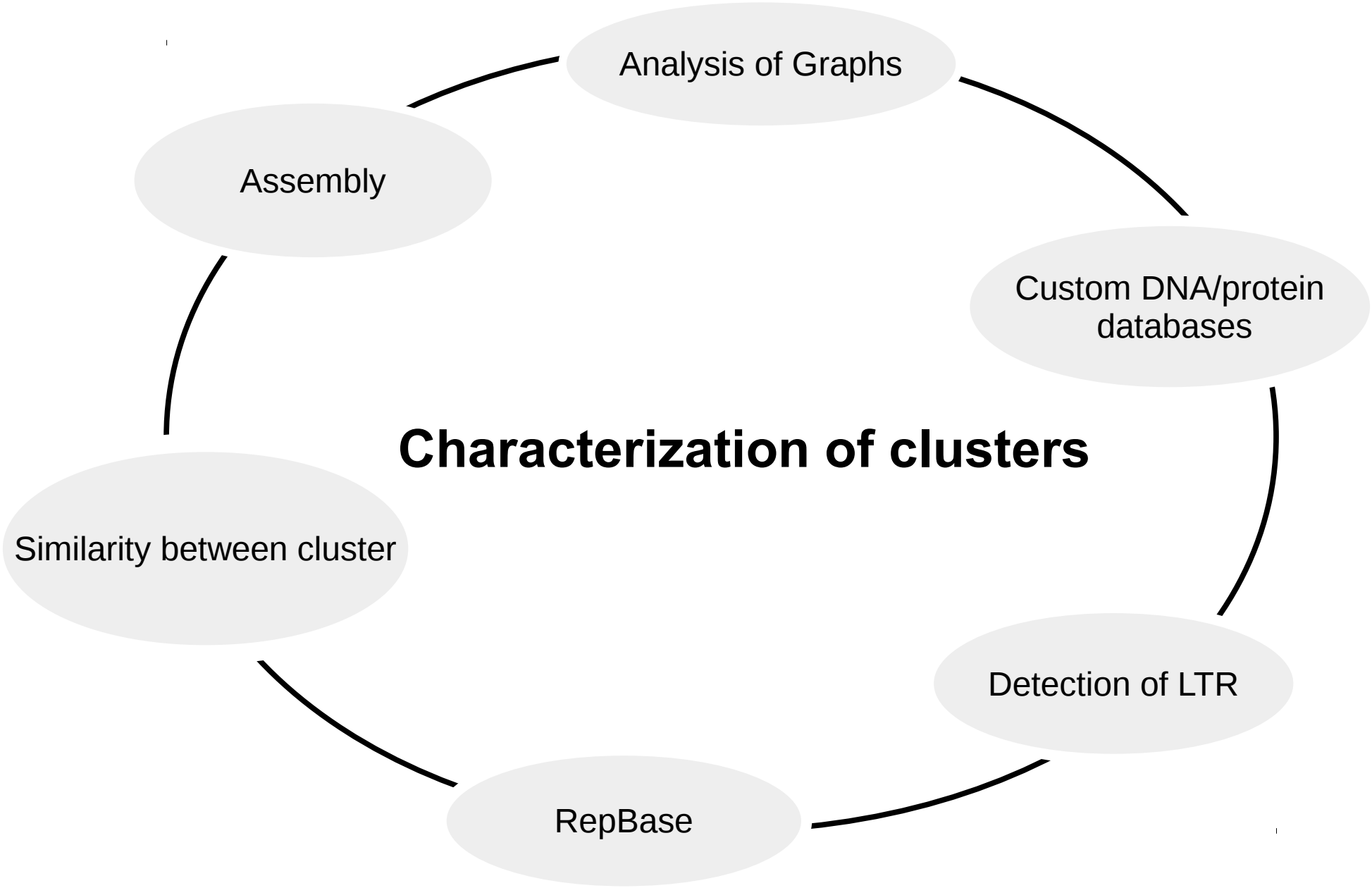


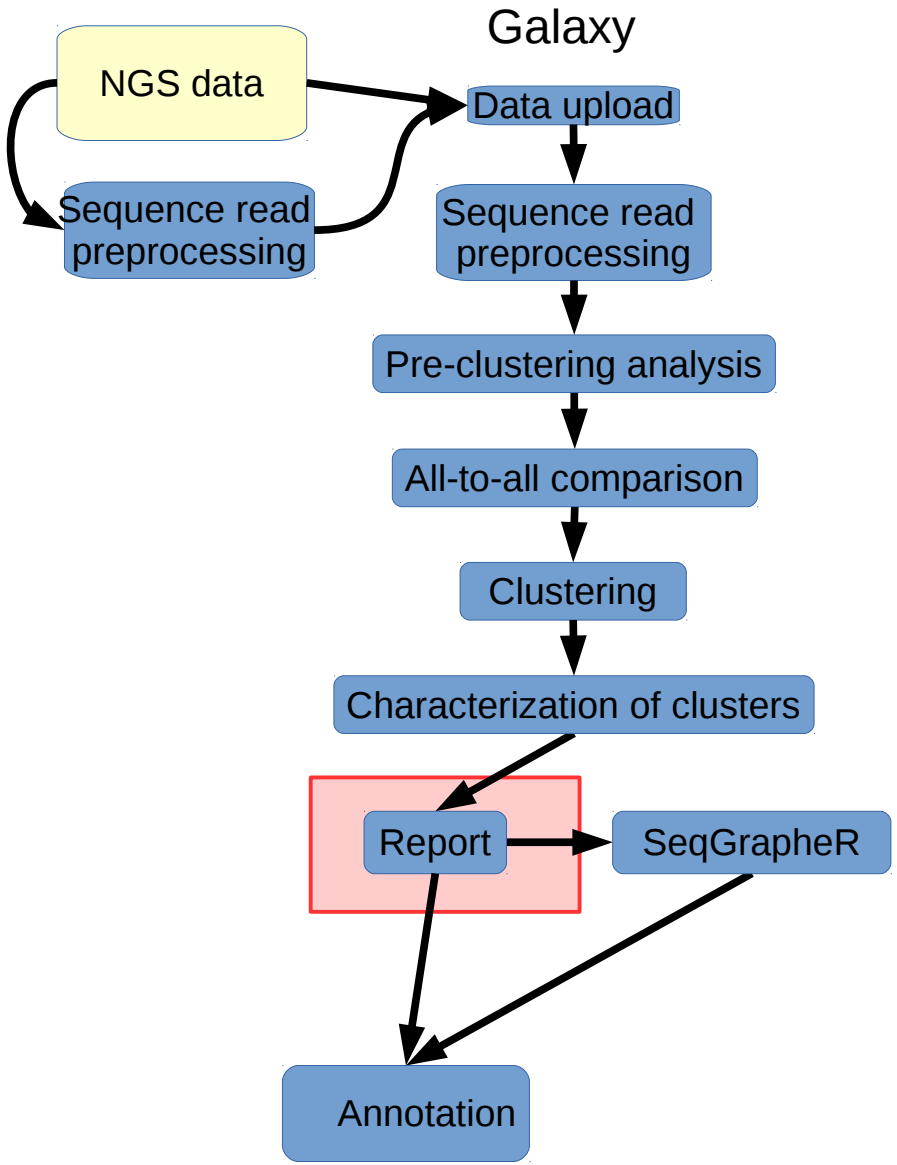


RepeatExplorer

Discover repeats in your next generation sequencing data







Output

- **HTML report** – user friendly graphical overview of top clusters
- **Archive with complete results**
- **Contigs** assembled from clusters – can be used directly in follow up analysis of protein domains
- **Log file –** information about pipeline progress

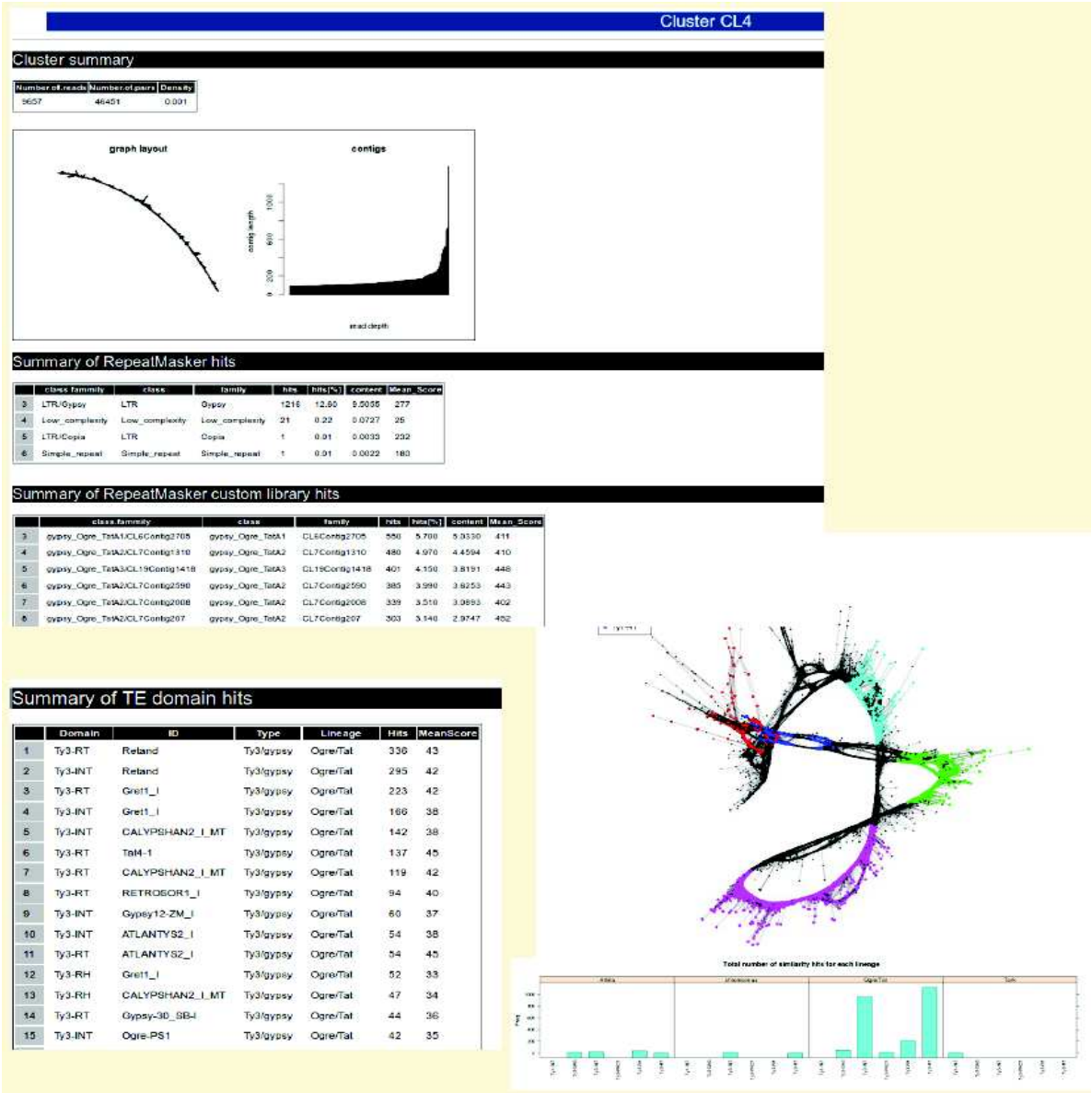
RepeatExplorer

Discover repeats in your next generation sequencing data

HTML report – user friendly graphical overview of top clusters

	cluster	total length [bp]	number of reads	Genome proportion[%]	cumulative GP [%]	Repeat Masker	Domain hits	Repeat Masker custom library	Layout	All missing mates [%]	Missing mates with no similarity hit [%]	Portion of similarity hits to other clusters[%]	Outside reads with similarity [%]
1	CL1	3339900	37110	7.850	7.8	Low_complexity (49hits, 0.0381%) RC.Helitron (9hits, 0.0119%) Simple_repeat (17hits, 0.00979%) LTR.Gypsy (5hits, 0.00596%) LINE.L1 (2hits, 0.00443%) LTR.Copia (2hits, 0.00305%)		SatA_Parvisepalum.CL1Contig965 (**18600hits, 49.3%**) SatA_Parvisepalum.CL1Contig940 (**8181hits, 21.6%**) SatA_Parvisepalum.CL1Contig886 (**7019hits, 18.5%**) SatA_Parvisepalum.CL1Contig393 (**2.....**		68.3	0.51	37.540	89.600
2	CL2	2954610	32829	6.950	14.8	Low_complexity (11hits, 0.00941%) RC.Helitron (3hits, 0.00437%) Simple_repeat (4hits, 0.00257%) DNA.CMC.EnSpm (1hits, 0.0019%) LTR.Copia (1hits, 0.00146%) LTR.Gypsy (1hits, 0.00129%)		SatA_Parvisepalum.CL1Contig965 (**17779hits, 53.2%**) SatA_Parvisepalum.CL1Contig940 (**8246hits, 24.7%**) SatA_Parvisepalum.CL1Contig393 (**5389hits, 16.1%**) SatA_Parvisepalum.CL1Contig971 (**1.....**		76.9	0.47	66.370	112.900
3	CL3	1022670	11363	2.400	17.2	Low_complexity (**3739hits, 15.5%**) Simple_repeat (102hits, 0.628%) LTR.Gypsy (6hits, 0.0306%)	DTH-CD1 NA NA (1 hits 0.0088%)	UnknownG.CL26Contig1825 (**784hits, 5.69%**) UnknownG.CL26Contig1920 (**529hits, 4.18%**) SatD1.CL4Contig1816 (**481hits, 3.73%**) SatD1.CL4Contig3001 (201hits, 1.42%) SatF.CL8Contig4082 (144.....		35.9	13.06	0.074	0.378
4	CL4	869130	9657	2.040	19.2	LTR.Gypsy (**1216hits, 9.51%**) Low_complexity (21hits, 0.0727%) LTR.Copia (1hits, 0.00334%) Simple_repeat (1hits, 0.00219%)	Ty3-RT Ty3/gypsy Ogre/Tat (1118 hits 11.6%) Ty3-INT Ty3/gypsy Ogre/Tat (956 hits 9.9%) Ty3-RH Ty3/gypsy Ogre/Tat (208 hits 2.15%) Ty3-GAG Ty3/gypsy Ogre/Tat (50 hits	gypsy_Ogre_TatA1.CL6Contig2705 (**550hits, 5.03%**) gypsy_Ogre_TatA2.CL7Contig1310 (**480hits, 4.46%**) gypsy_Ogre_TatA3.CL19Contig1418 (**401hits, 3.82%**) gypsy_Ogre_TatA2.CL7Contig2590 (**385h.....**		29.1	15.16	0.015	0.072

HTML report – individual clusters



Cluster CL4

Cluster summary

Number.of.reads	Number.of.pairs	Density
9657	46451	0.001

graph layout

contigs

Summary of RepeatMasker hits

	class.family	class	family	hits	hits[%]	content	Mean.Score
3	LTR/Gypsy	LTR	Gypsy	1216	12.60	9.5056	277
4	Low_complexity	Low_complexity	Low_complexity	21	0.22	0.0727	25
5	LTR/Copia	LTR	Copia	1	0.01	0.0033	232
6	Simple_repeat	Simple_repeat	Simple_repeat	1	0.01	0.0022	180

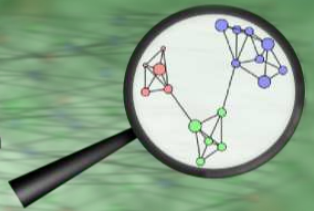
Summary of RepeatMasker custom library hits

	class.family	class	family	hits	hits[%]	content	Mean.Score
3	gypsy_Ogre_TatA1/CL6Contig2705	gypsy_Ogre_TatA1	CL6Contig2705	550	5.700	5.9330	411
4	gypsy_Ogre_TatA2/CL7Contig1310	gypsy_Ogre_TatA2	CL7Contig1310	480	4.970	4.4594	410
5	gypsy_Ogre_TatA3/CL19Contig1418	gypsy_Ogre_TatA3	CL19Contig1418	401	4.150	3.8191	448
6	gypsy_Ogre_TatA2/CL7Contig2590	gypsy_Ogre_TatA2	CL7Contig2590	385	3.990	3.8253	443
7	gypsy_Ogre_TatA2/CL7Contig2008	gypsy_Ogre_TatA2	CL7Contig2008	339	3.510	3.9893	402
8	gypsy_Ogre_TatA2/CL7Contig207	gypsy_Ogre_TatA2	CL7Contig207	303	3.140	2.9747	452

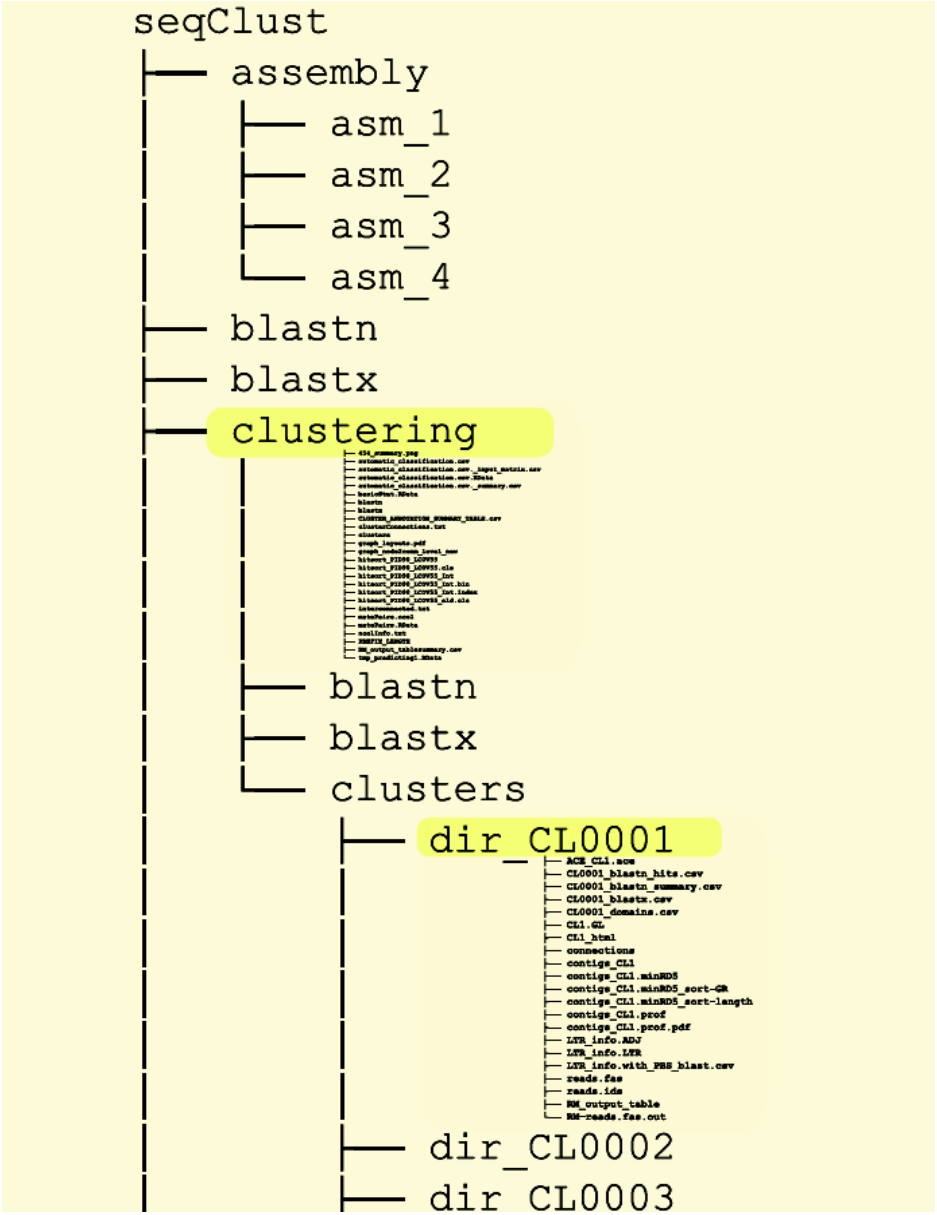
Summary of TE domain hits

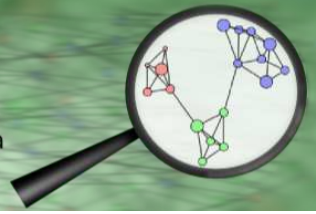
	Domain	ID	Type	Lineage	Hits	MeanScore
1	Ty3-RT	Reland	Ty3/gypsy	Ogre/Tat	336	43
2	Ty3-INT	Reland	Ty3/gypsy	Ogre/Tat	295	42
3	Ty3-RT	Gret1_I	Ty3/gypsy	Ogre/Tat	223	42
4	Ty3-INT	Gret1_I	Ty3/gypsy	Ogre/Tat	166	38
5	Ty3-INT	CALYPSHAN2_I_MT	Ty3/gypsy	Ogre/Tat	142	38
6	Ty3-RT	Tat4-1	Ty3/gypsy	Ogre/Tat	137	45
7	Ty3-RT	CALYPSHAN2_I_MT	Ty3/gypsy	Ogre/Tat	119	42
8	Ty3-RT	RETROSOR1_I	Ty3/gypsy	Ogre/Tat	94	40
9	Ty3-INT	Gypsy12-ZM_I	Ty3/gypsy	Ogre/Tat	60	37
10	Ty3-INT	ATLANTYS2_I	Ty3/gypsy	Ogre/Tat	54	38
11	Ty3-RT	ATLANTYS2_I	Ty3/gypsy	Ogre/Tat	54	45
12	Ty3-RH	Gret1_I	Ty3/gypsy	Ogre/Tat	52	33
13	Ty3-RH	CALYPSHAN2_I_MT	Ty3/gypsy	Ogre/Tat	47	34
14	Ty3-RT	Gypsy-30_SB-I	Ty3/gypsy	Ogre/Tat	44	36
15	Ty3-INT	Ogre-PS1	Ty3/gypsy	Ogre/Tat	42	35

Total number of similarity hits for each lineage



Archive with complete results



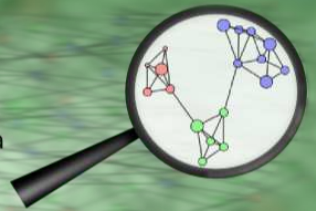


>CLxContigY (length[bp]-read_depth-genome_representation)

>CL1Contig1 (126-9.0-1128)
GACCAATTTTAAAATGTGAAATTAATGATCCCACATTGAAATTGGACAAATTATAAGCTA
TGTGGACTTTTATTTTGGTTAATTTGGACACCAATTGACCTTCAATTTGATTGGTCAAAC
TTAGGT
>CL1Contig2 (114-6.3-719)
TAATCTGACCAAATTTGGACCCAATCTTTGAACCTTCTAATTAAAGTTCAATTGGTGTCC
AAATCAACTCAAATAAAATTCCATATAGCTTAGAATTTGCCCAAATTCAATGTC
>CL1Contig3 (106-1.6-174)
TTCAAAATAACCTATTTAGGCTTTTAAATGGAGTAAGTTGAACTTTAATTGACCTTCAAT
TGGAAGGGTCAAAGTTTGAGTCAAATTTAGGGTCAATTTTGACCAA
>CL1Contig4 (105-3.2-335)
AATTGTAAGCTATATTGAATTTTAATTGAGTTTATTTGGACACCAATTGACCTTCATATT
CAAAGATTGGGTCCAAATTAGGTTTGATTTTGACCAACCTCAAAA
>CL1Contig5 (120-2.2-267)

RepeatExplorer

Discover repeats in your next generation sequencing data



>CLxContigY (length[bp]-read_depth-genome_representation)

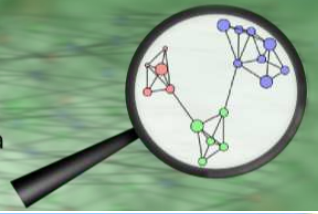
```
>CL1Contig1 (126-9.0-1128)
GACCAATTTTAAAATGTGAAATTAATGATCCCACATTGAAATTGGACAAATTATAAGCTA
TGTGGACTTTTATTTTGGTTAATTTGGACACCAATTGACCTTCAATTTGATTGGTCAAAC
TTAGGT
>CL1Contig2 (114-6.3-719)
TAATCTGACCAAATTTGGACCCAATCTTTGAACCTTCTAATTAAAGTTCAATTGGTGTCC
AAATCAACTCAAATAAAATTCCATATAGCTTAGAATTTGCCCAAATTCAATGTC
>CL1Contig3 (106-1.6-174)
TTCAAAATAACCTATTTAGGCTTTTAAATGGAGTAAGTTGAACTTTAATTGACCTTCAAT
TGGAAGGGTCAAAGTTTGAGTCAAATTTAGGGTCAATTTTGACCAA
>CL1Contig4 (105-3.2-335)
AATTGTAAGCTATATTGAATTTTAATTGAGTTTATTTGGACACCAATTGACCTTCATATT
CAAAGATTGGGTCCAAATTAGGTTTGATTTTGACCAACCTCAAAA
>CL1Contig5 (120-2.2-267)
```

Protein domain analysis

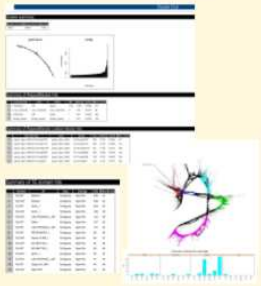
ChIP-seq-mapper

RepeatExplorer

Discover repeats in your next generation sequencing data



HTML Report



Contigs

Archive with complete results

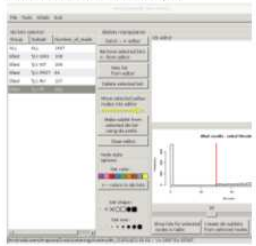
```
seqClust
├── assembly
│   ├── asm_1
│   ├── asm_2
│   ├── asm_3
│   └── asm_4
├── blastn
├── blastx
├── clustering
│   ├── blastn
│   ├── blastx
│   └── clusters
│       ├── dir_CL0001
│       ├── dir_CL0002
│       ├── dir_CL0003
│       ├── dir_CL0004
│       ├── dir_CL0005
│       ├── dir_CL0006
│       ├── dir_CL0007
│       ├── dir_CL0008
│       ├── dir_CL0009
│       ├── dir_CL0010
│       ├── dir_CL0011
│       ├── dir_CL0012
│       ├── dir_CL0013
│       ├── dir_CL0014
│       ├── dir_CL0015
│       ├── dir_CL0016
│       ├── dir_CL0017
│       └── dir_CL0018
├── prerun
└── sequences
```

Log file

Annotation(Manual/Automatic)

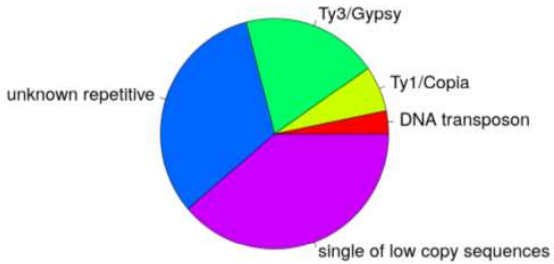
SeqGrapheR

Interactive graph analysis

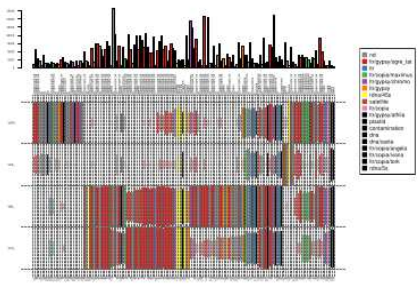


Reconstruction of complete
Probe, Primer design

Genome composition



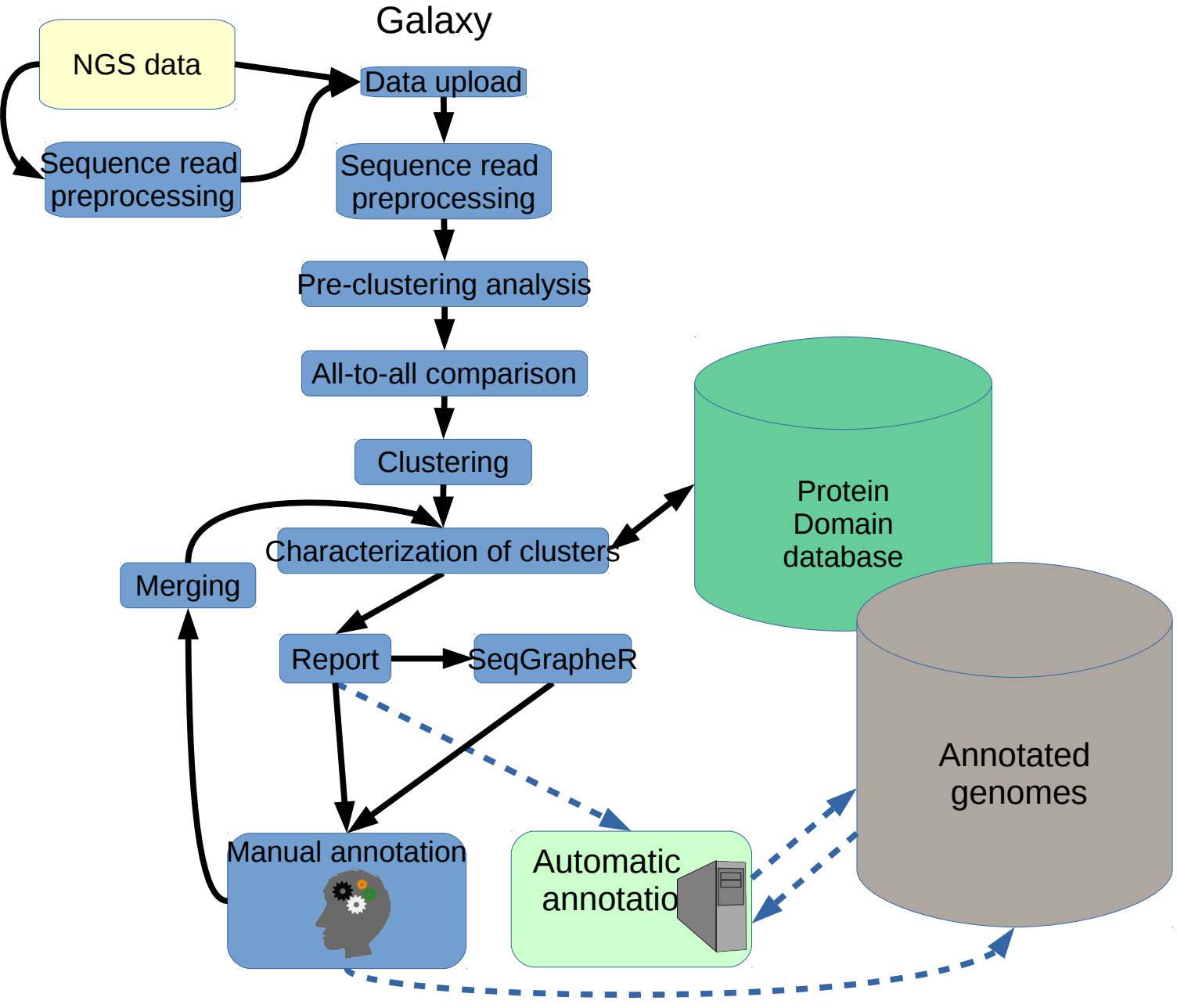
Comparative analysis

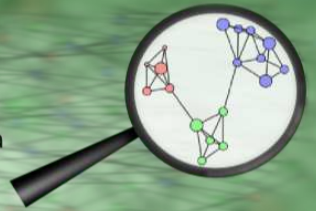


RepeatExplorer

Discover repeats in your next generation sequencing data

Automatic annotation





RepeatExplorer availability:

source code
https://bitbucket.org/repeatexplorer/

- Galaxy server – Graphical user interface

www.repeatexplorer.org
https://galaxy-elixir.cerit-sc.cz/ ELIXIR-CZ instance

- Your custom instance

- Command line