RepeatExplorer

Discover repeats in your next generation sequencing data

3rd Workshop on the Application of Next Generation Sequencing to Repetitive DNA Analysis in Plants

May 27-29, 2014

Institute of Plant Molecular Biology, České Budějovice, Czech Republic

Outline

- Principles
- Implementation
- Analysis work-flow
- New tools
- Reproducibility
- Replicates

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

Contributors:

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

RepeatExplorer

Discover repeats in your next generation sequencing data

Implementation

Galaxy server (http://galaxyproject.org)

- User friendly interface
- **Additional NGS related tools**
- Workflows / modularity
- Data and workflows sharing
- Can be installed on your own Galaxy server
- Keep a **history** as you try & retry tools = **reproducibility**

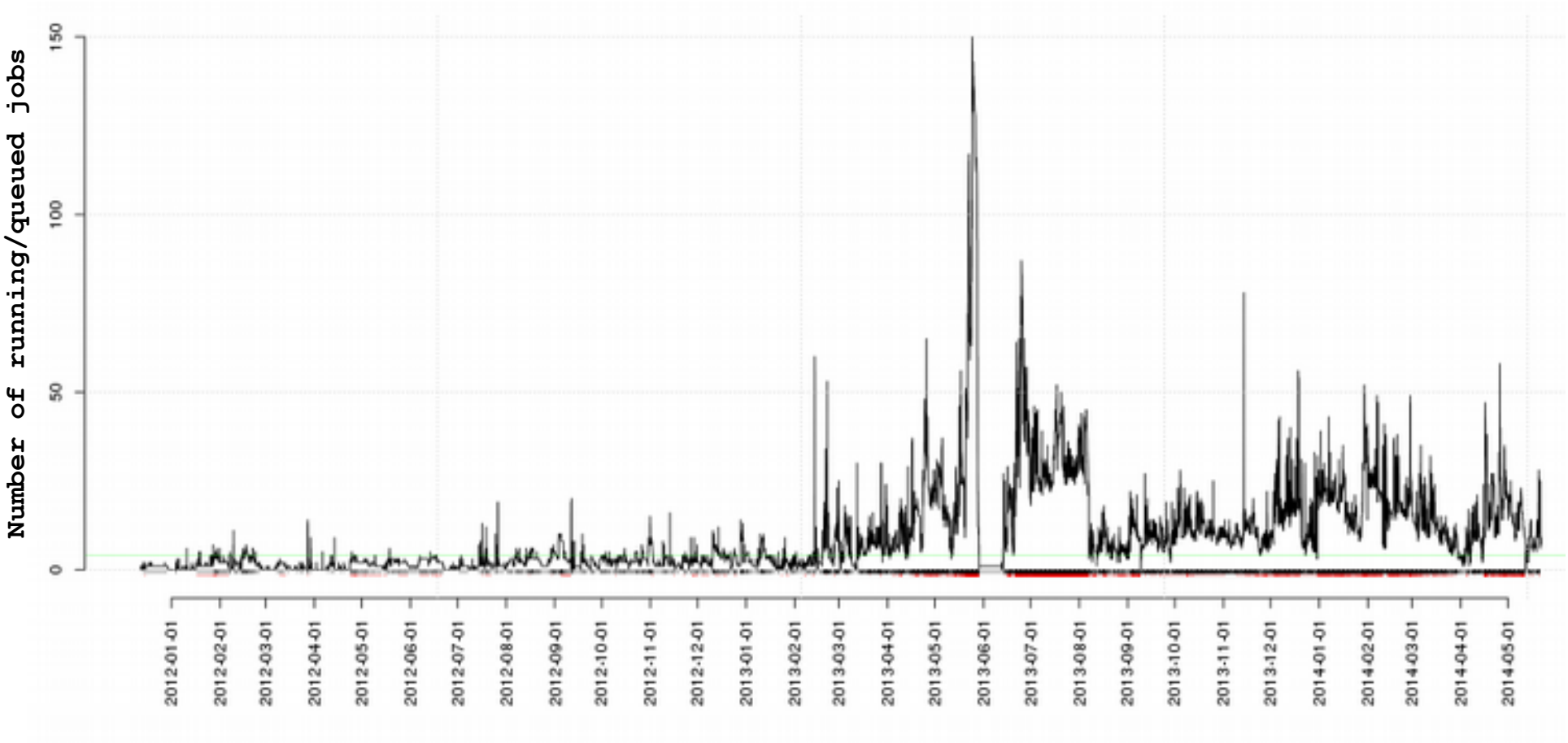
Command line scripts

- High throughput analysis
- Disk space efficient
- Flexible

RepeatExplorer

Discover repeats in your next generation sequencing data

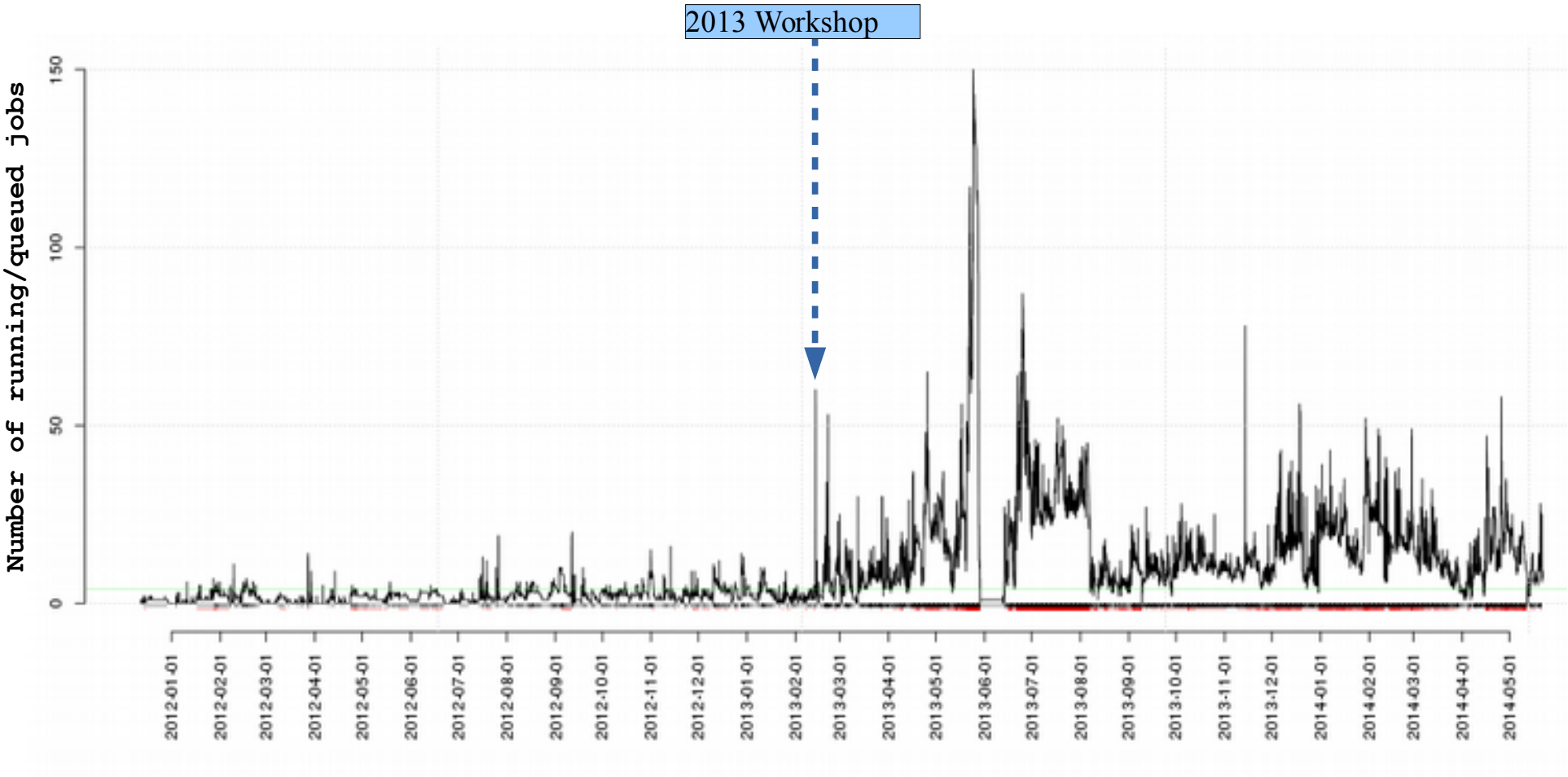
www.repeatexplorer.org
repeatexplorer.umbr.cas.cz



RepeatExplorer

Discover repeats in your next generation sequencing data

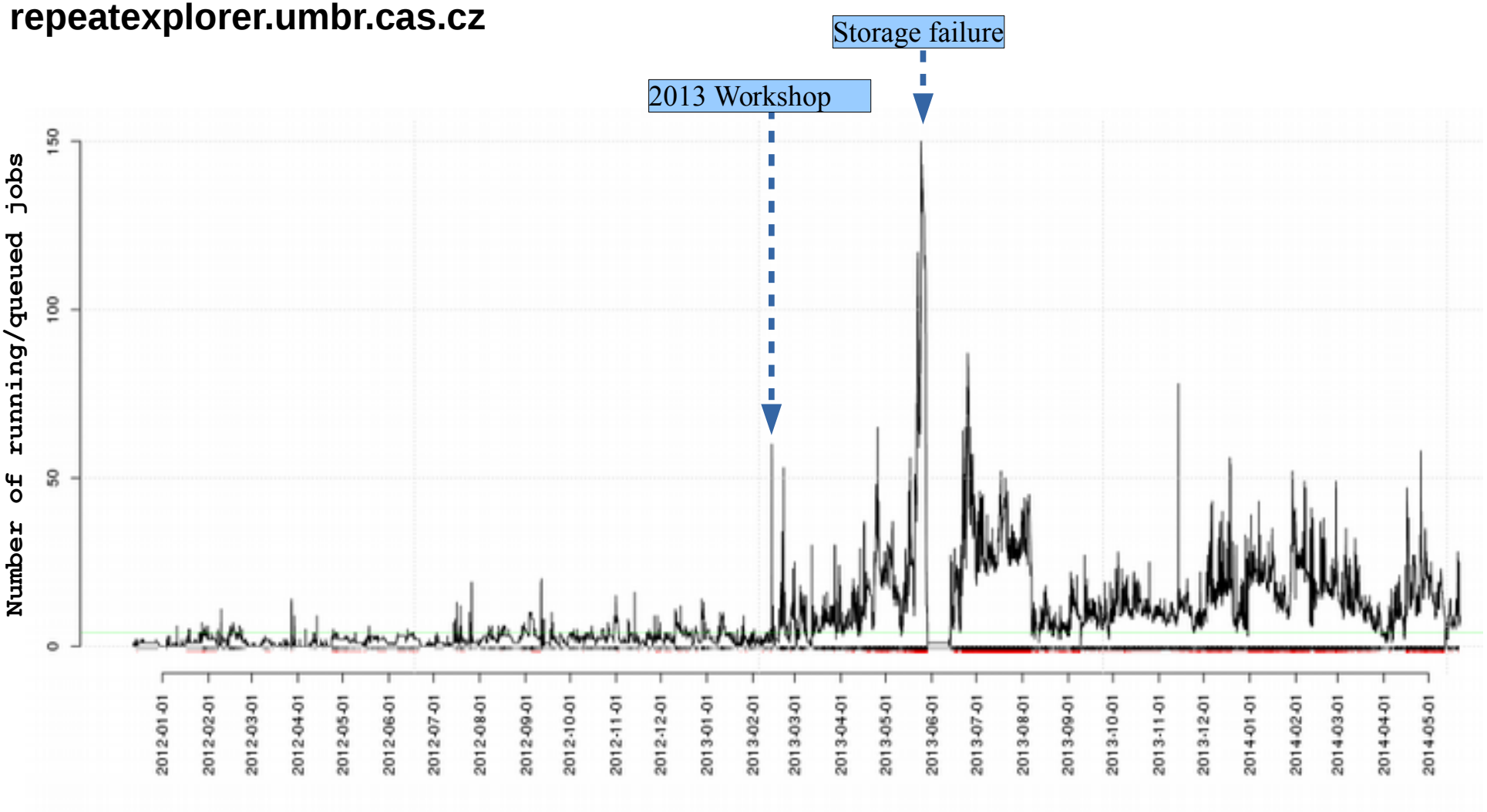
www.repeatexplorer.org
repeatexplorer.umbr.cas.cz



RepeatExplorer

Discover repeats in your next generation sequencing data

www.repeatexplorer.org
repeatexplorer.umbr.cas.cz

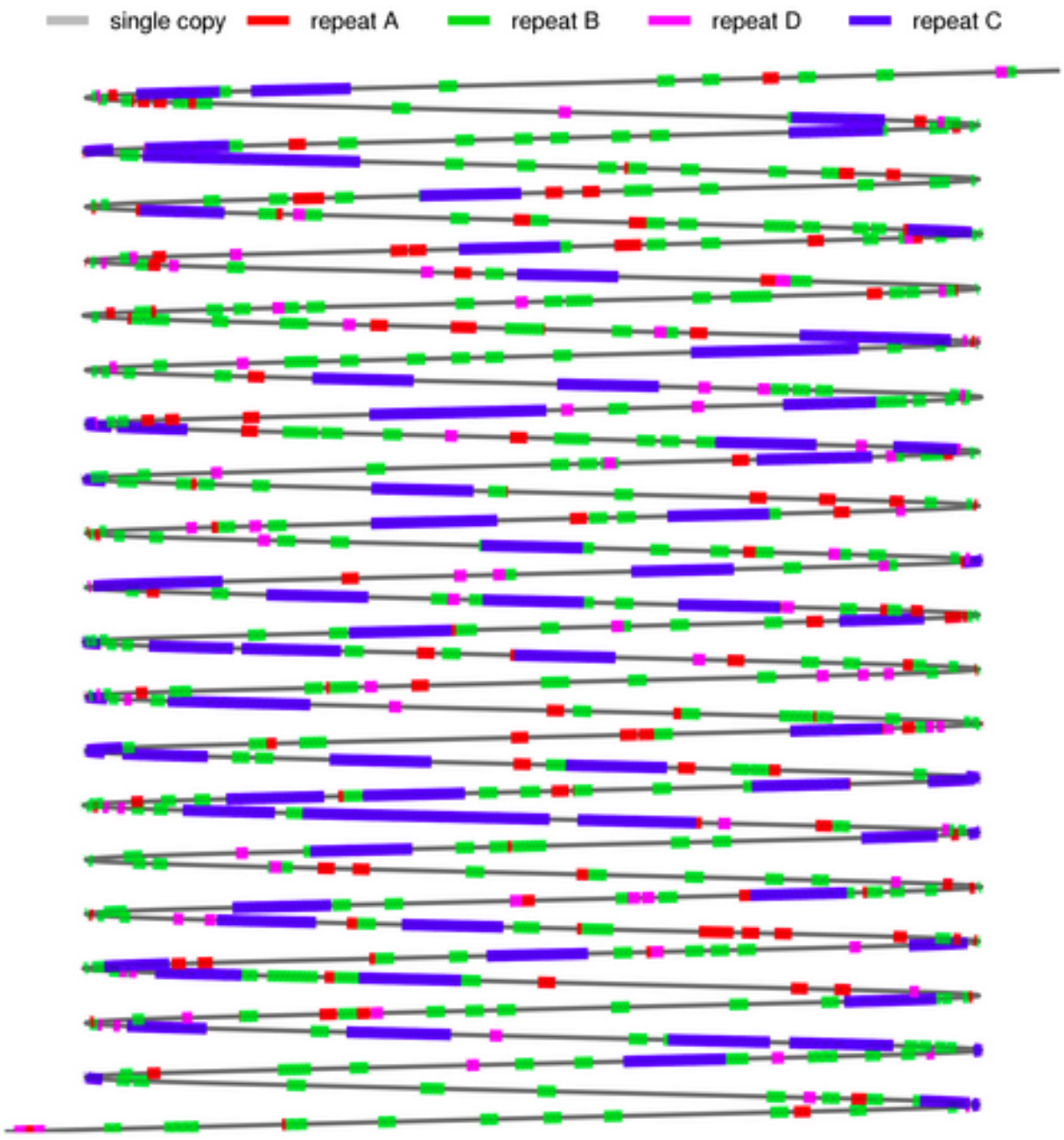


RepeatExplorer

Discover repeats in your next generation sequencing data

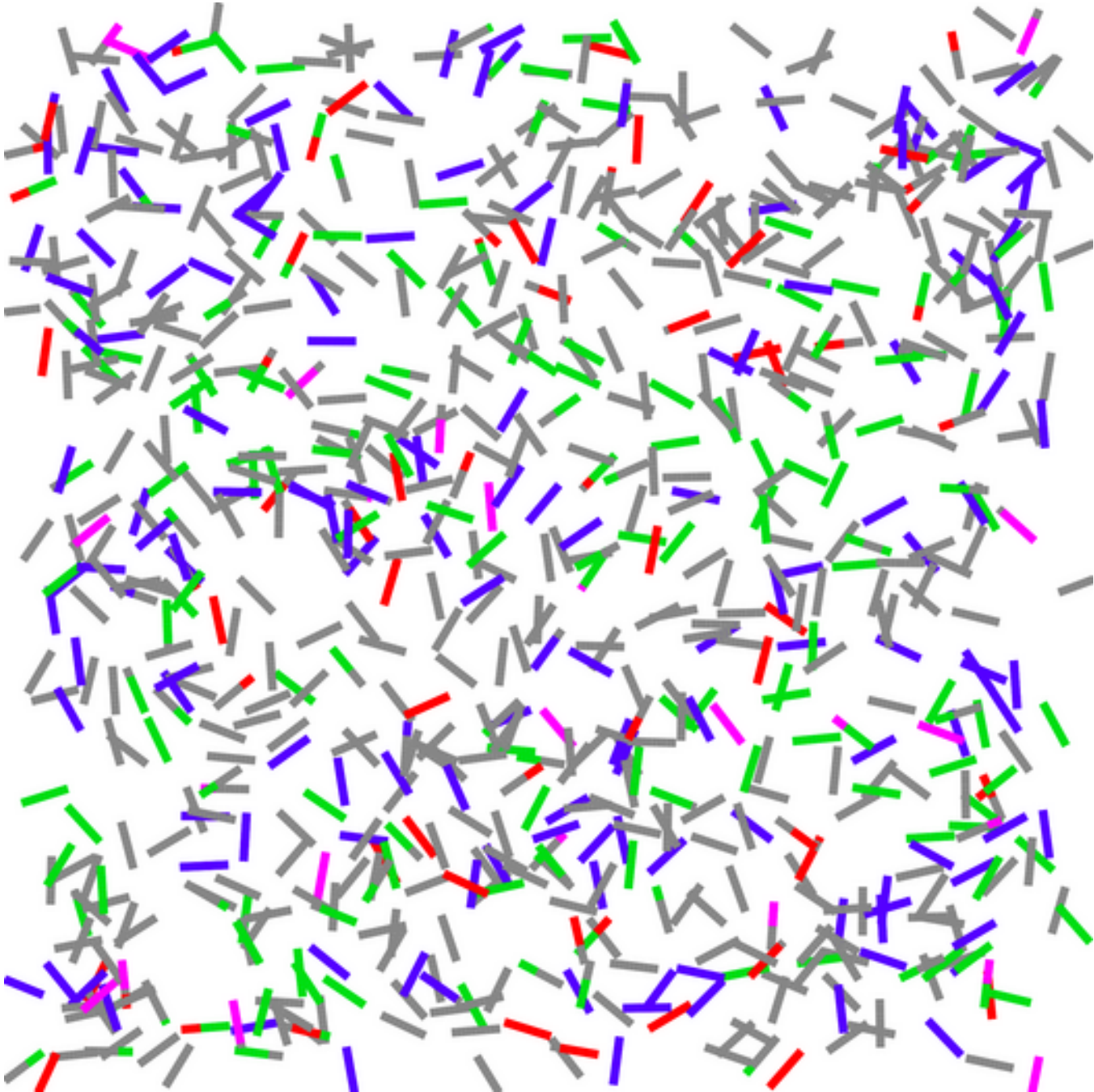
Principles

Genome



Principles

Shotgun sequencing

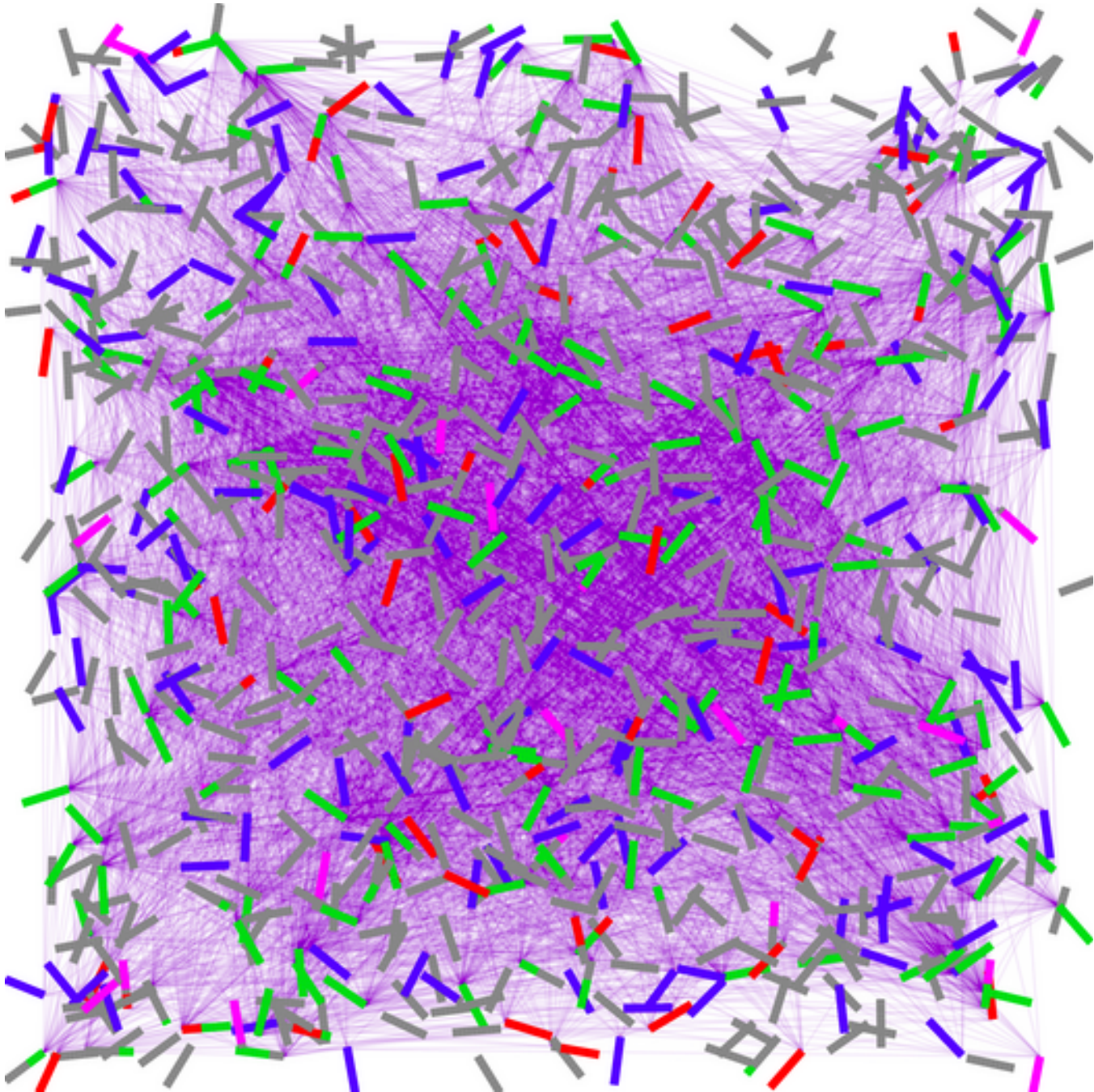


RepeatExplorer

Discover repeats in your next generation sequencing data

Principles

All-to-All similarity search

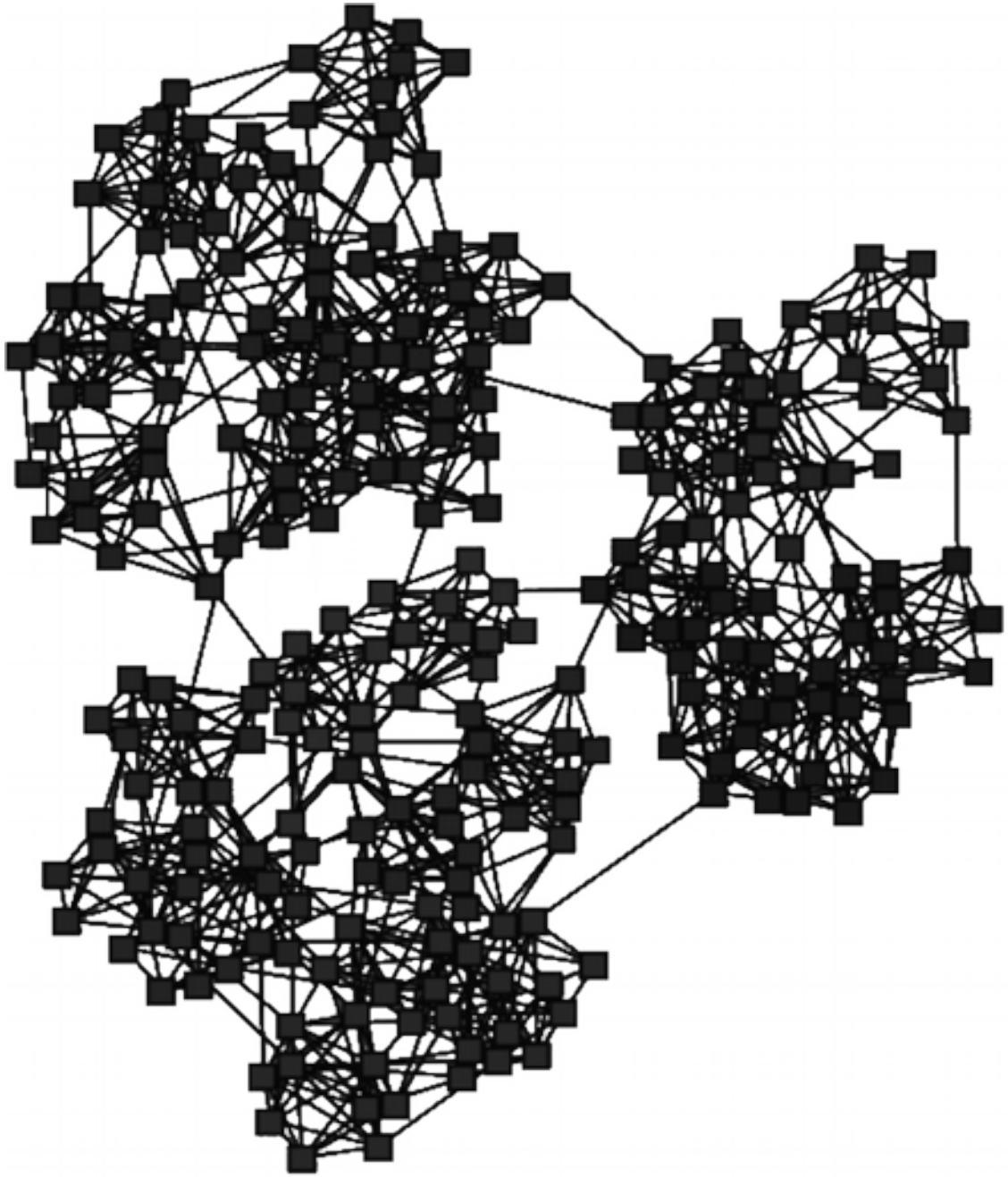


RepeatExplorer

Discover repeats in your next generation sequencing data

Principles

Graph Based Clustering

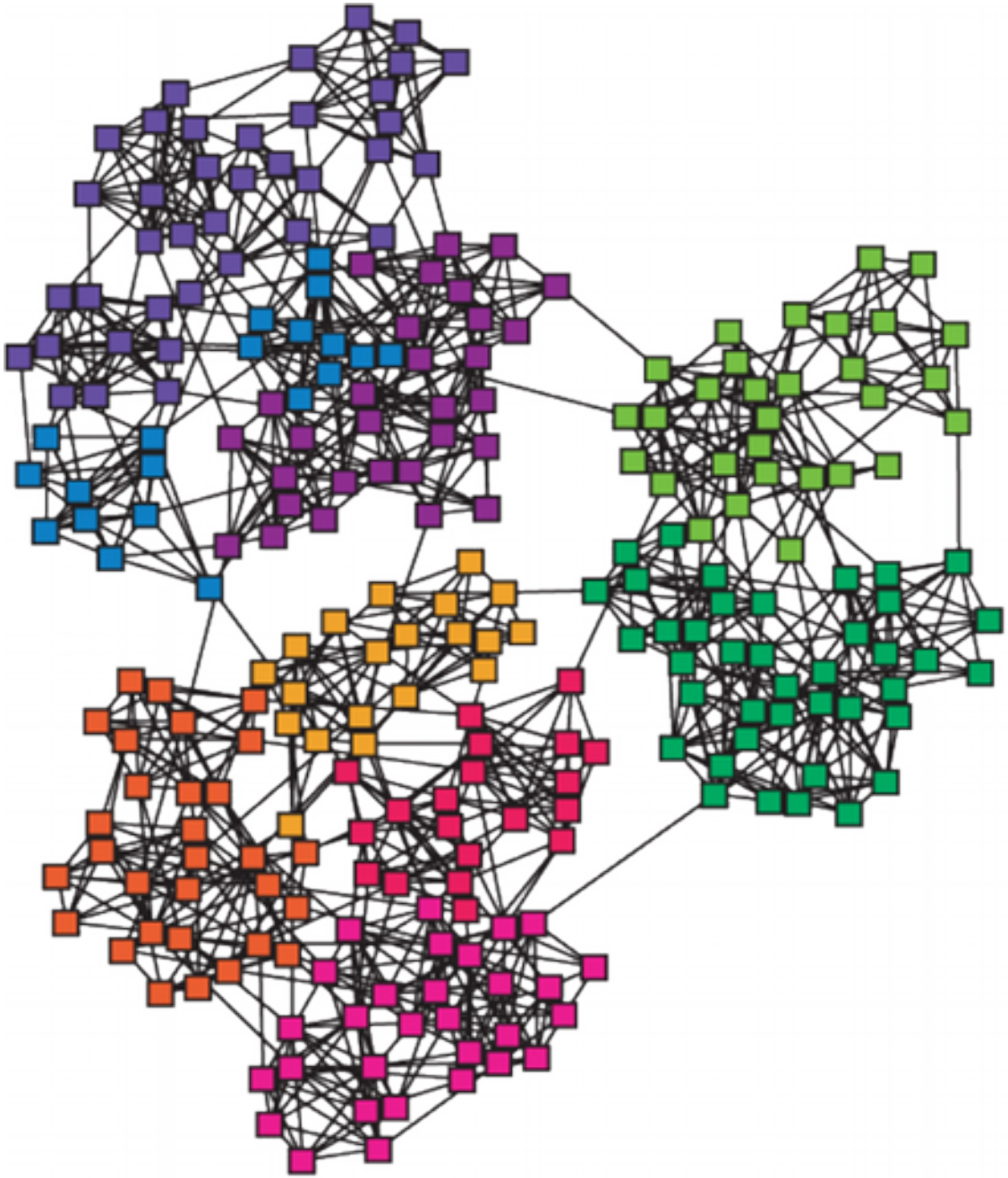


Graph Based Clustering

Finds clusters in network

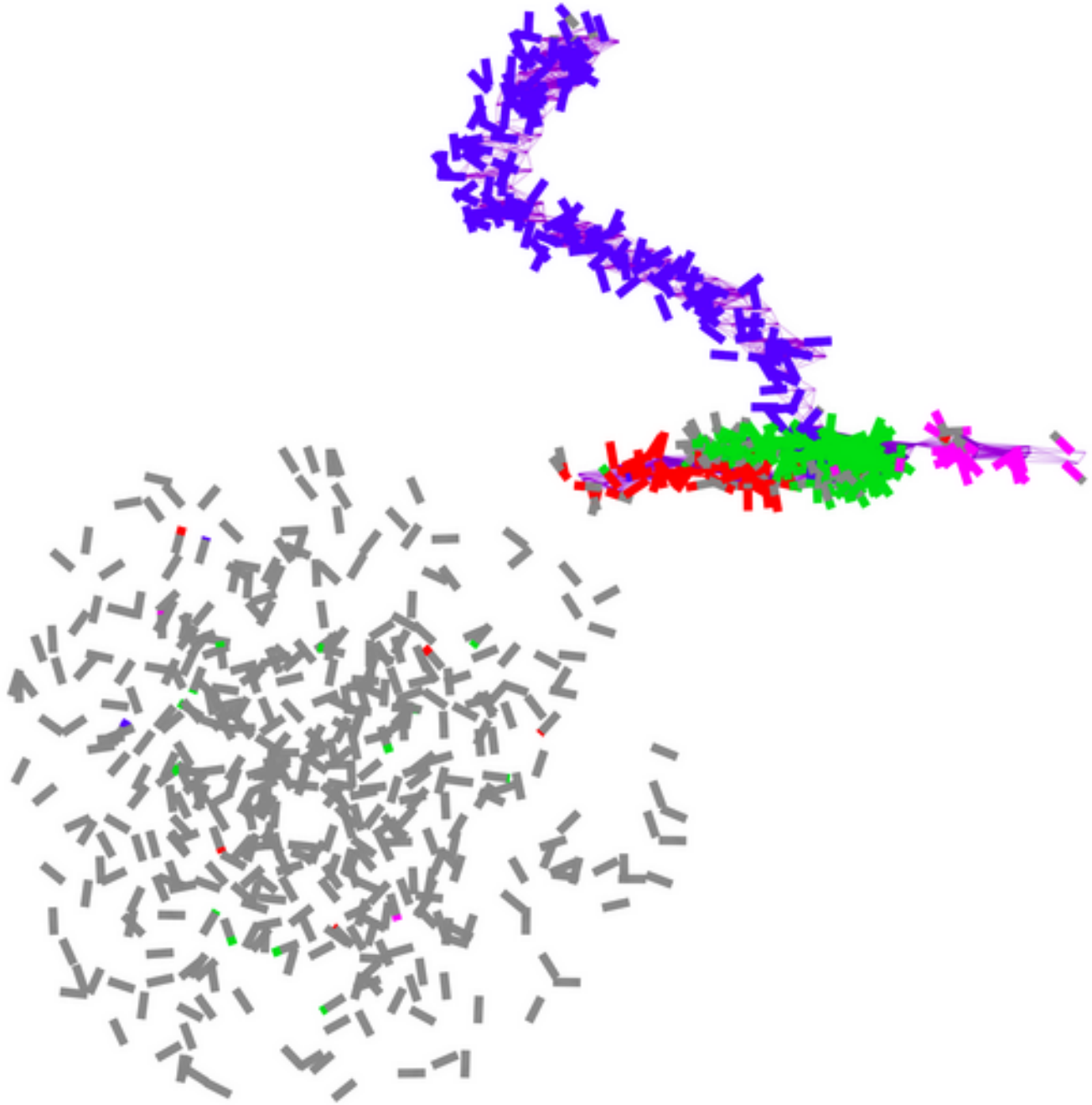
Modularity – quantify the quality of division a network into communities (clusters)

Community (clusters) – high number of connection between members, low number of connection to members of other communities



Principles

Identification of clusters

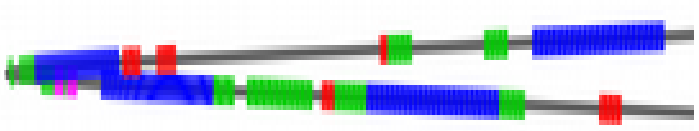


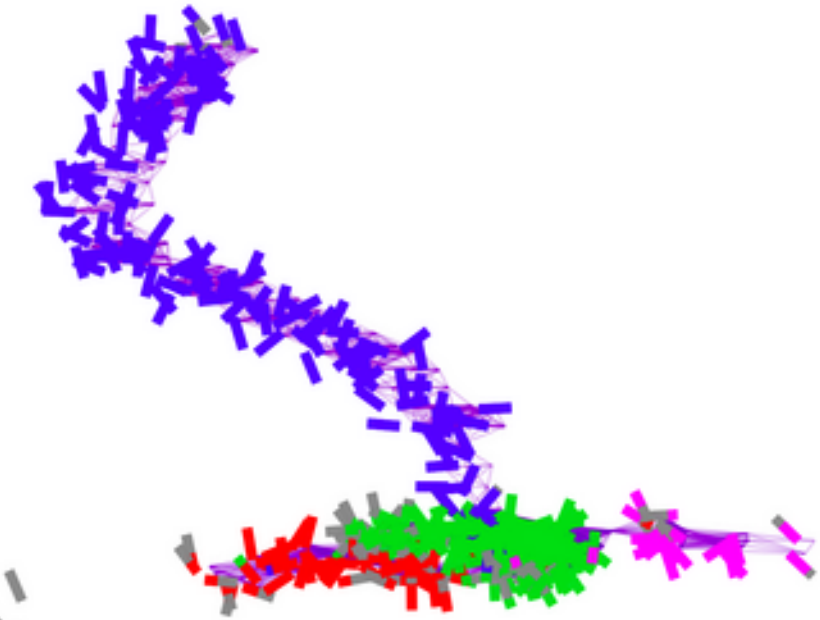
RepeatExplorer

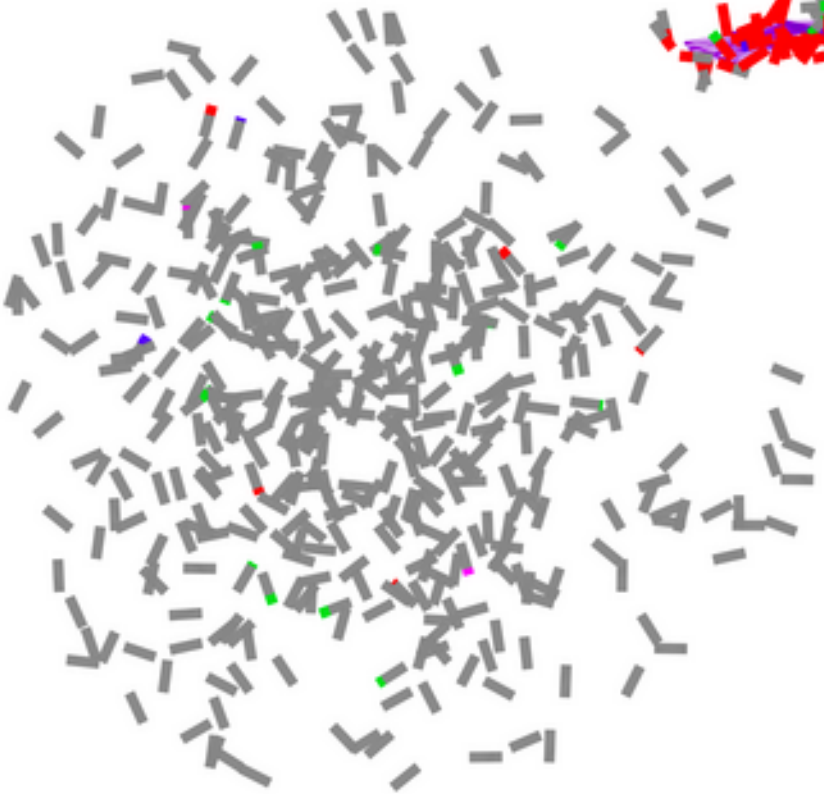
Discover repeats in your next generation sequencing data

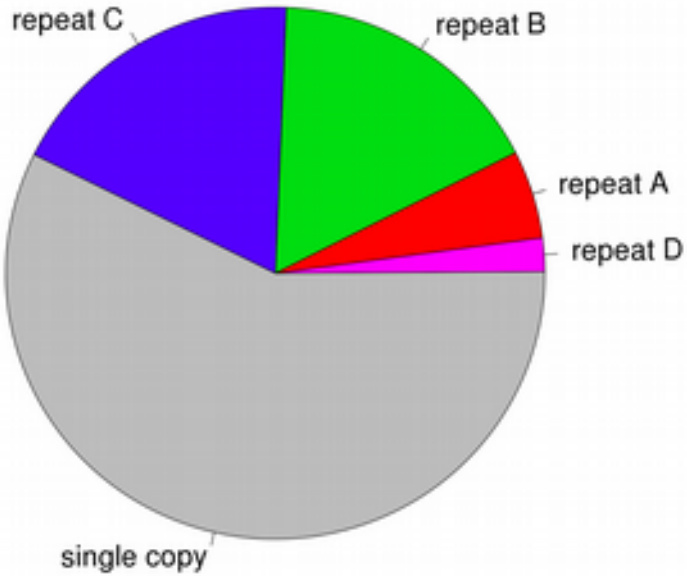
Principles

Identification of clusters









Proportion of repeat in the genome

RepeatExplorer

Discover repeats in your next generation sequencing data

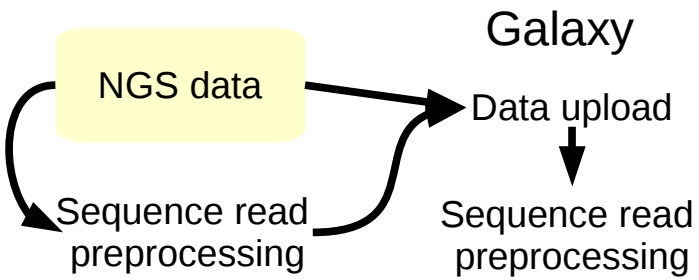
Analysis work-flow

NGS data

Sequence read:

- 100 nt or more
- Paired
- Uniform length

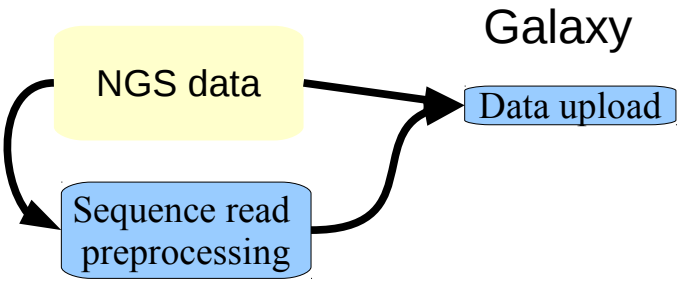
Analysis work-flow



Sequence read:

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

Analysis work-flow

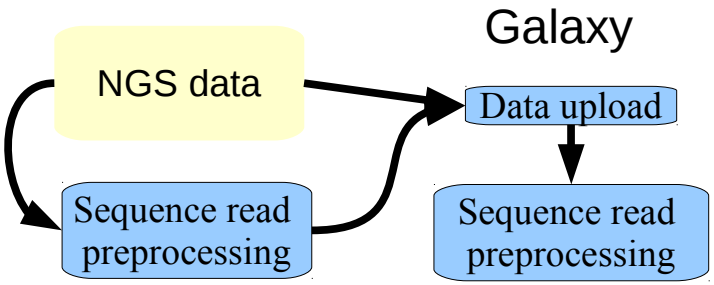


Upload methods

- Web interface (not suitable for large files)
- FTP - speed, stability, multiple files, resume
  - Filezilla (GUI)
  - Curl (command line)

- EBI SRA – direct download from databases

Analysis work-flow



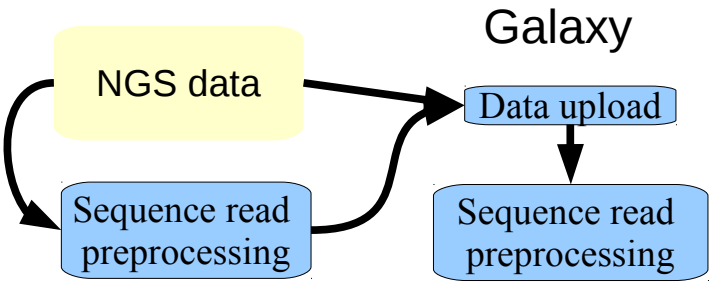
Sequence read:

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

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



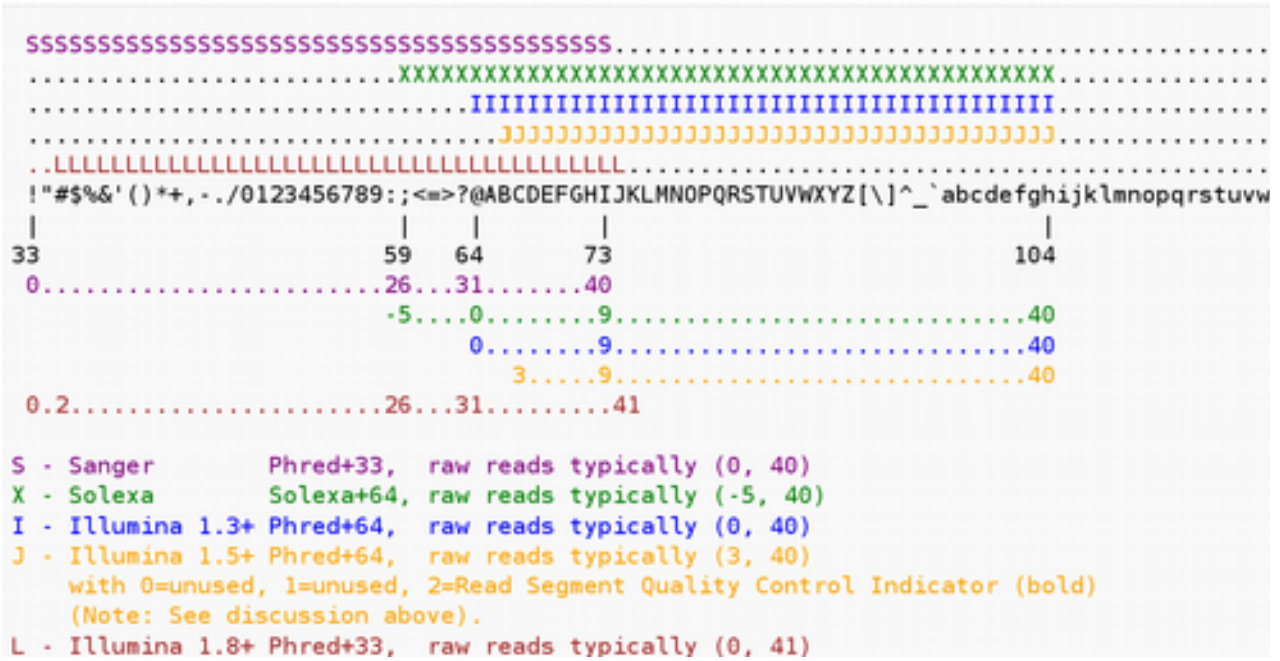
Sequence read:

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

Fastq format

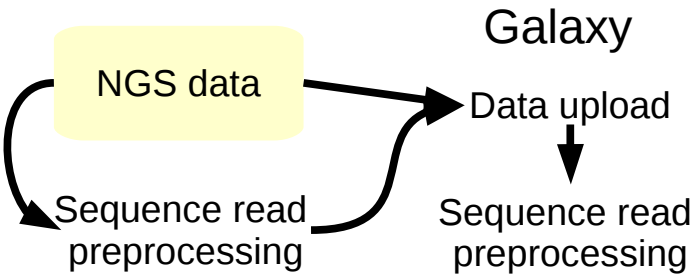
```
@SRR001666.1 071112_SLXA-
EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-
EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Quality coding



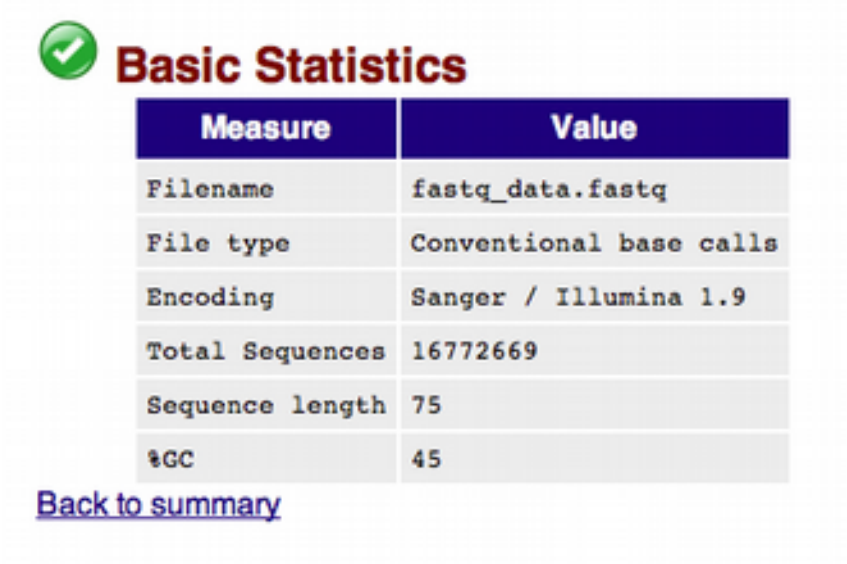
Source: wikipedia

Analysis work-flow



Sequence read:

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



Basic Statistics

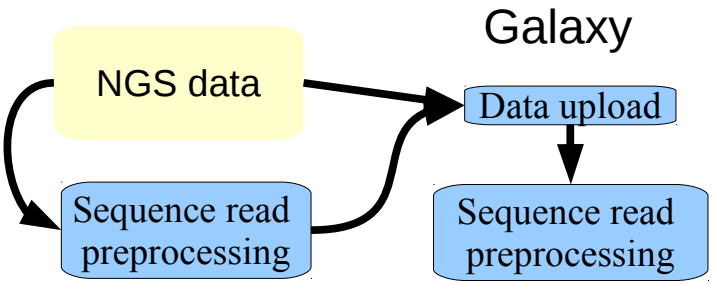
Measure	Value
Filename	fastq_data.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	16772669
Sequence length	75
%GC	45

Back to summary

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

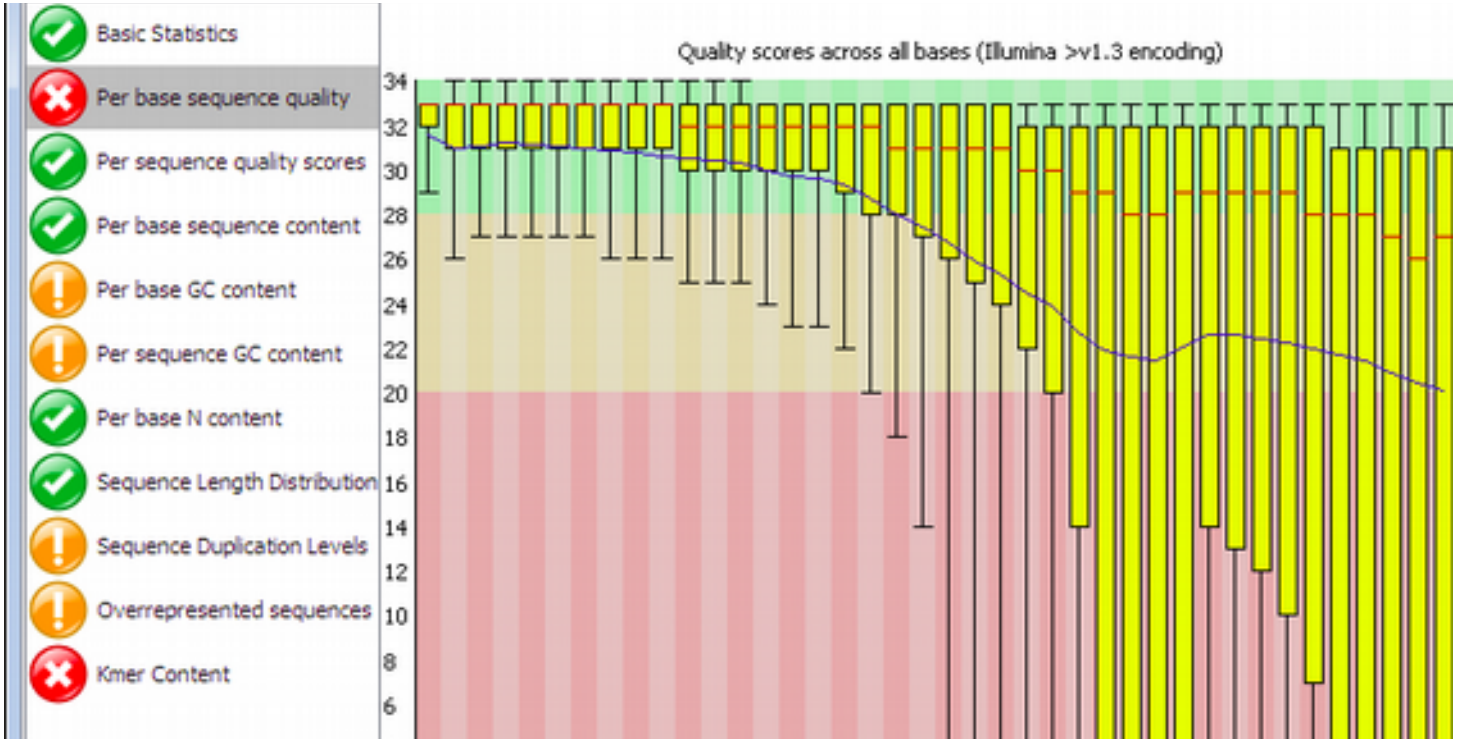


Sequence read:

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

FastQC

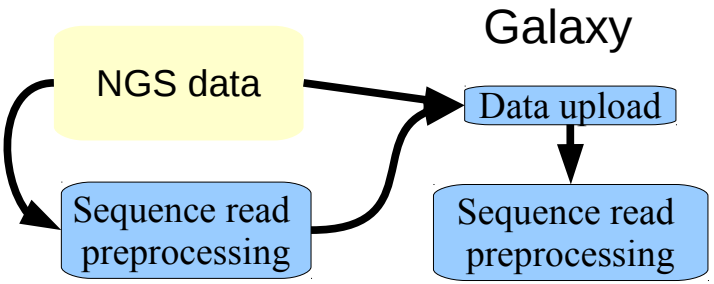
- Galaxy
- GUI based
- Command line



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Sequence read:

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

cutadapt

A tool that removes adapter sequences from DNA sequencing reads
https://code.google.com/p/cutadapt/

- command line
- available in galaxy

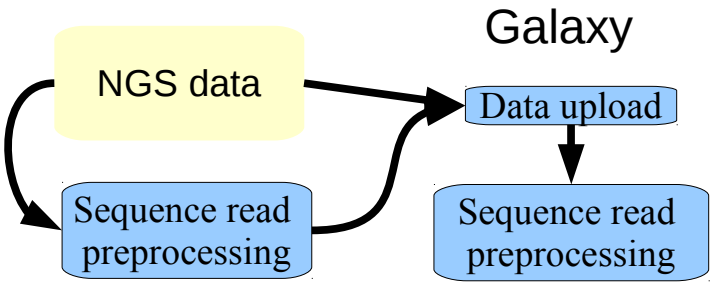
Options:

- adapter sequence
- position (5', 3', anywhere)
- stringency

Sampling:

Reproducible pseudo-random samples
Single/paired reads

Analysis work-flow



Sequence read:

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

Quality filtering criteria – FASTX-toolkit

Quality cut-off value

Quality Score Comparison

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
S - Sanger       Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3 Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa       Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

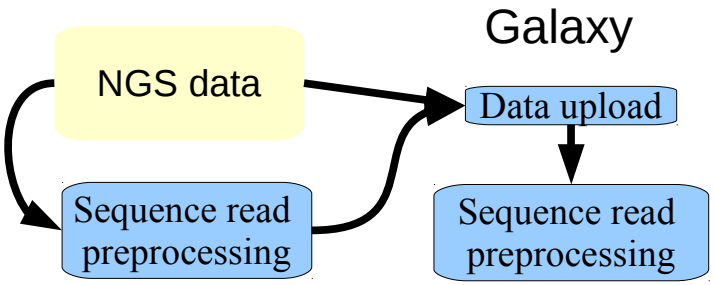
Percent of bases in sequence that must have quality equal to / higher than cut-off value

Proportion of Ns

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

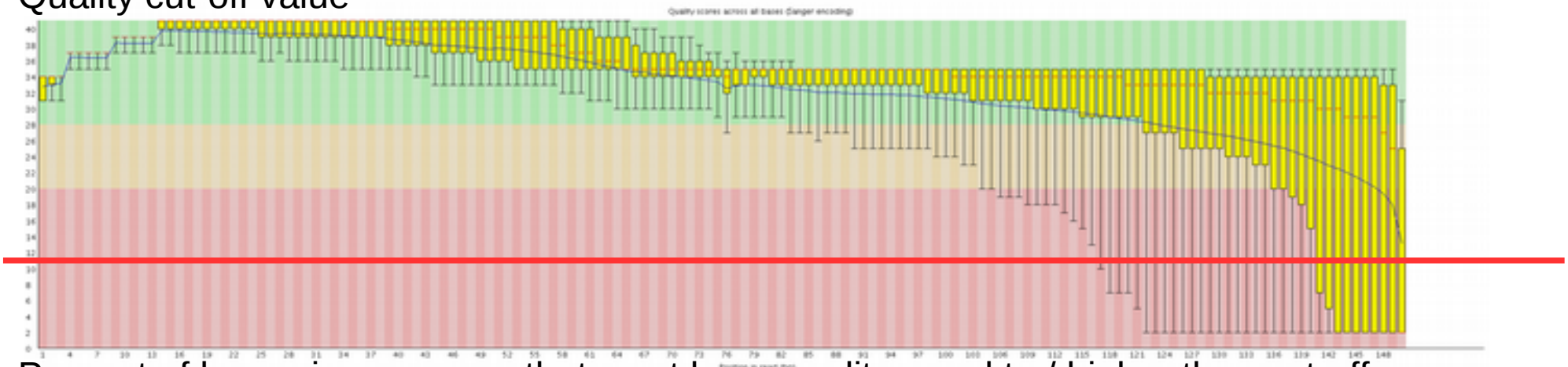


Sequence read:

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

Quality filtering criteria – FASTX-toolkit

Quality cut-off value



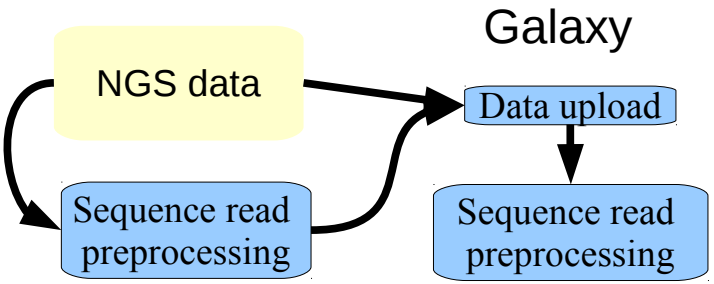
Percent of bases in sequence that must have quality equal to / higher than cut-off value

Proportion of Ns

RepeatExplorer

Discover repeats in your next generation sequencing data

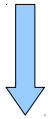
Analysis work-flow



Sequence read:

- 100 nt or more
- Paired
- Uniform length
- Quality control, Grooming
- 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
```



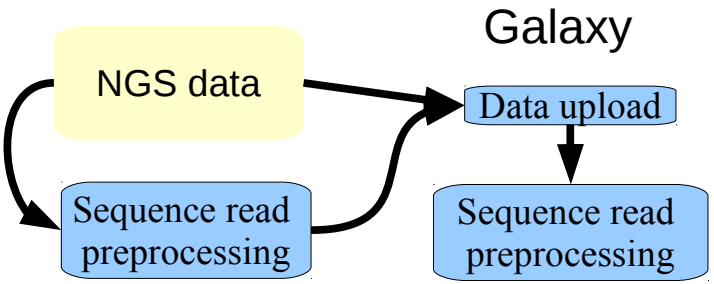
```
>1
acgacagctgactaatgc
>2
cttcgaggctacacgagct
>3
actatcgacactgccggcgcg
```

Simple clustering

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Sequence read:

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

Comparative analysis of multiple genomes/replicates/individuals

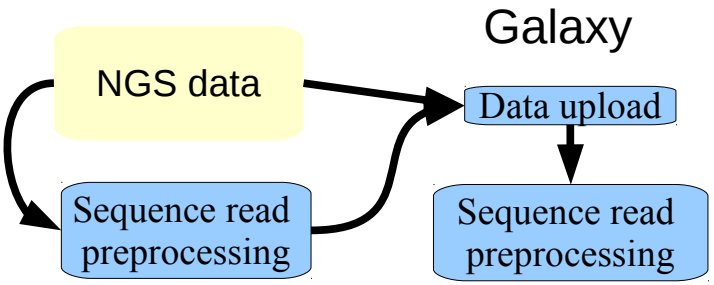
```
Genome AB
>FUXEH4V01EAG68....
acgacagctgactaatgc
>FUXEH4V01BKPDK
cttcgaggctacacgagct
>FUXEH4V01AJAJV
actatcgacactgccggcgcg
...


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

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

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Sequence read:

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

Pair-end reads

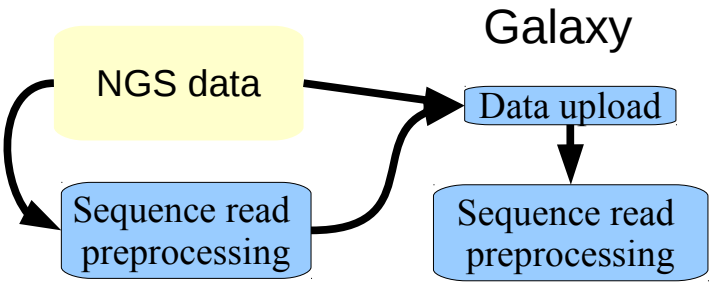
```
Forward reads
>H:1252:2230 1:N:0:CGATGT
acgacagctgactaatgc
>H:1262:1230 1:N:0:CGATGT
cttcgaggctacacgagct
>H:1252:1330 1:N:0:CGATGT
actatcgacactgccggcgcg
...

Reverse reads
>H:1252:2230 2:N:0:CGATGT
gccccgtcgccgtccgtgtcg
>H:1262:1230 2:N:0:CGATGT
tgtgtgcccgtctgcgcgccccc
>H:1252:1330 2:N:0:CGATGT
atatgctatgcgcgc
...
```

interlacing

```
Pair-end reads clustering:
>1f
acgacagctgactaatgc
>1r
gccccgtcgccgtccgtgtcg
>2f
cttcgaggctacacgagct
>2r
tgtgtgcccgtctgcgcgccccc
>3f
actatcgacactgccggc
>3r
atatgctatgcgcgc
```

Remove reads with no pair

Analysis work-flow



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

Pair-end reads

Fastq interlacer – Galaxy built in tool
- Does not expect ordered sequences
- Memory demanding

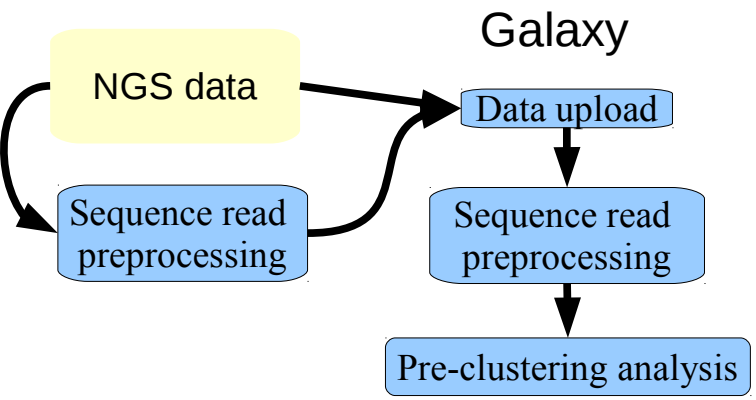
Fasta interlacer – RepeatExplorer tool
- Assumes ordered sequences
- Low memory requirements
- Fast

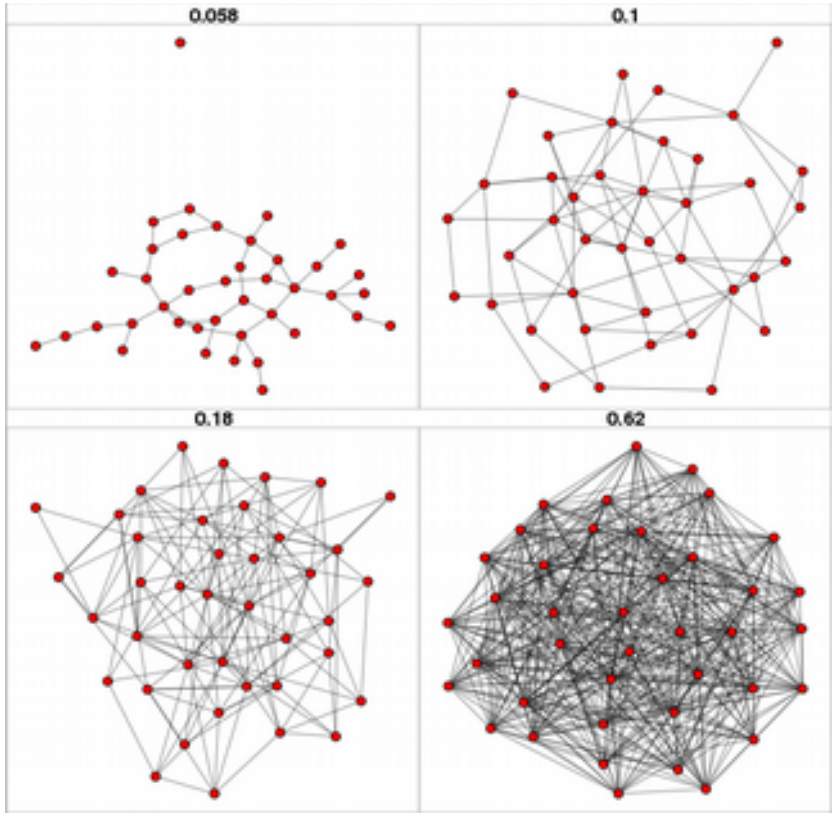
Original Illumina reads are always ordered!

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow





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

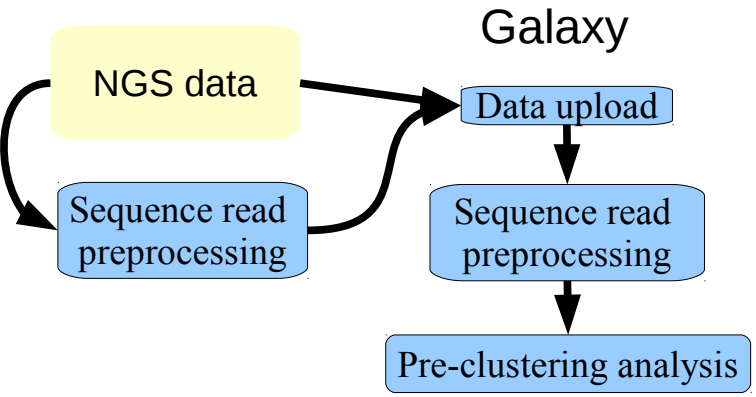
Density corresponds to probability that two randomly chosen sequences from genome will be similar

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

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



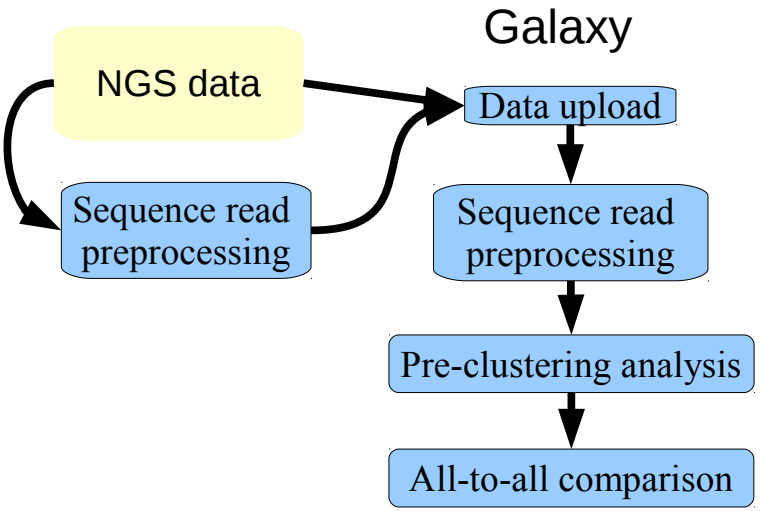
Maximum number of reads which can be processed depends on genome compositions:

- Number of various repeats
- Copy number
- Length
- Diversity

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

Analysis work-flow



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!

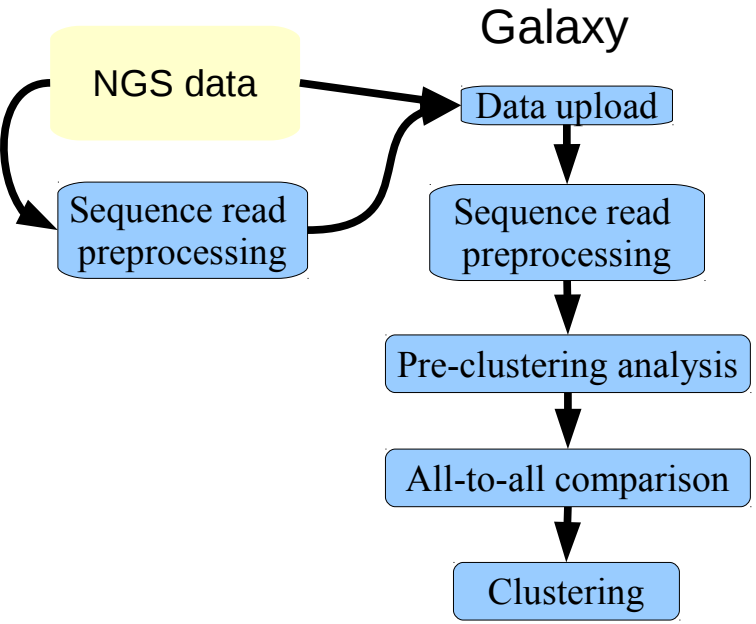
Length of overlap must be adjusted for reads shorter Then 100 nt

55% of length and 90% similarity is hardcoded – cannot be changed

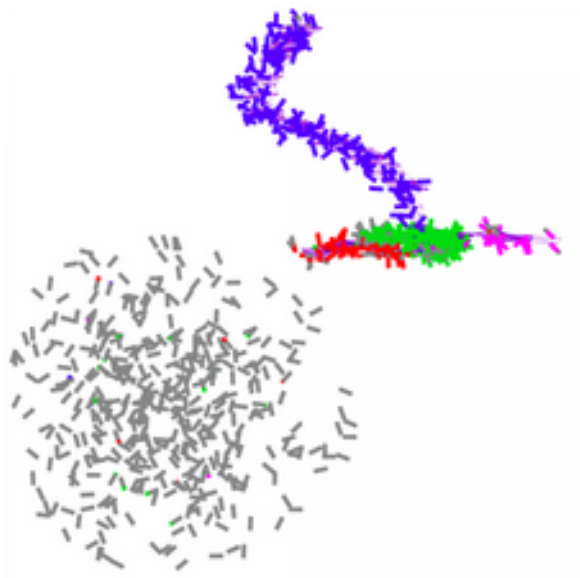
RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



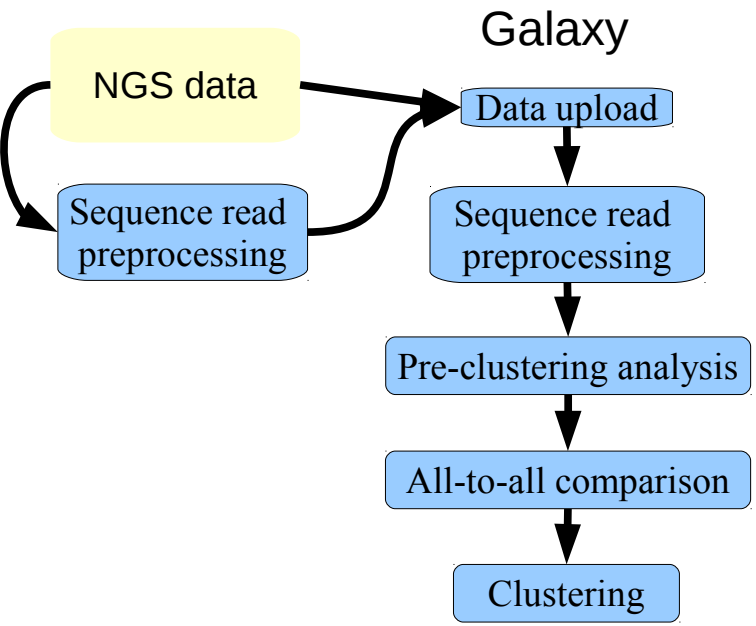
Louvain clustering



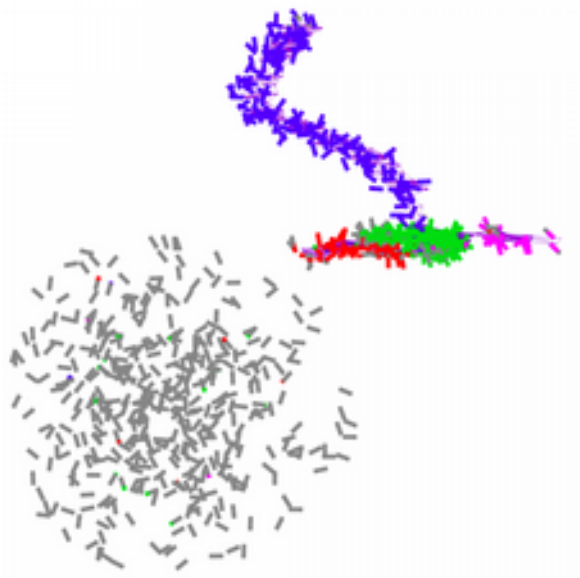
RepeatExplorer

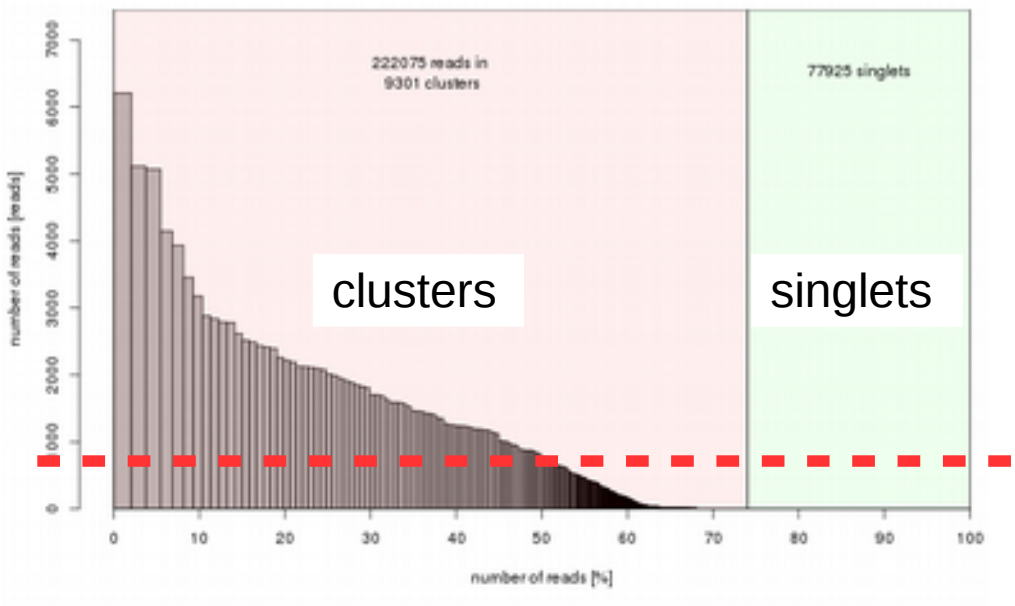
Discover repeats in your next generation sequencing data

Analysis work-flow



Louvain clustering

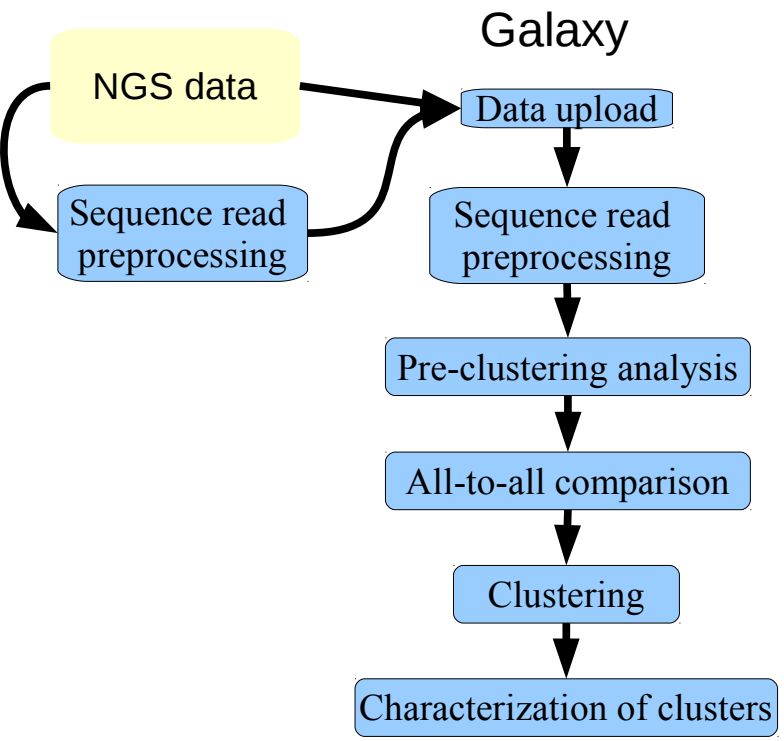




RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



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

Methods:

- **RepeatMasker**

RepBase (Default)

Custom database (Optional)
Use specified format!

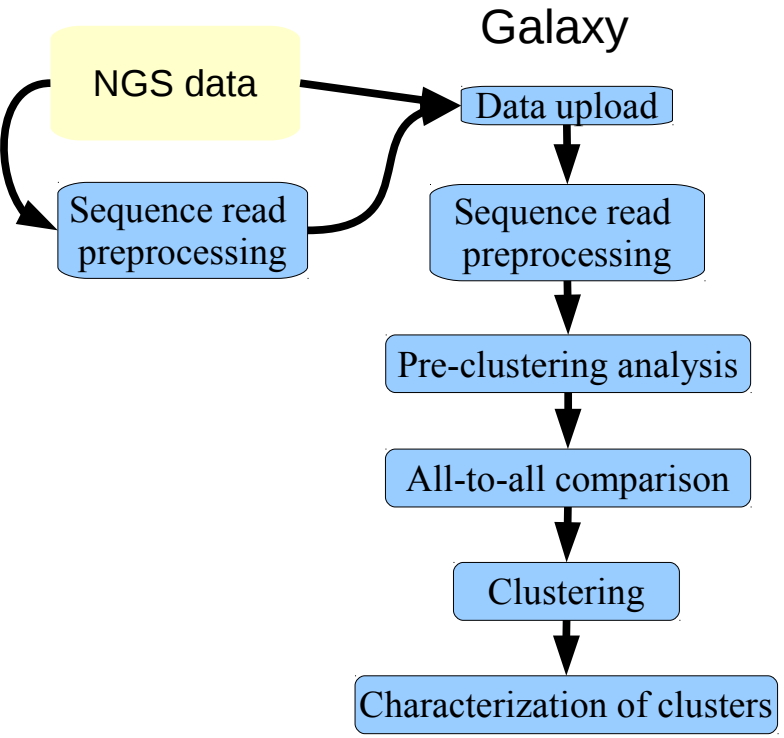
```
>reapeatname#class/subclass
acgatcgatcghagsctcgacagtcgatca
>reapeatname#class/subclass
acgatcgatcghagsctcgacagtcgatca
```

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

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



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

Methods:

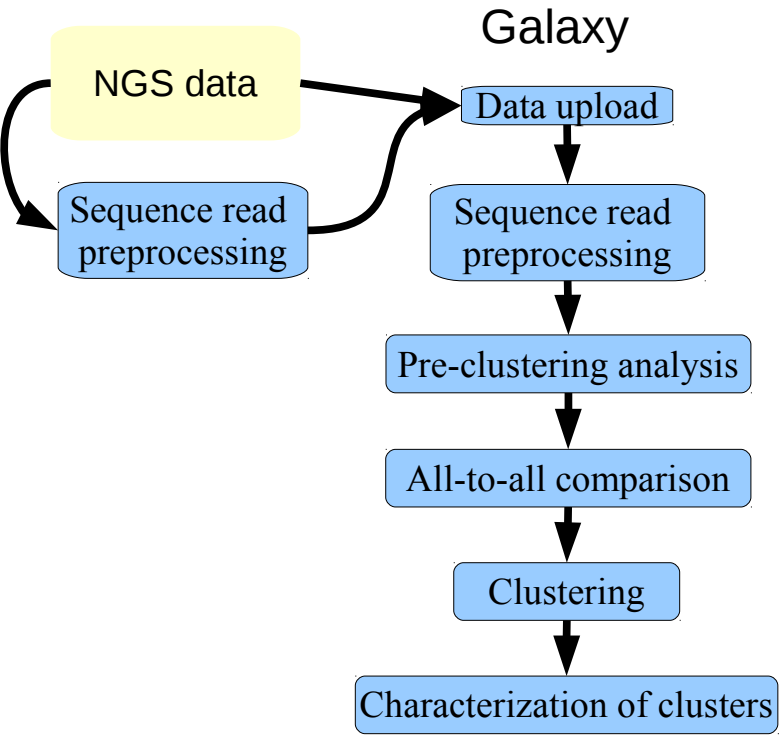
- RepeatMasker
- **RPS BLAST against conserved domain database**

Very slow!
This feature will be removed in future

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



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

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains

Protein sequences are derived from conserved domains of transposable elements

Manually curated

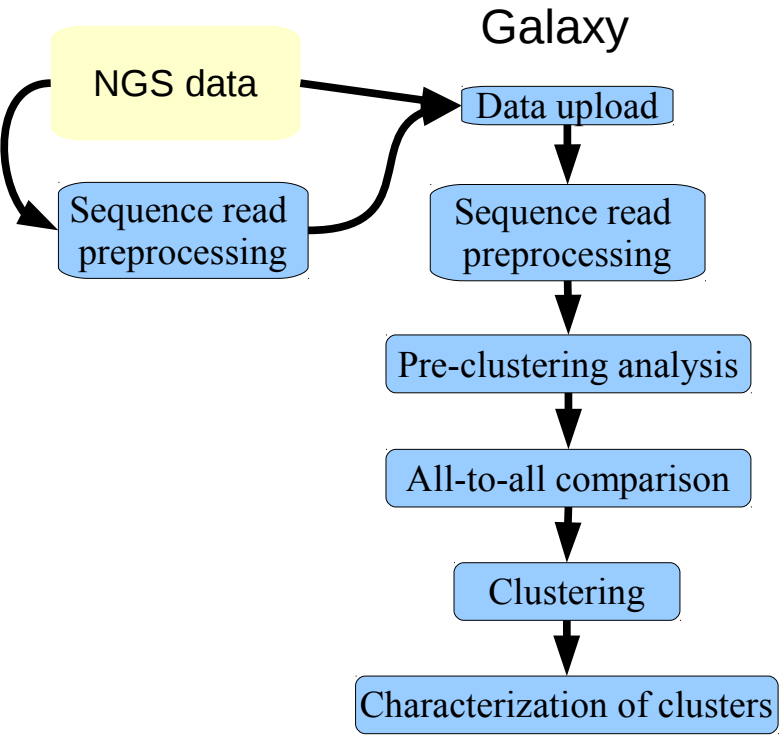
The most informative

Currently “plant” oriented

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



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

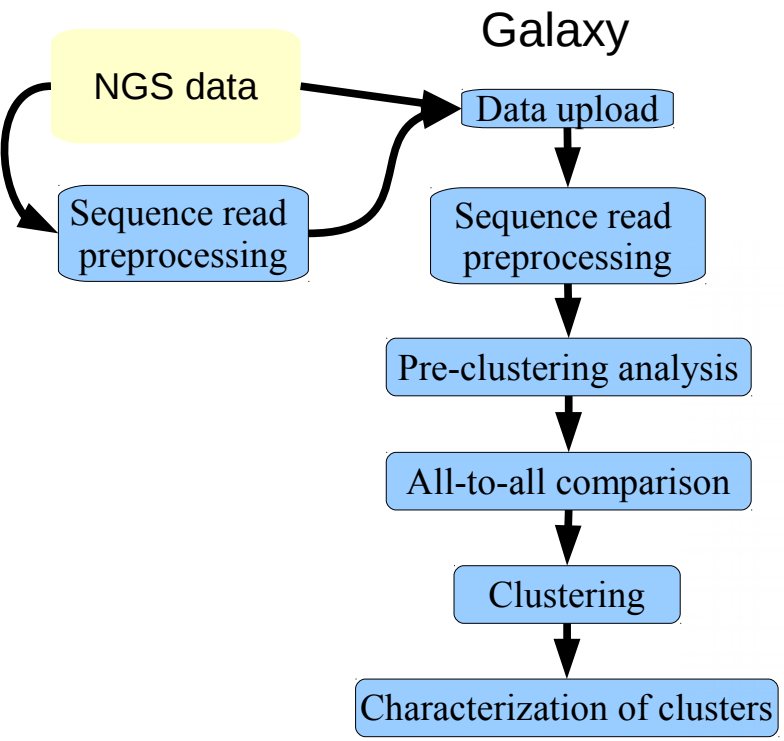
Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- **Detection of LTR**

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

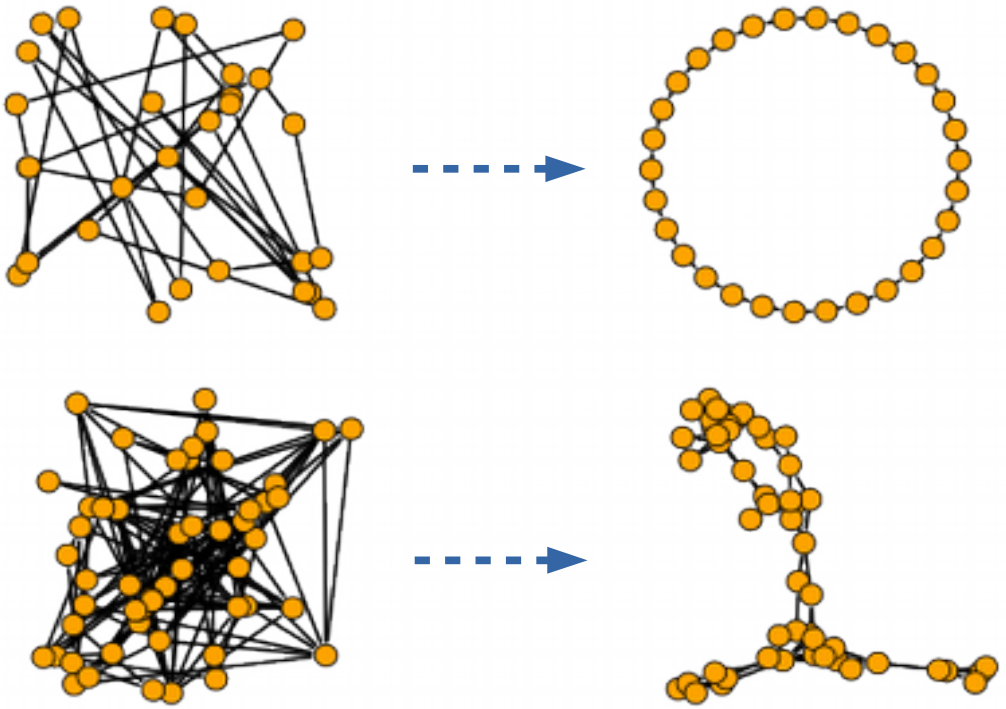


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

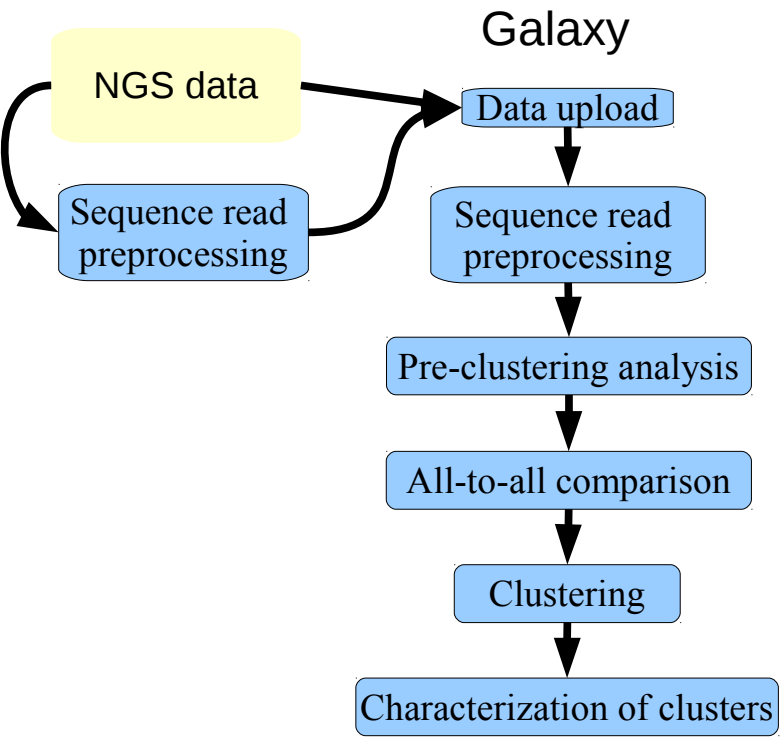
Fruchterman-Reingold



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

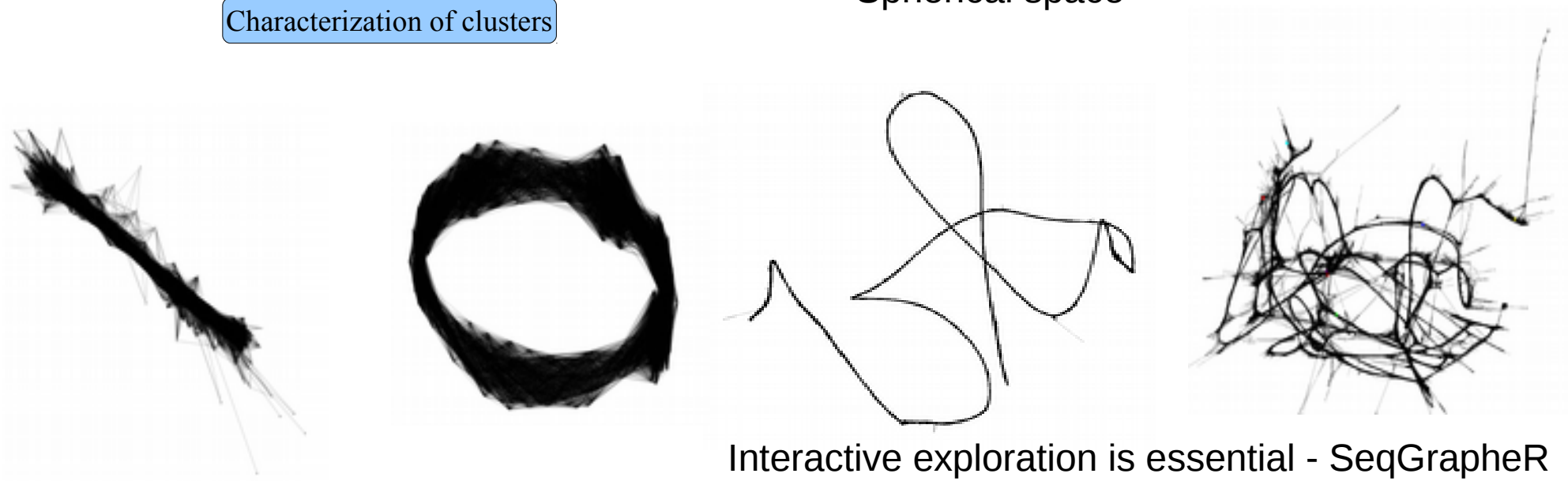


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

Fruchterman-Reingold algorithm
2D projection of 3D layout
Spherical space

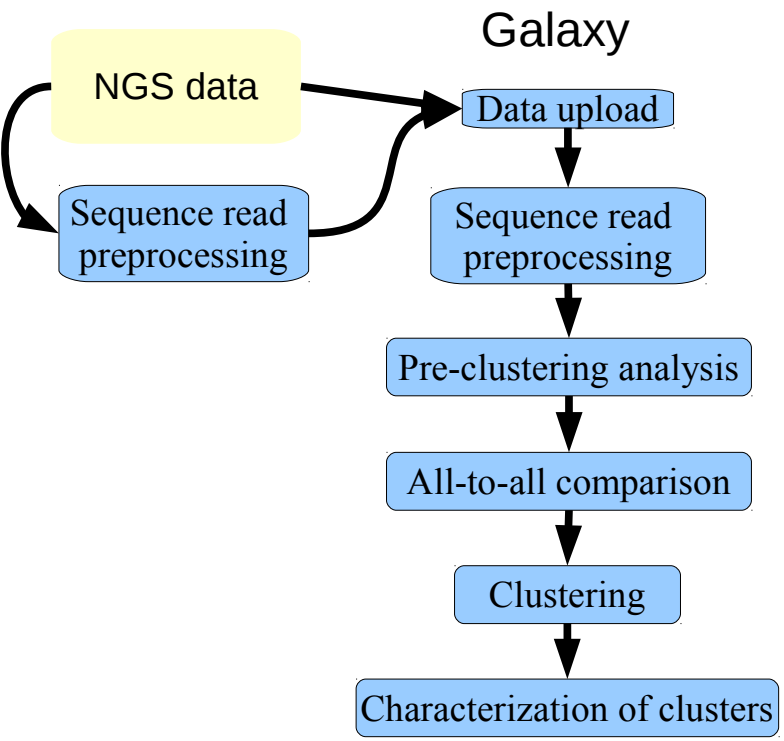


Interactive exploration is essential - SeqGrapheR

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

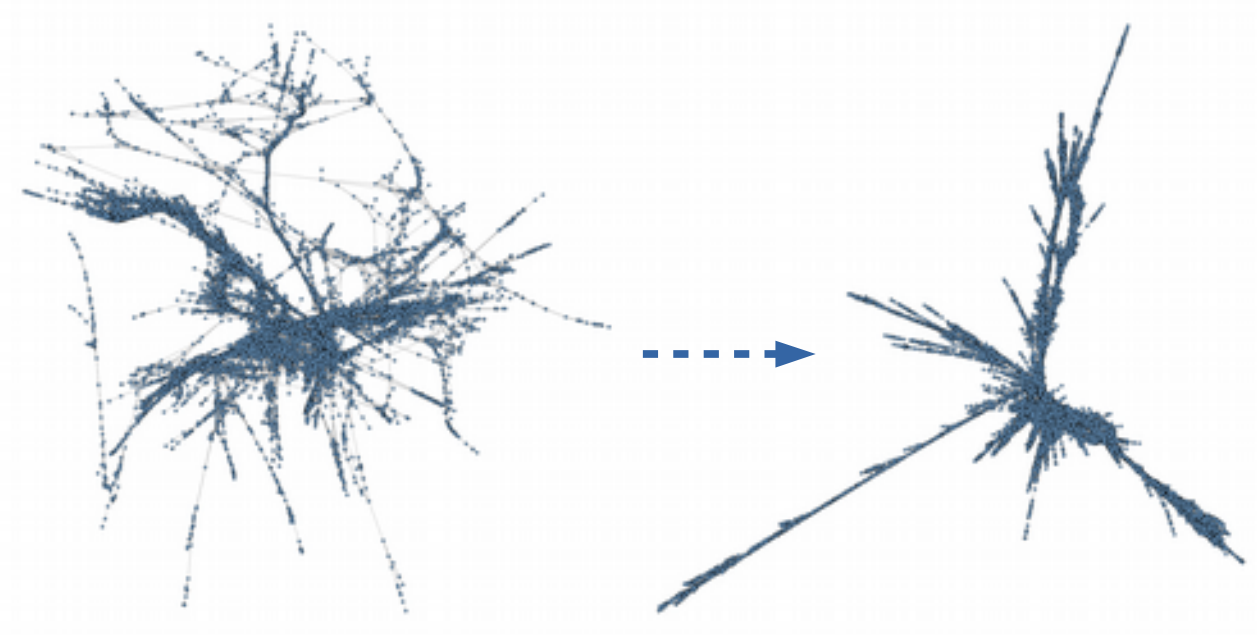


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

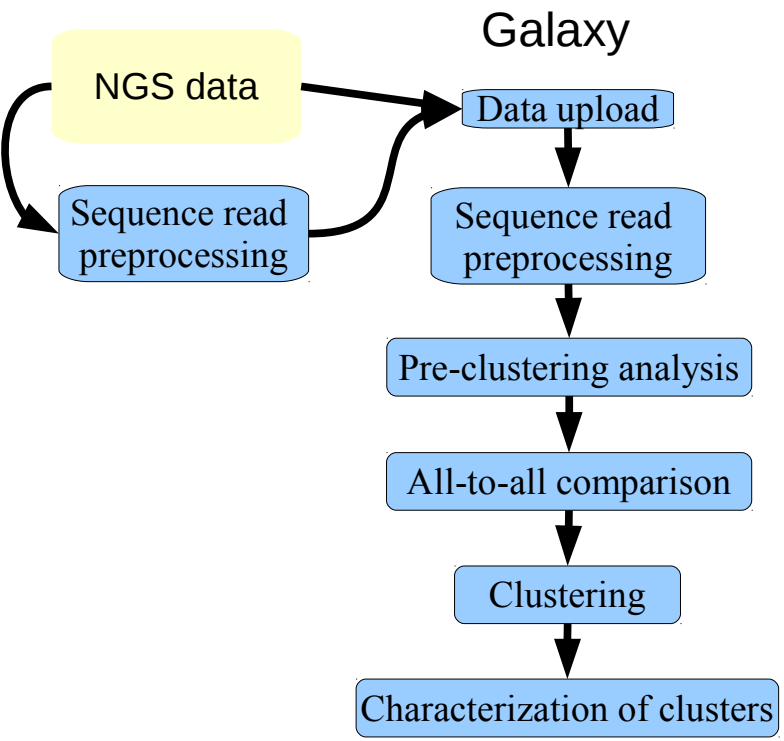
New layout algrithm from OGDF http://www.ogdf.net



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

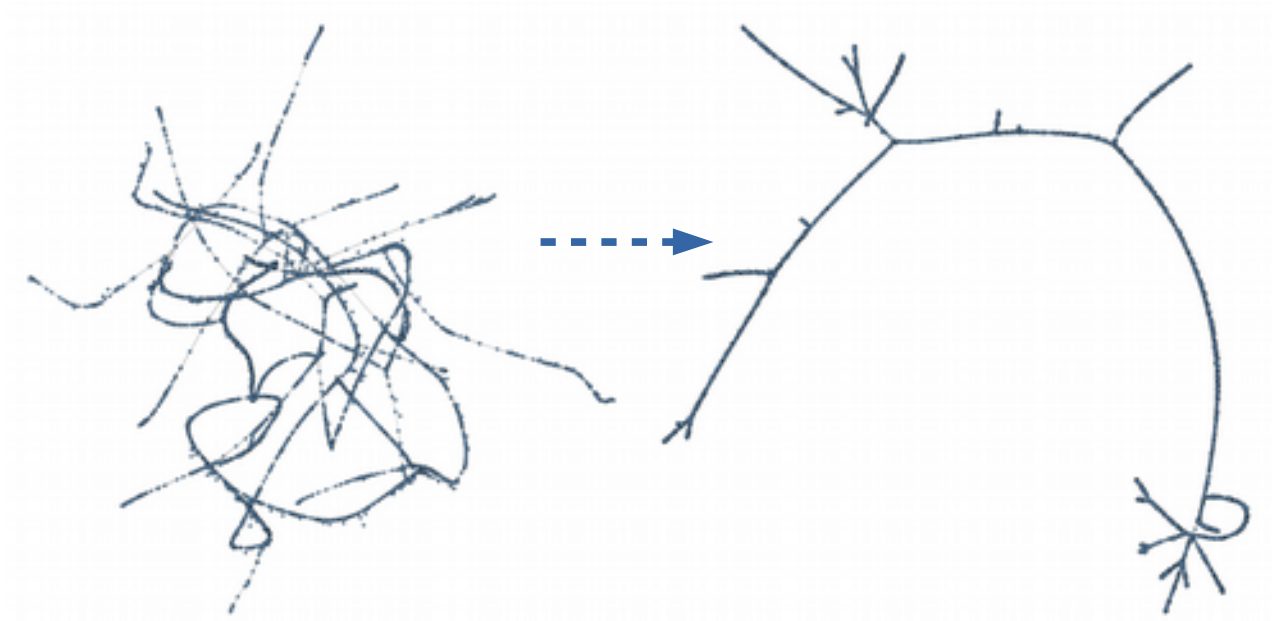


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

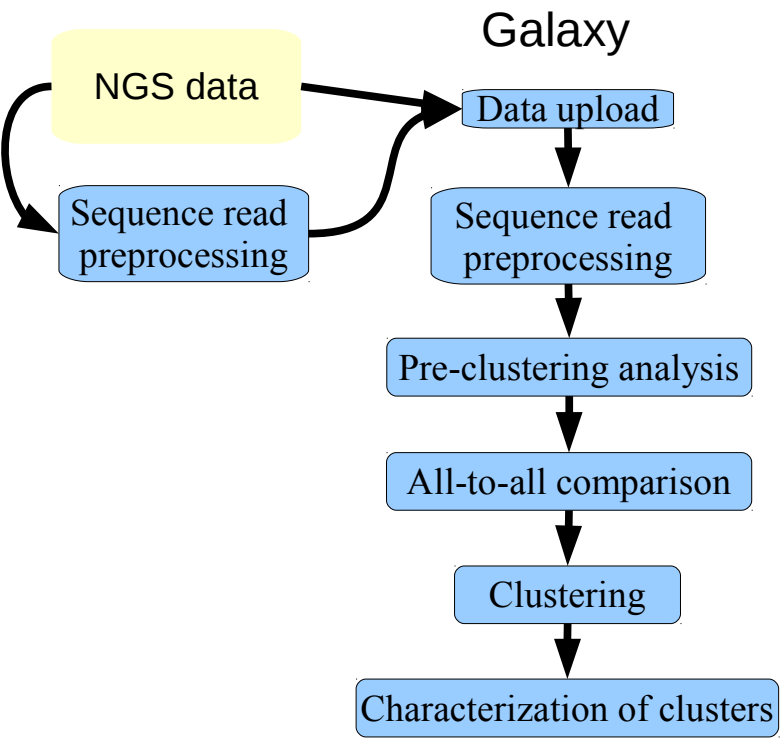
New layout algrithm from OGDF http://www.ogdf.net



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

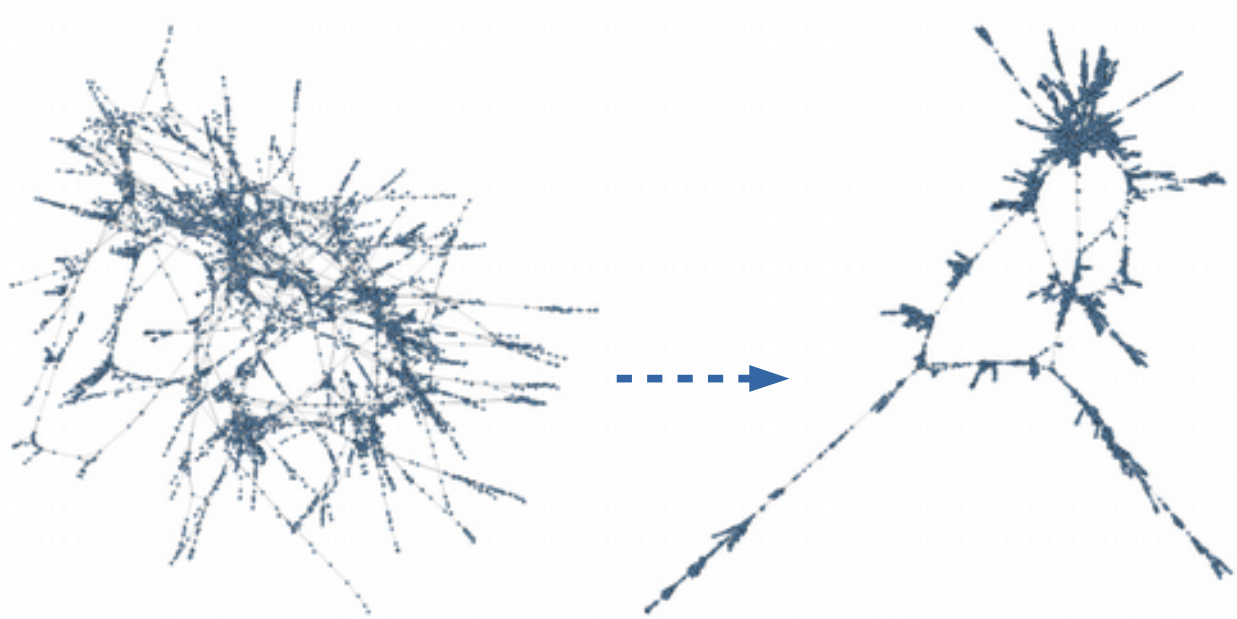


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

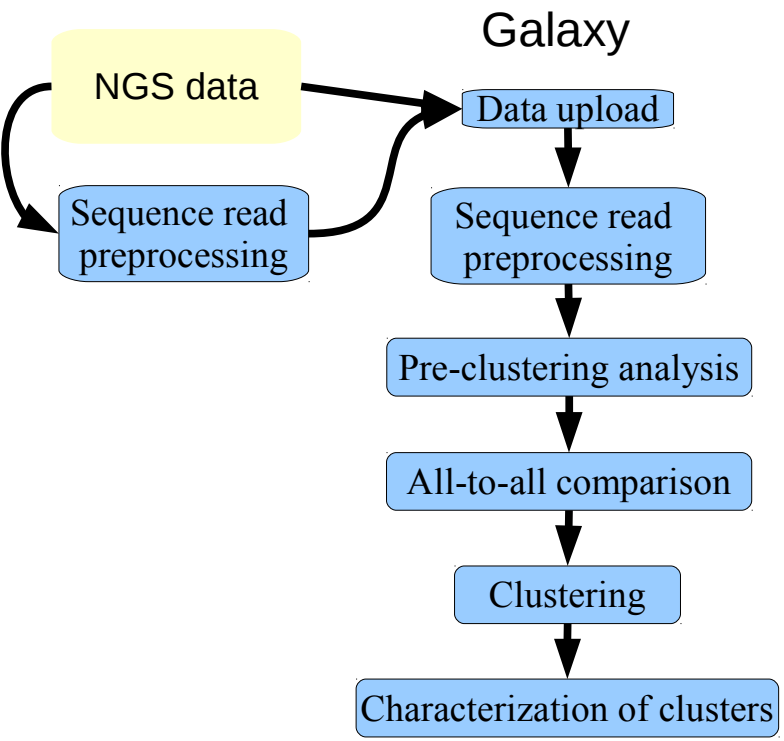
New layout algrithm from OGDF http://www.ogdf.net



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

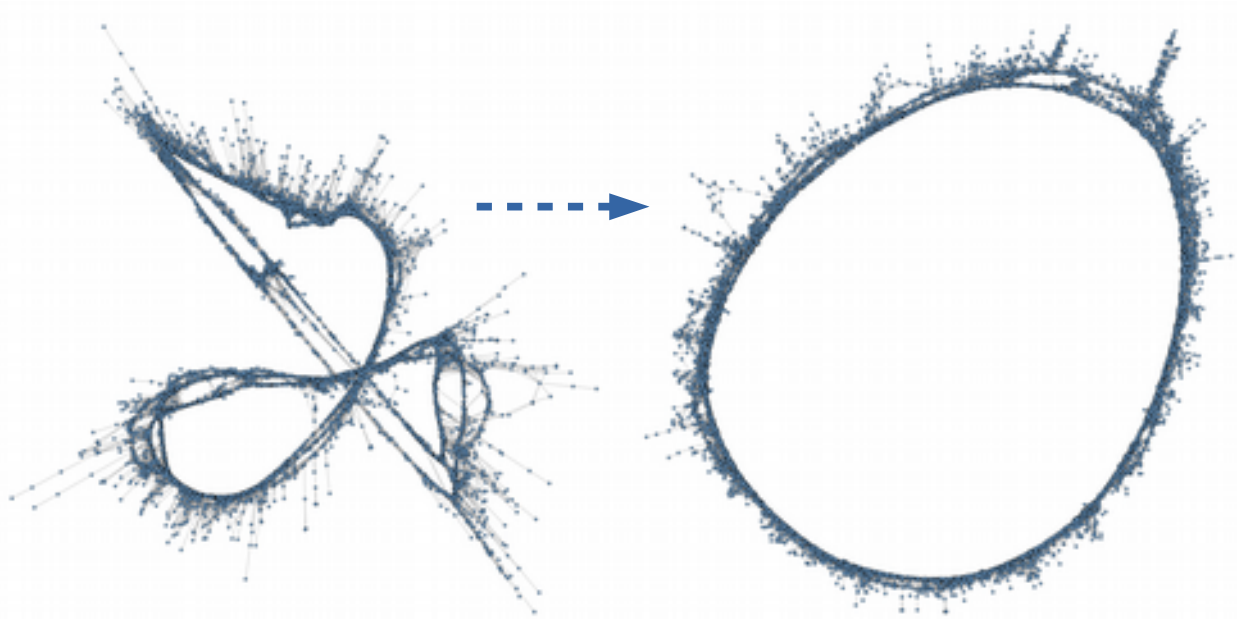


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

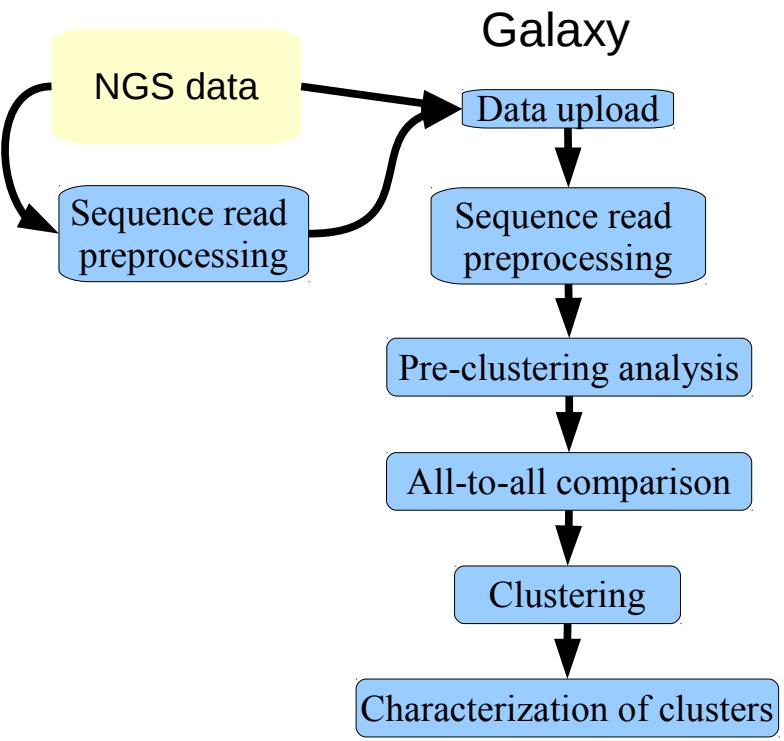
New layout algrithm from OGDF http://www.ogdf.net



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

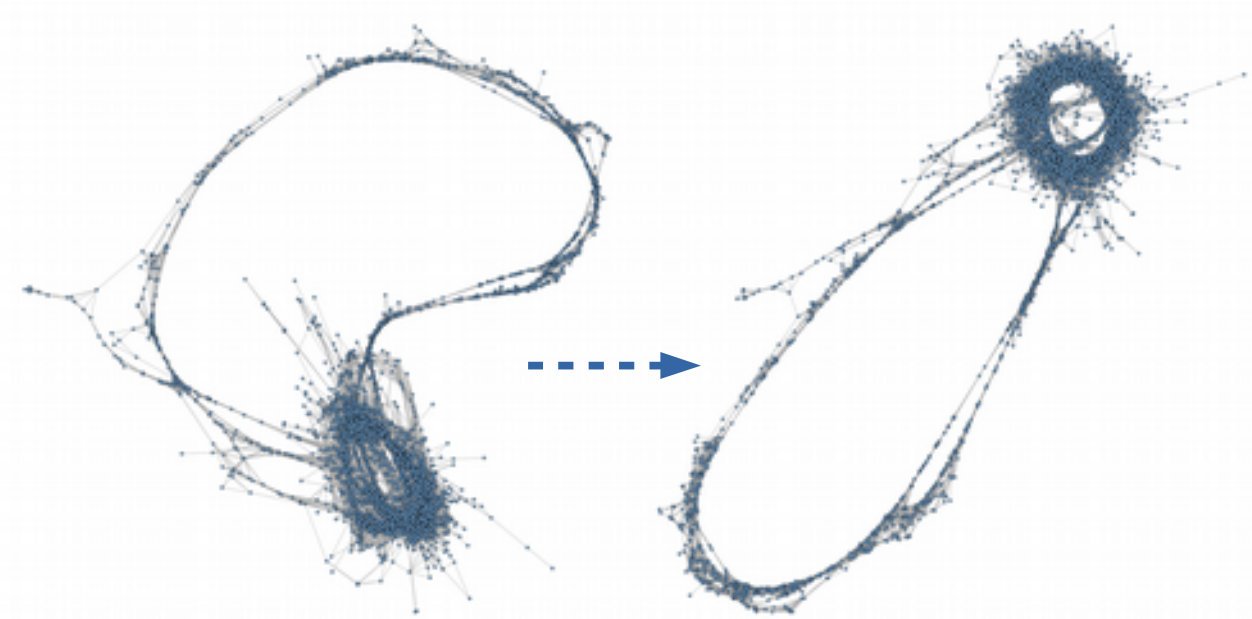


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

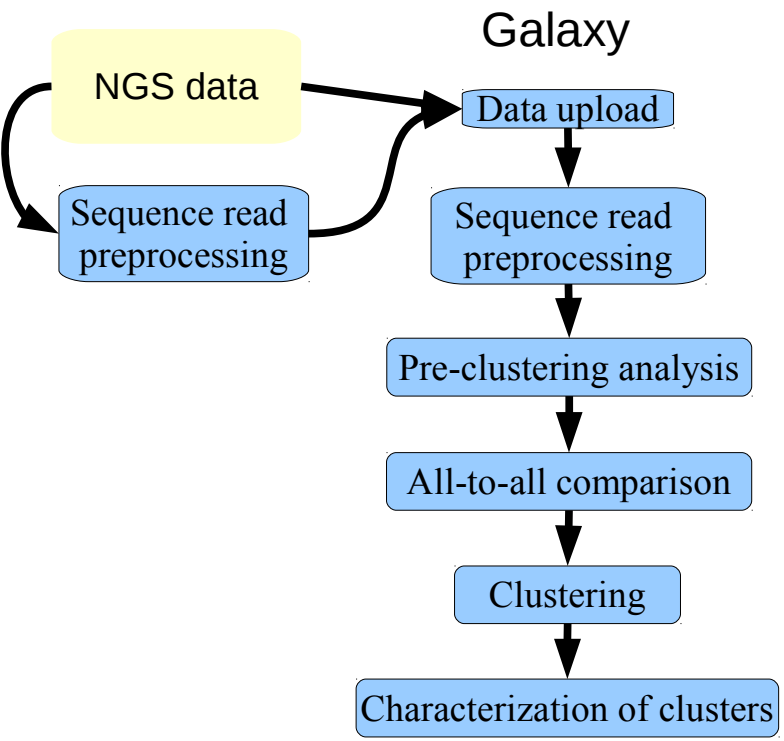
New layout algrithm from OGDF http://www.ogdf.net



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

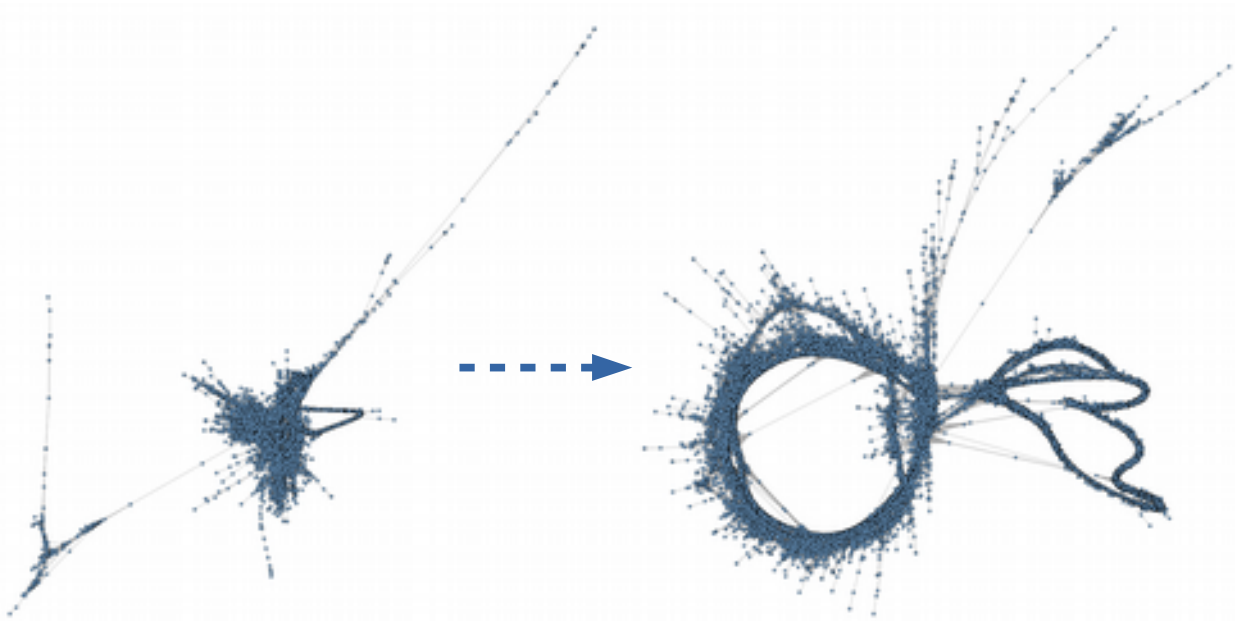


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- **Graph layout calculation**

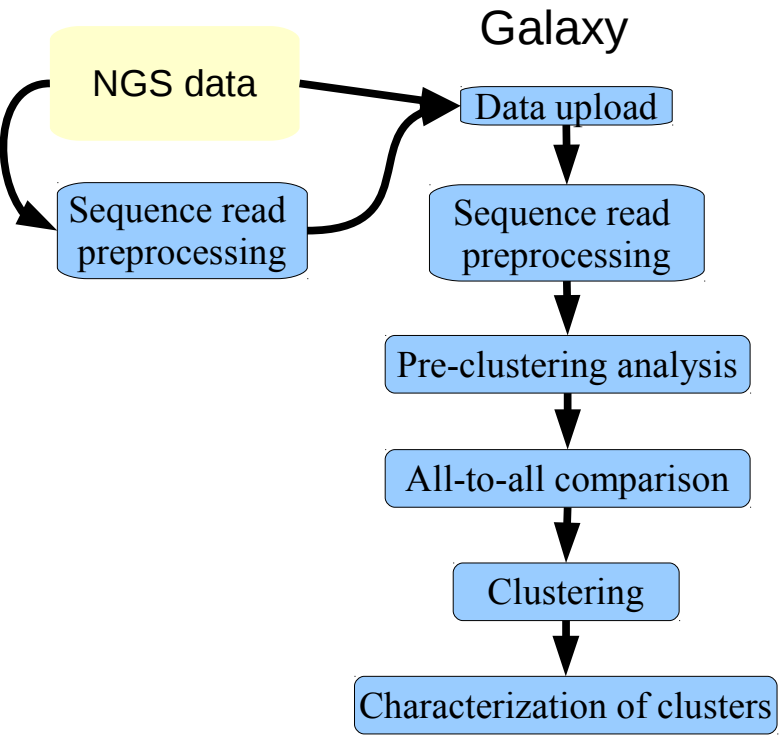
New layout algrithm from OGDF http://www.ogdf.net



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



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

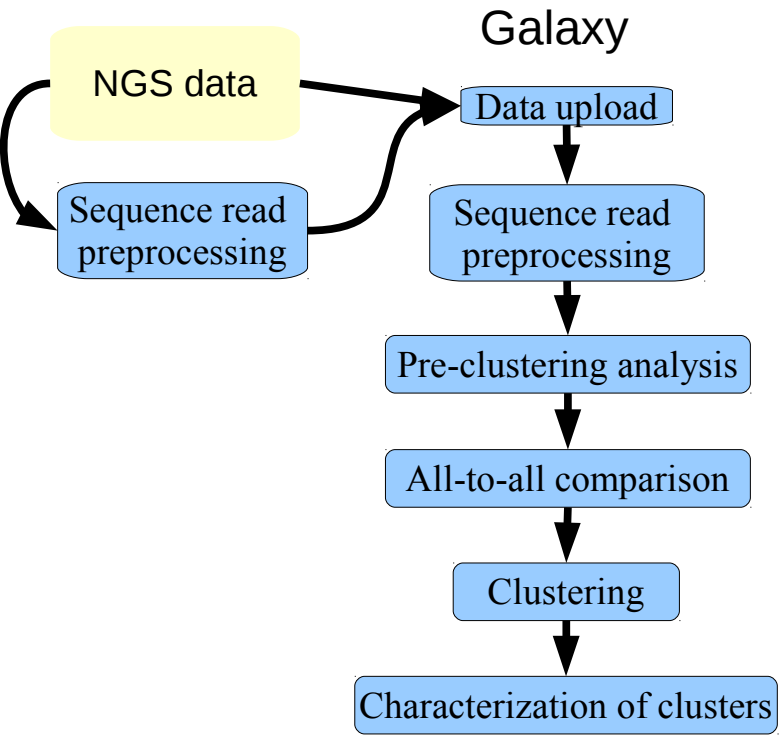
Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- Graph layout calculation
- **Similarity between clusters**

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

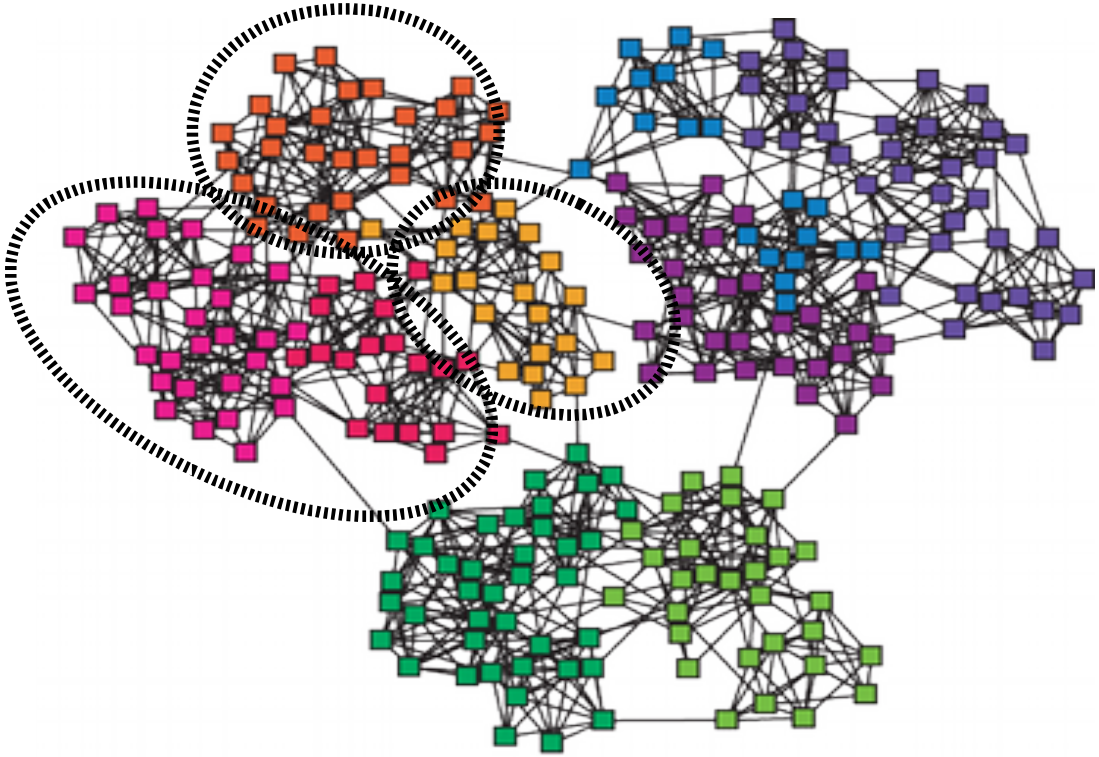


One type of repeat - often split up into multiple clusters

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

Methods:

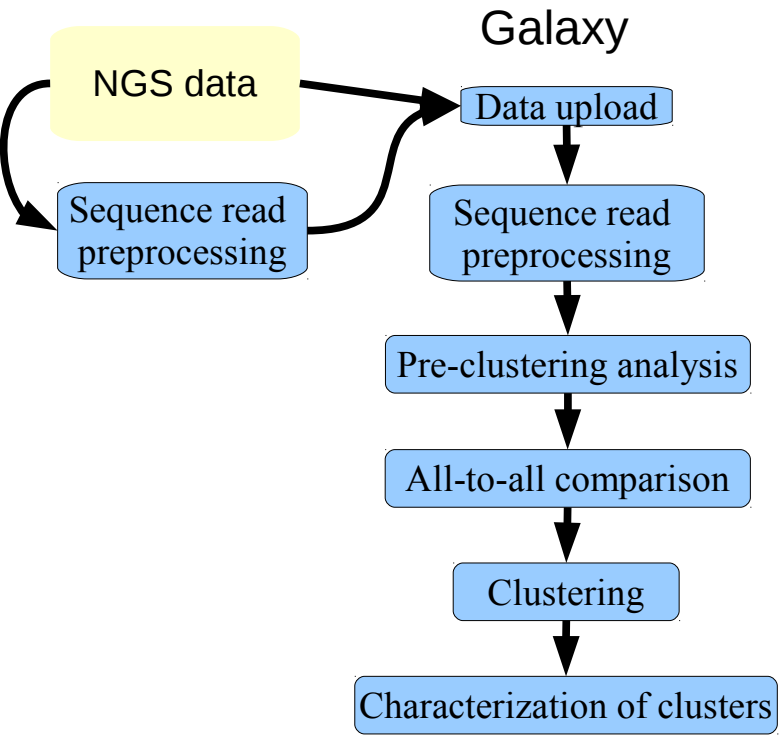
- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- Graph layout calculation
- **Similarity between clusters**



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

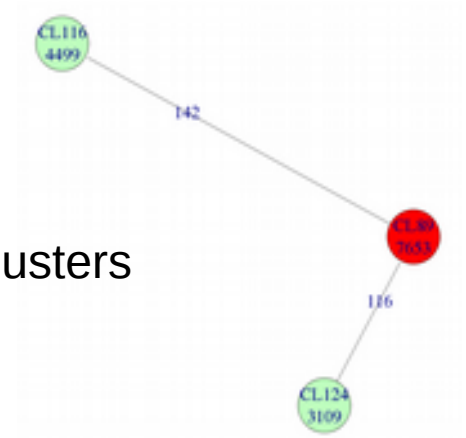


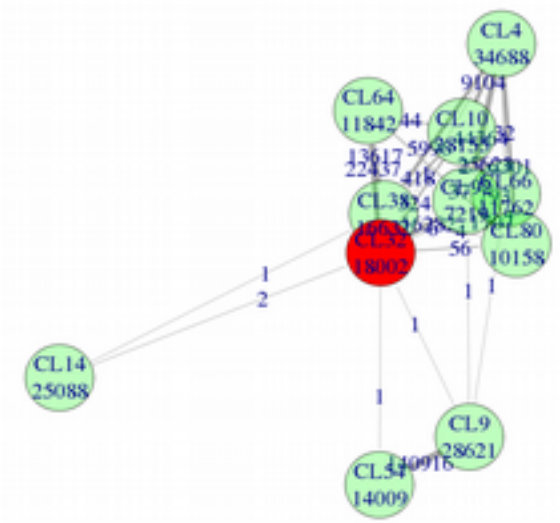
Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- Graph layout calculation
- **Similarity between clusters**

Number of similarity hits between clusters

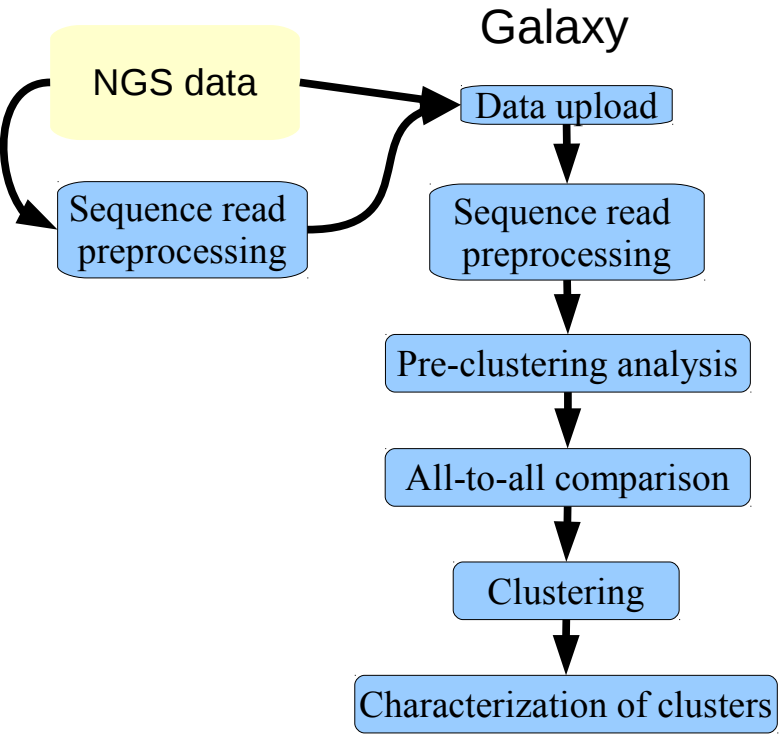




RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

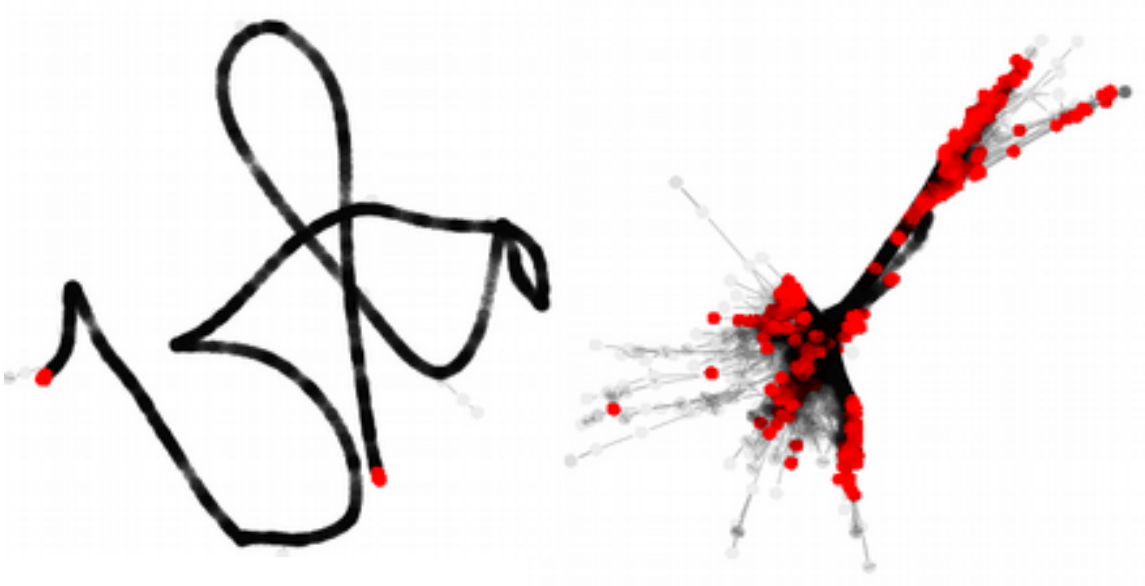


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- Graph layout calculation
- **Similarity between clusters**

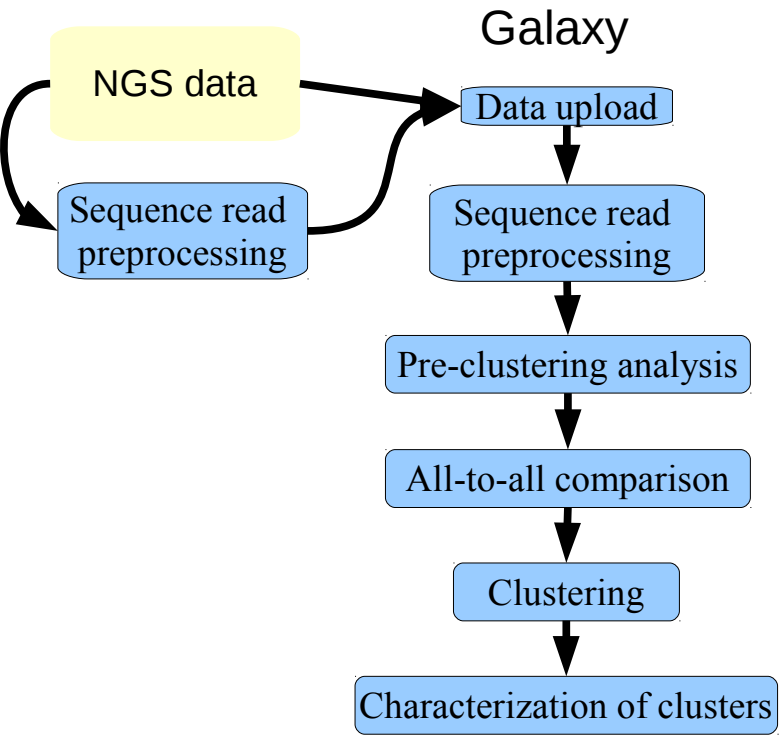
Distribution of similarity hits to other clusters



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

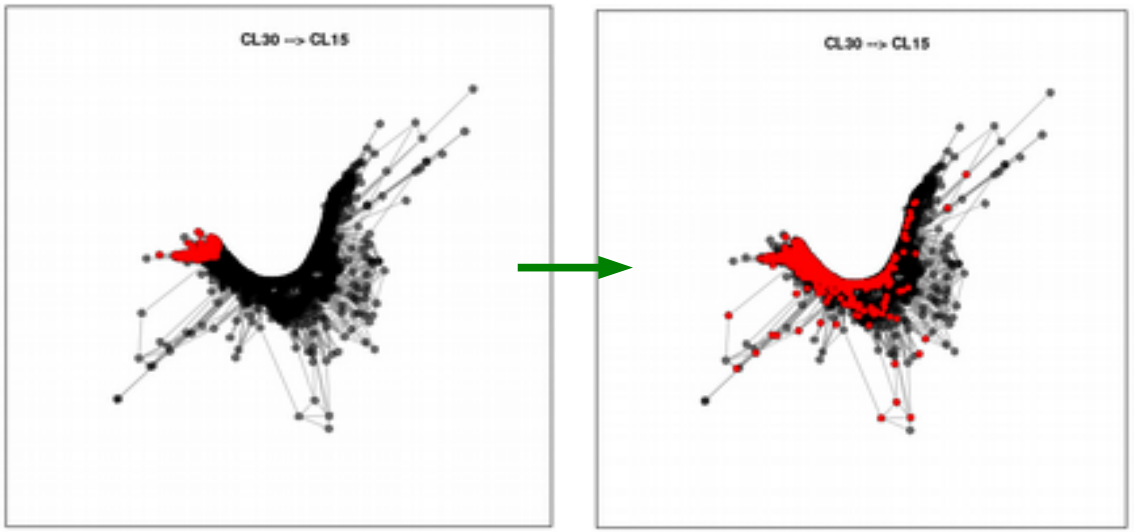


Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- Graph layout calculation
- **Similarity between clusters**

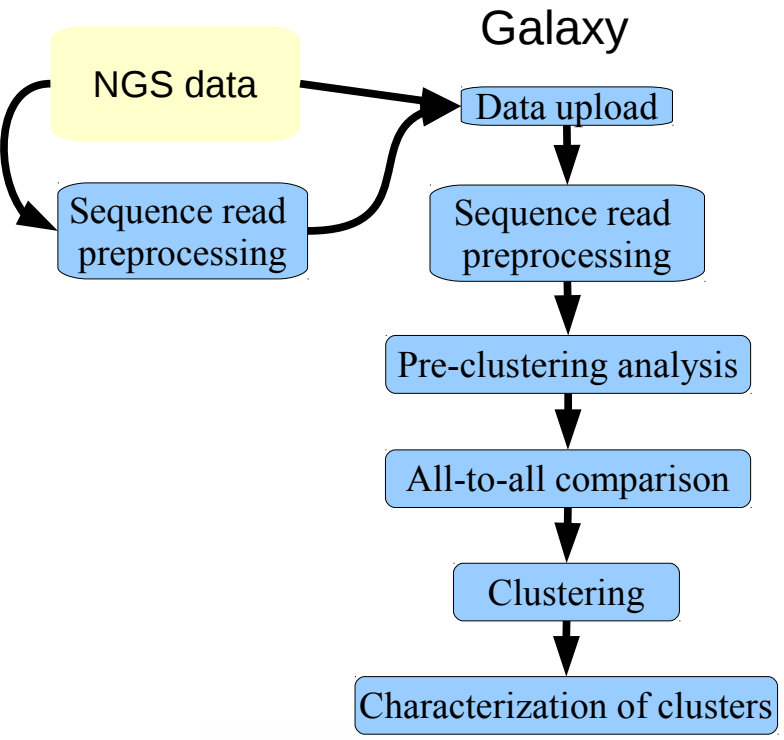
Similarity **Read pairs**



RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

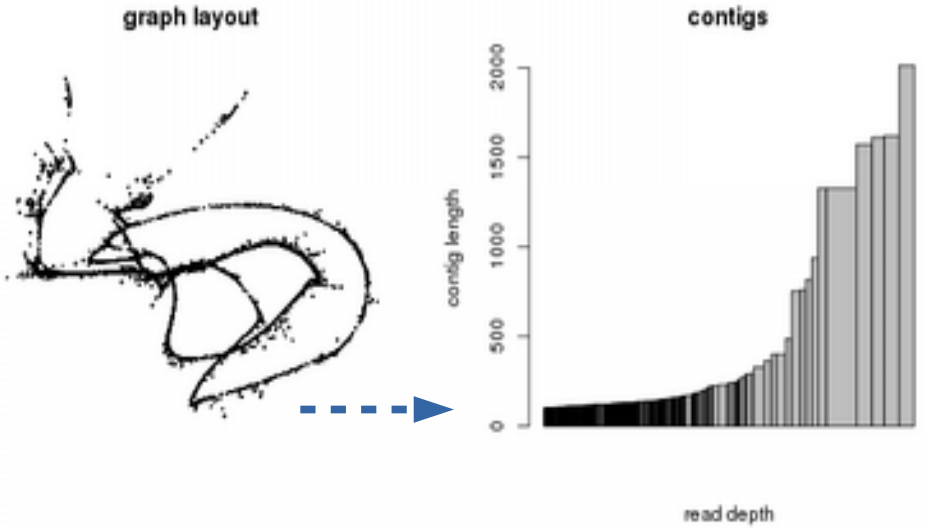


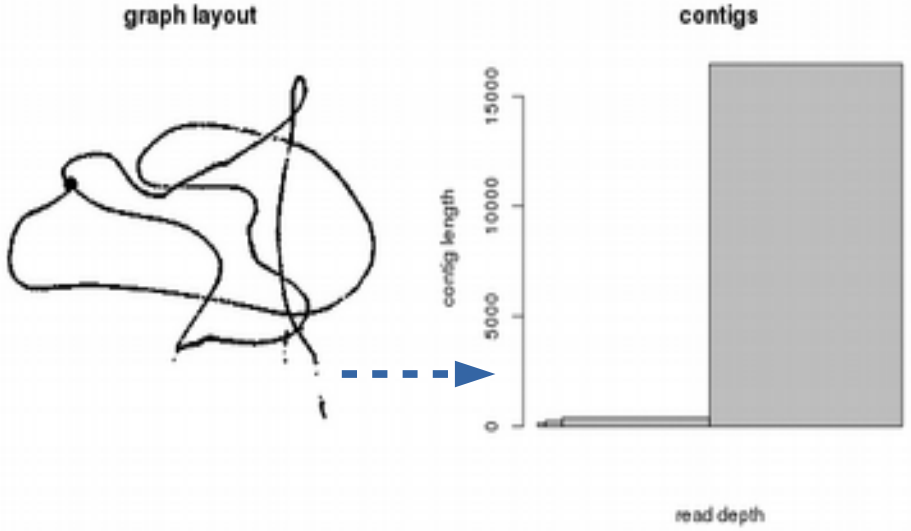
Characterization of clusters by various method to provide detailed information for annotation of clusters and identification of repeats

Methods:

- RepeatMasker
- RPS BLAST against conserved domain database
- Blastx search against database of protein domains
- Detection of LTR
- Graph layout calculation
- Similarity between clusters
- **Assembly**

CAP3 program
Each cluster is assembled separately

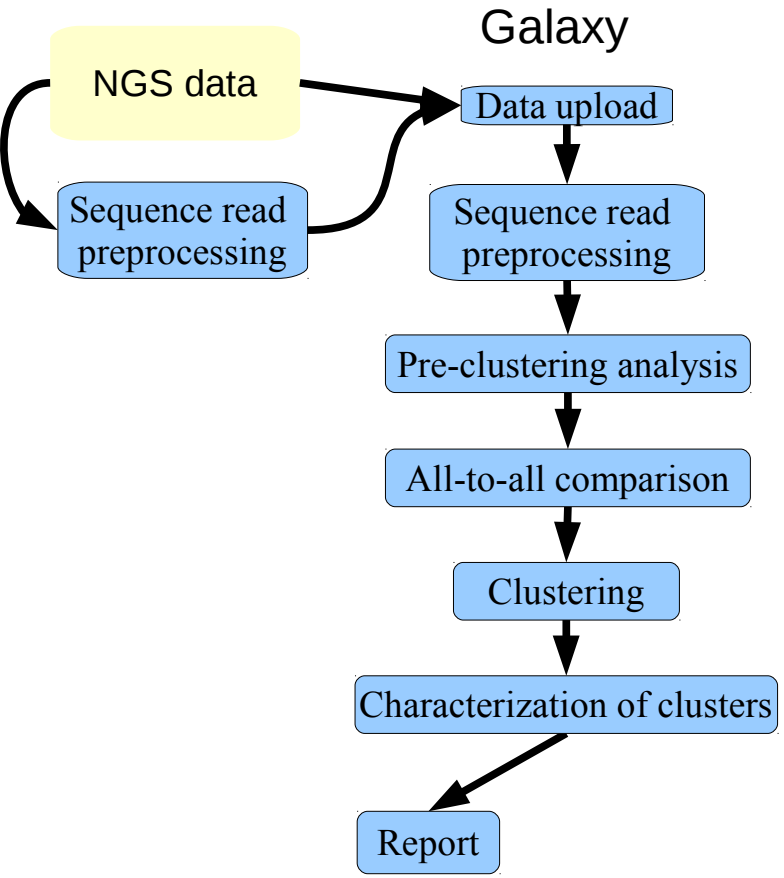




RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



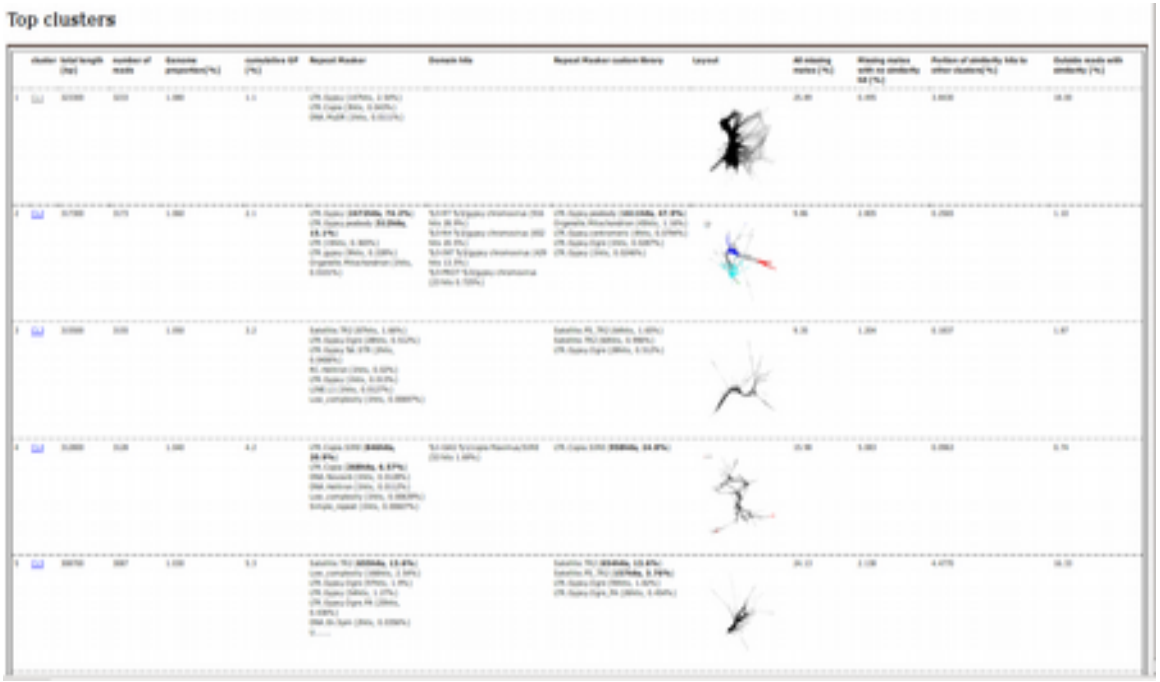
Output

Log file – information about pipeline progress

HTML report – user friendly graphical overview of top clusters

Archive with complete results

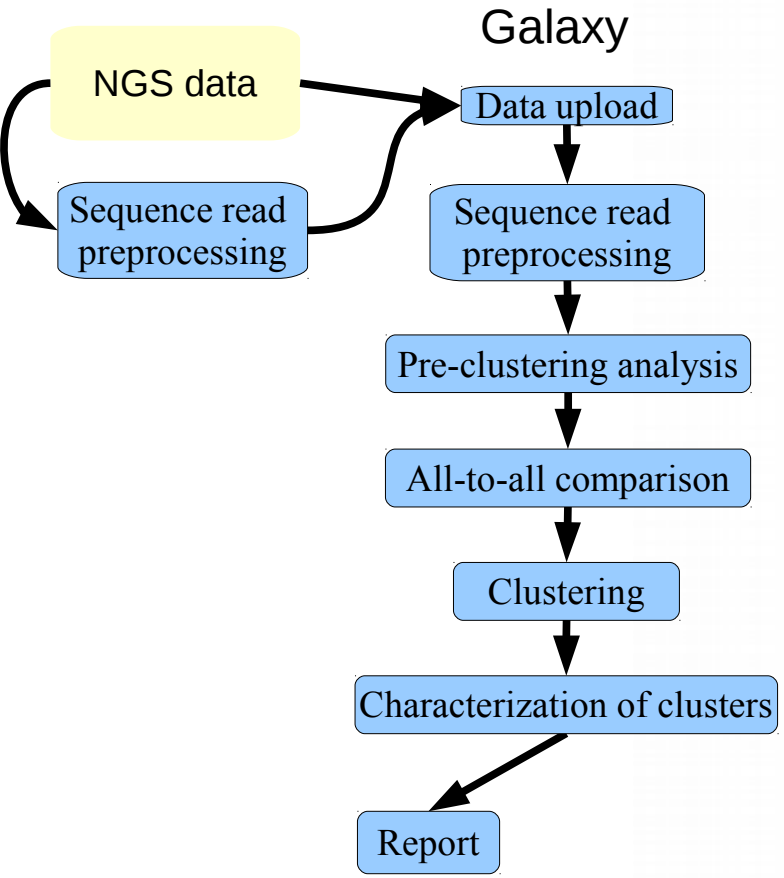
Contigs assembled from clusters – can be used directly in follow up analysis of protein domains

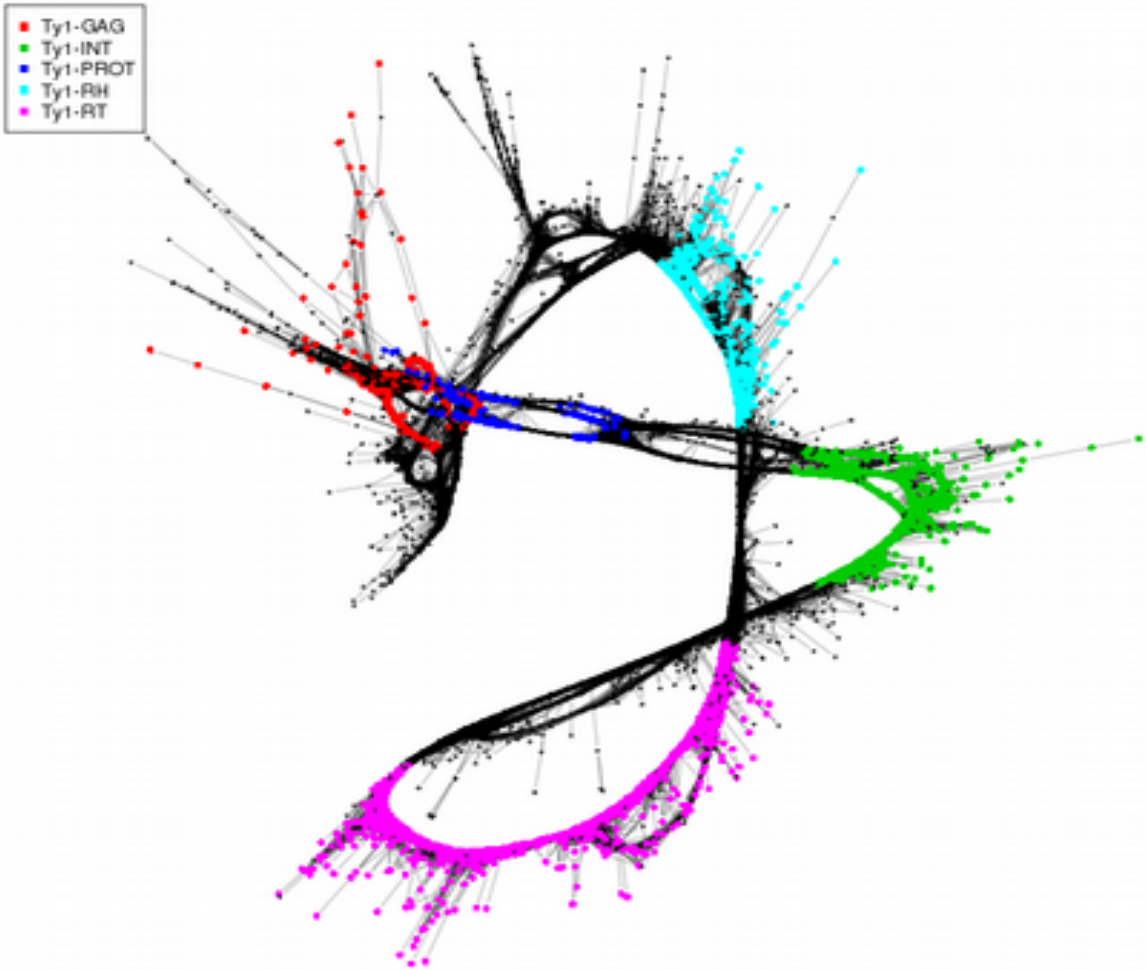


RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

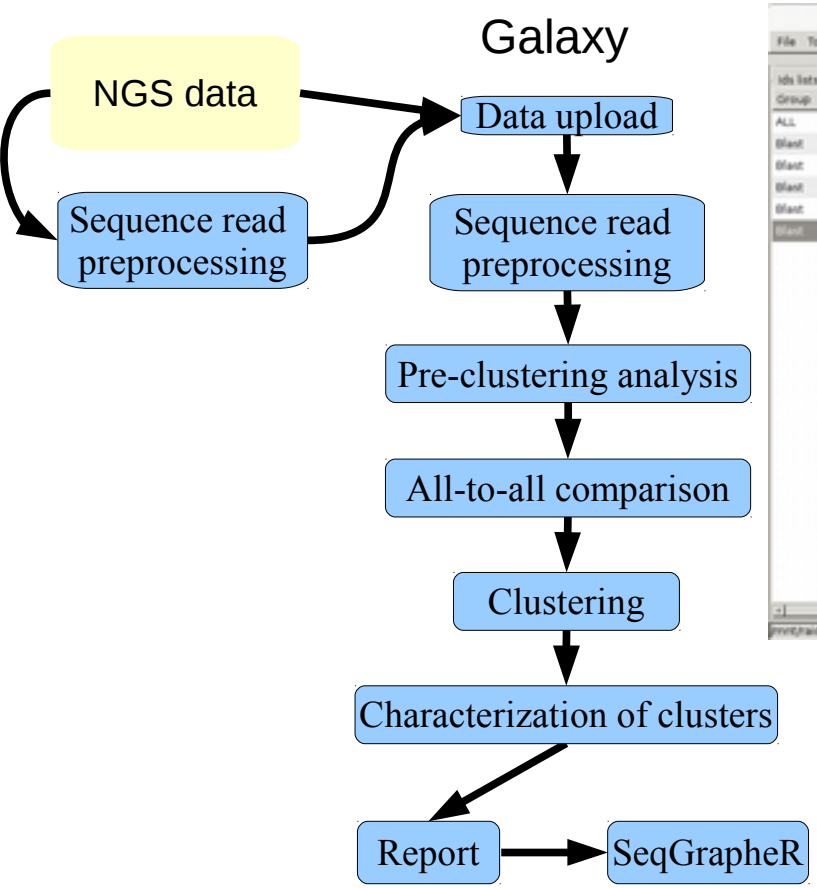




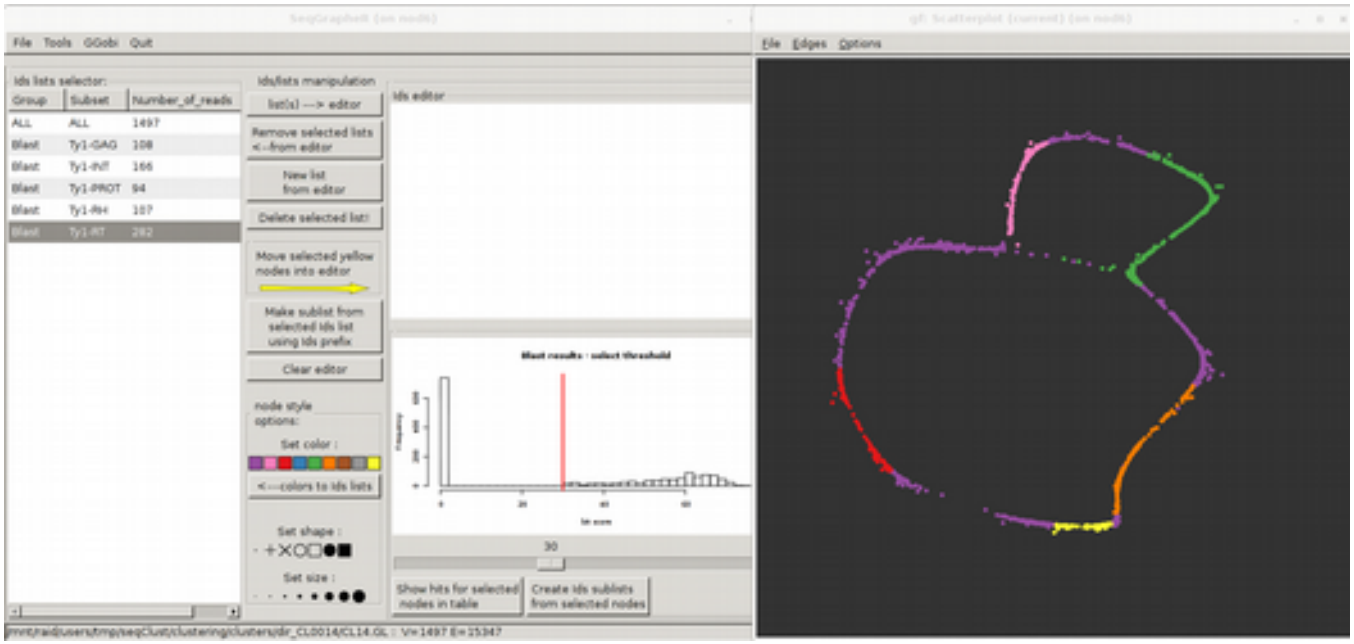
RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



SeqGrapheR

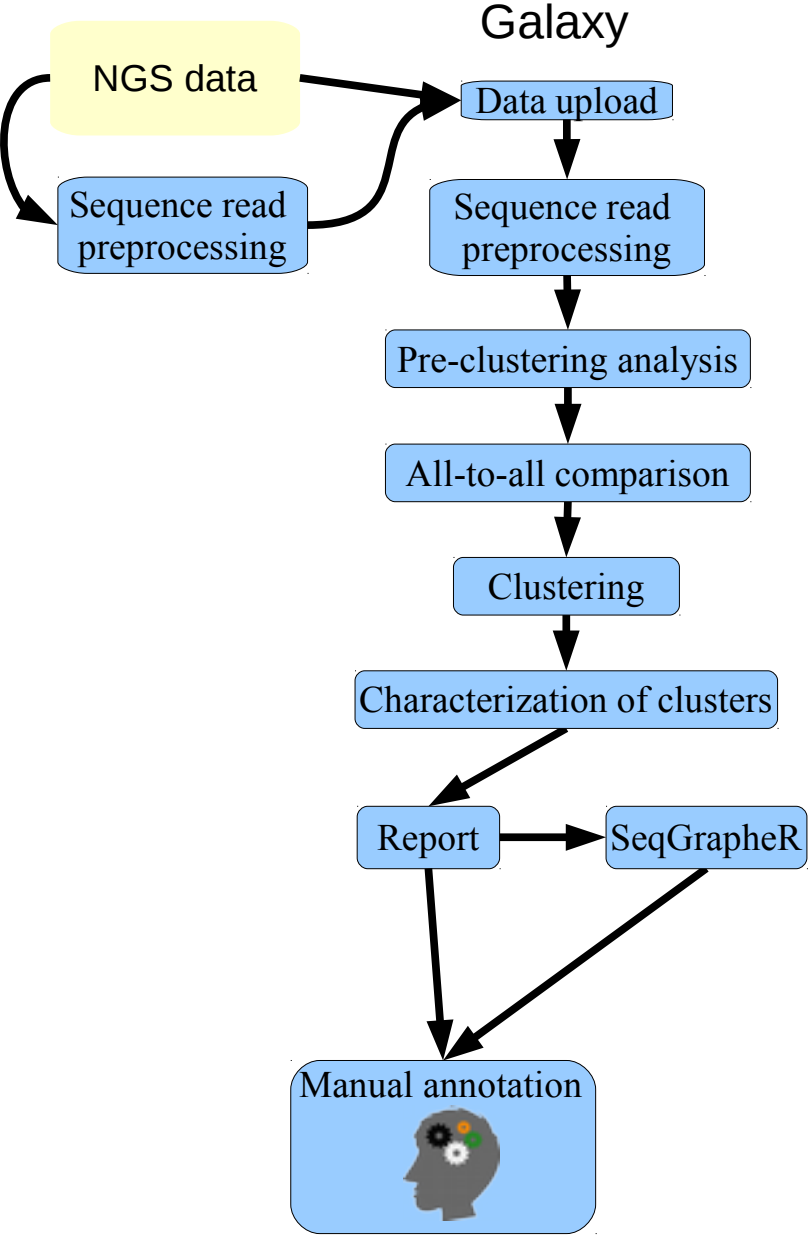


RepeatExplorer

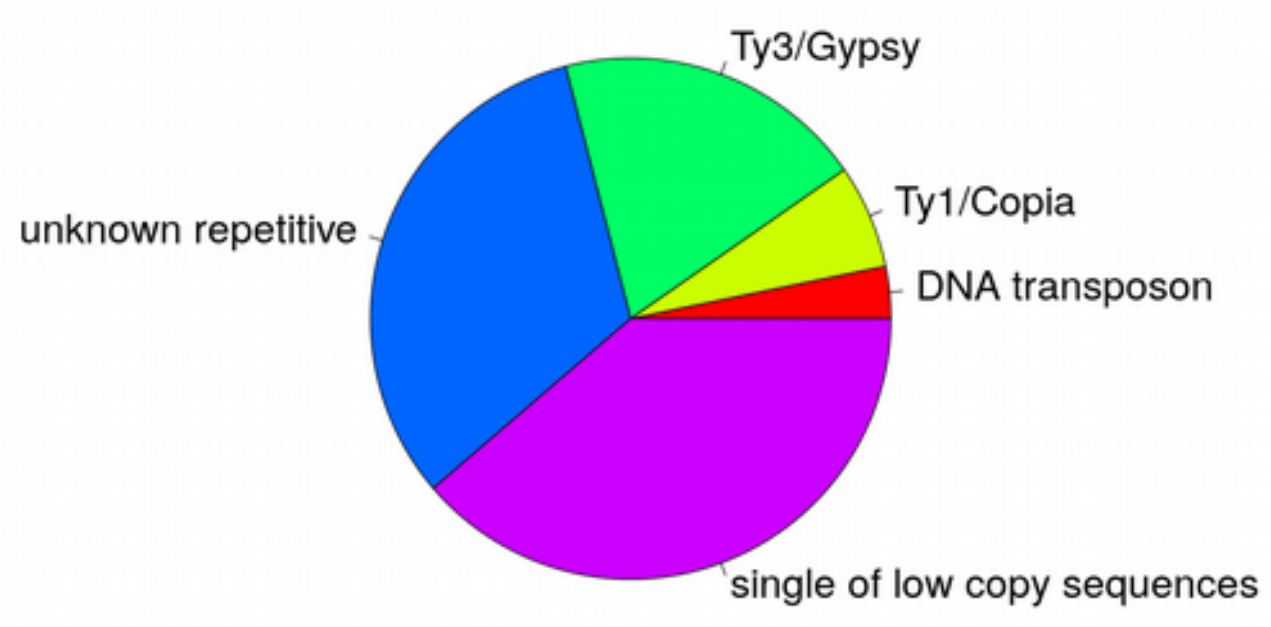
Discover repeats in your next generation sequencing data

Analysis work-flow

Galaxy



Manual annotation

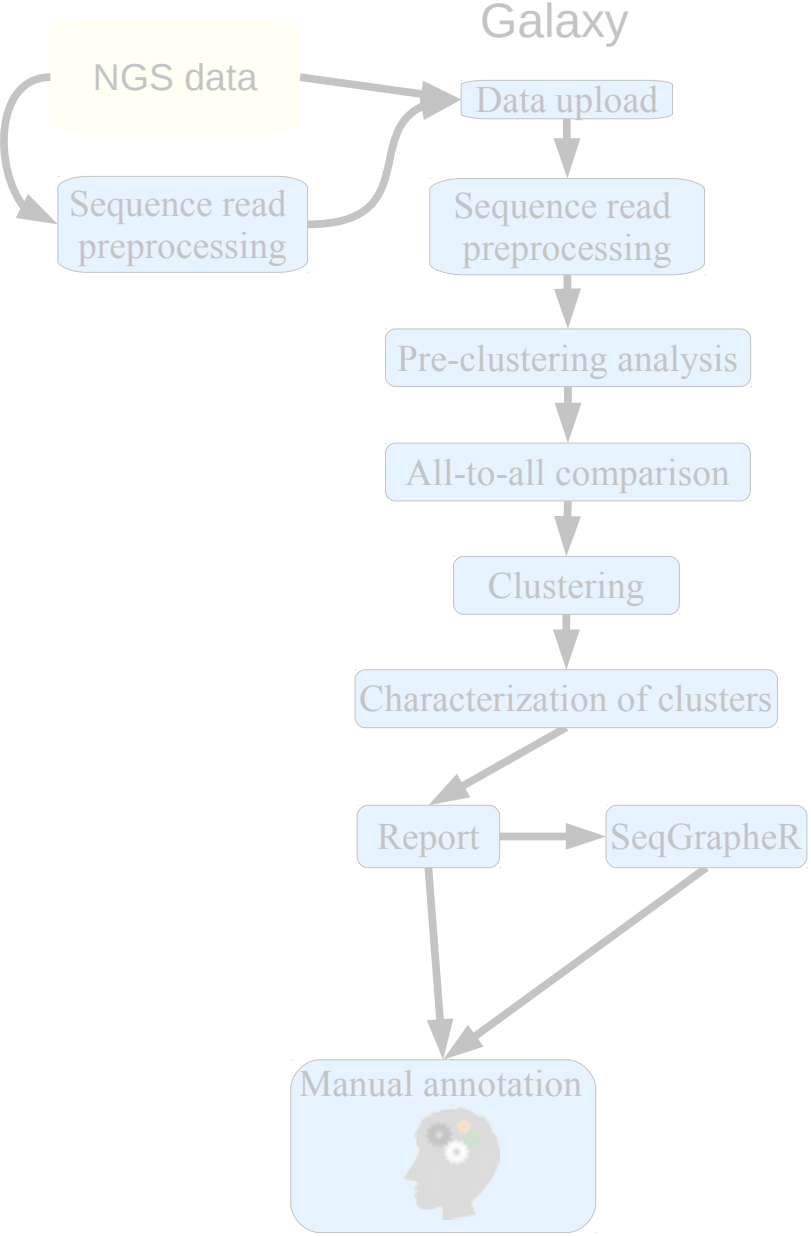


RepeatExplorer

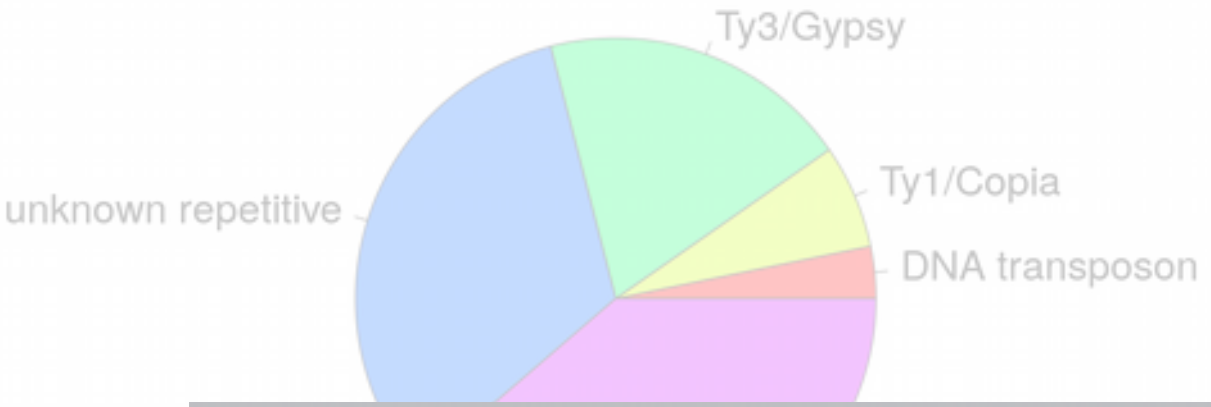
Discover repeats in your next generation sequencing data

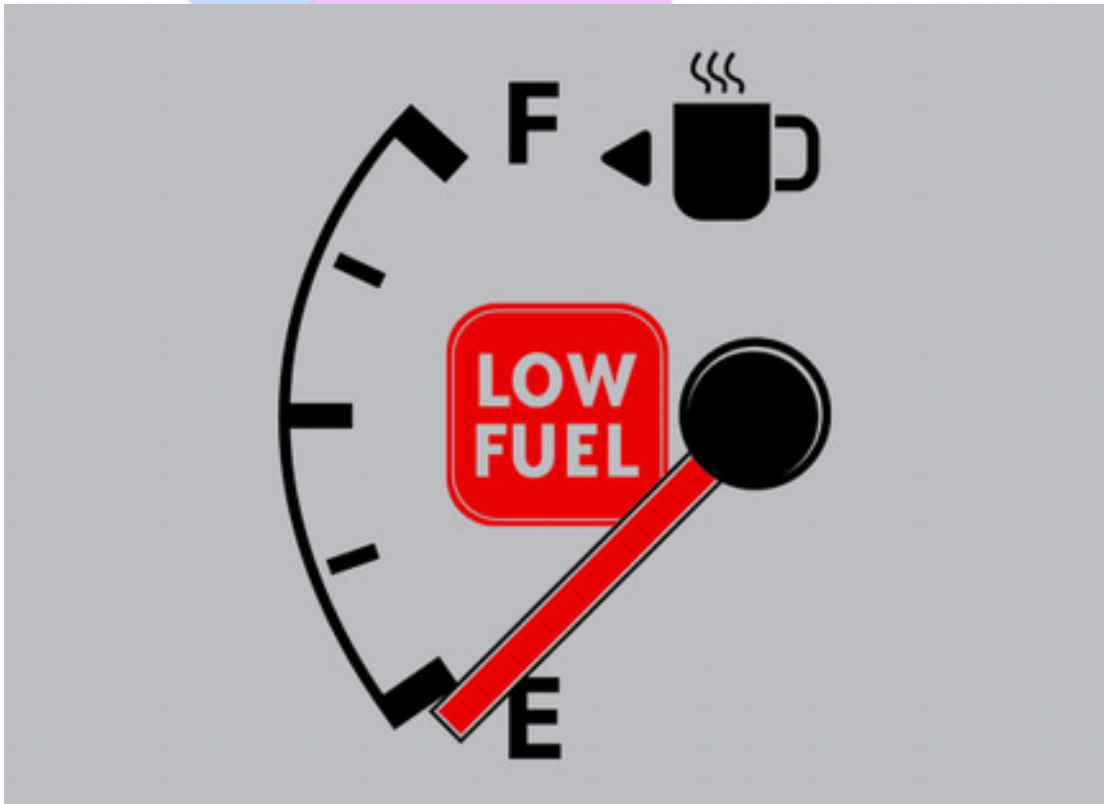
Analysis work-flow

Galaxy



Manual annotation





Advanced topics

Maximum number of reads which can be processed depends on genome compositions:

- Number of various repeats
- Copy number
- Length
- Diversity

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

Tandem repeat filtering

- In some genomes some abundant satellites produce majority similarity hits
- Most of the reads derived from tandem repeats are nearly identical
- We can remove most of the reads derived from tandem repeats without loosing information about other repeats
- Testing run with small sample (~ 300k) to identify problematic tandem repeats

Tandem repeat filtering

Advantages:

- More reads can be processed
- Shorter computational time

Filter = contigs assembled from reads derived from tandem repeat

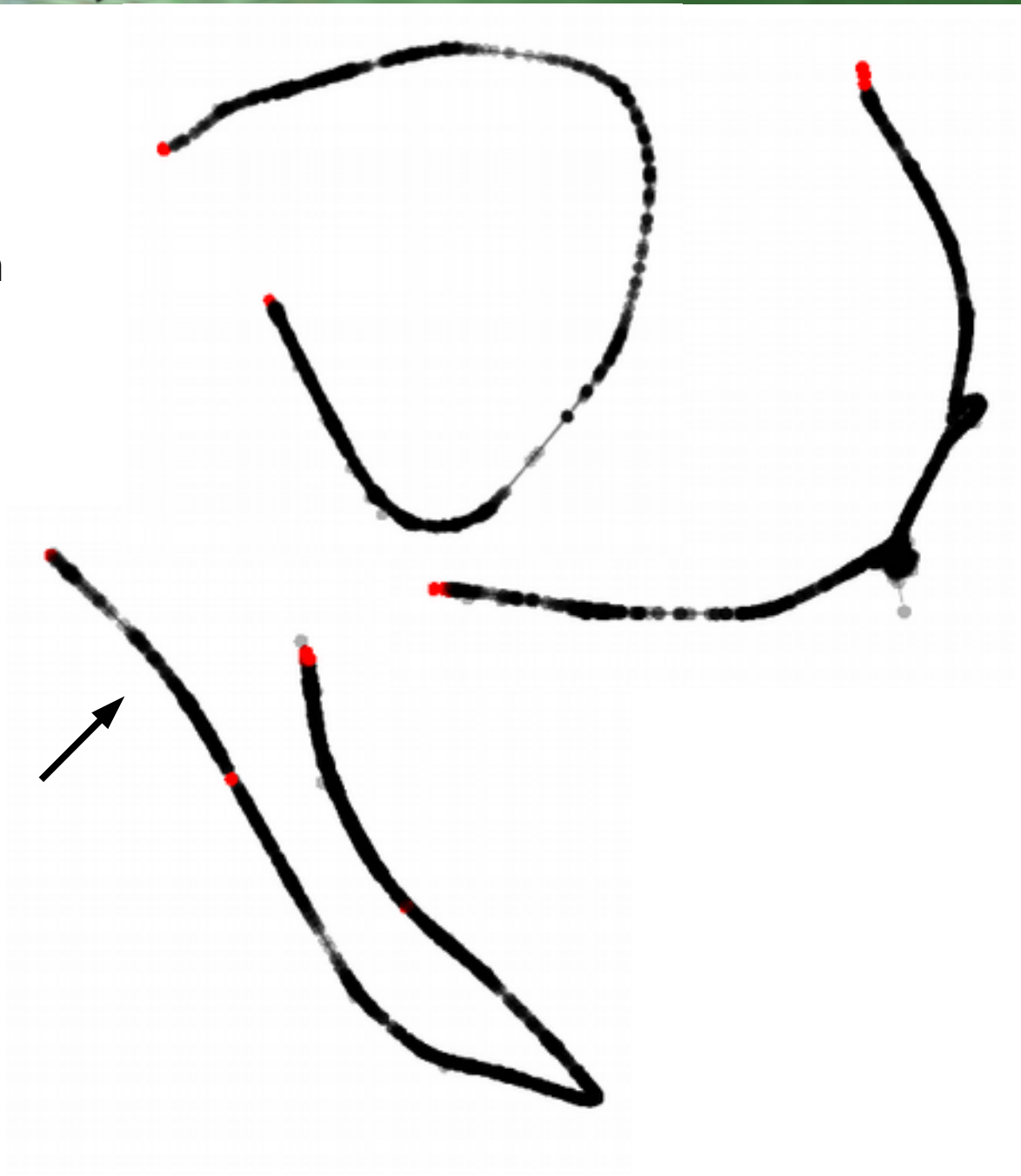
Species	filter	keep	Estimated number of reads	Reference repeat_quantification [%]				Computation time [hrs]
V. faba	-	-	3,076,204	Fok	3.403	TR9-like	1.21	4.17
	Fok, TR9-like	50%	4,322,187		3.4024		1.2	2.83
	Fok, TR9-like	10%	5,319,996		3.4327		1.13	2.42
V. sativa	-	-	1,047,006	VicTR.B	7.809			8.78
	VicTR-B	50%	1,993,513		7.8023			4.23
	VicTR-B	10%	4,751,557		7.793			2.35

Cluster fragmentation

Modularity corresponds to the fraction of edges that fall within communities minus the corresponding value in the random graph

Disadvantage of modularity:
Long elements splitting

Depending on initial condition, Graph can be splitted in multiple ways

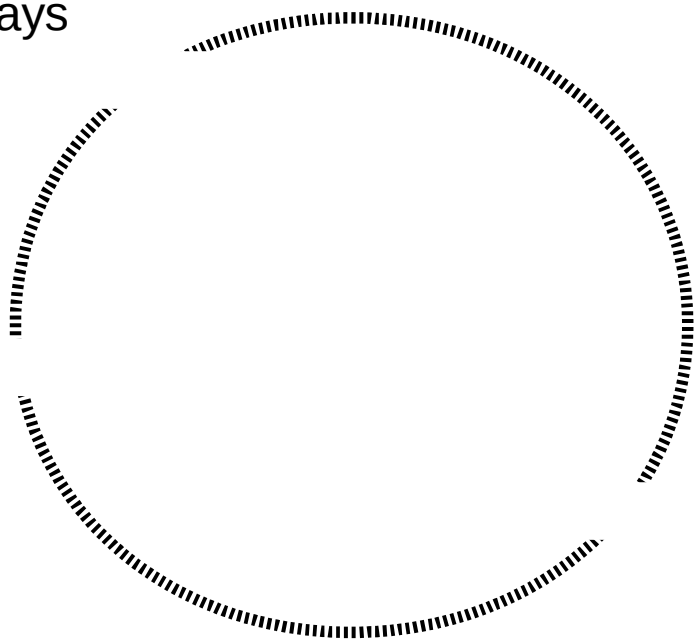


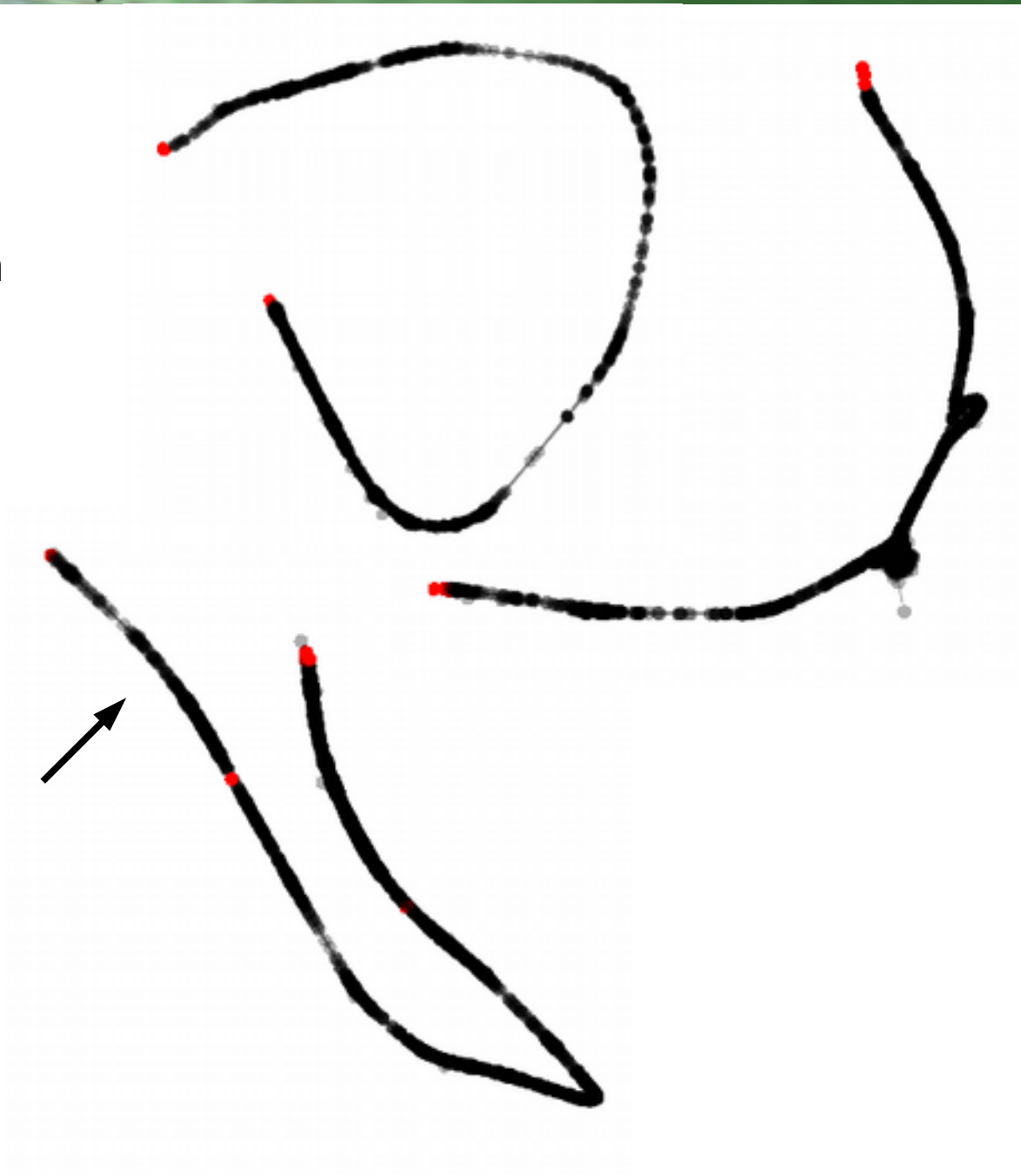
Cluster fragmentation

Modularity corresponds to the fraction of edges that fall within communities minus the corresponding value in the random graph

Disadvantage of modularity:
Long elements splitting

Depending on initial condition, Graph can be splitted in multiple ways



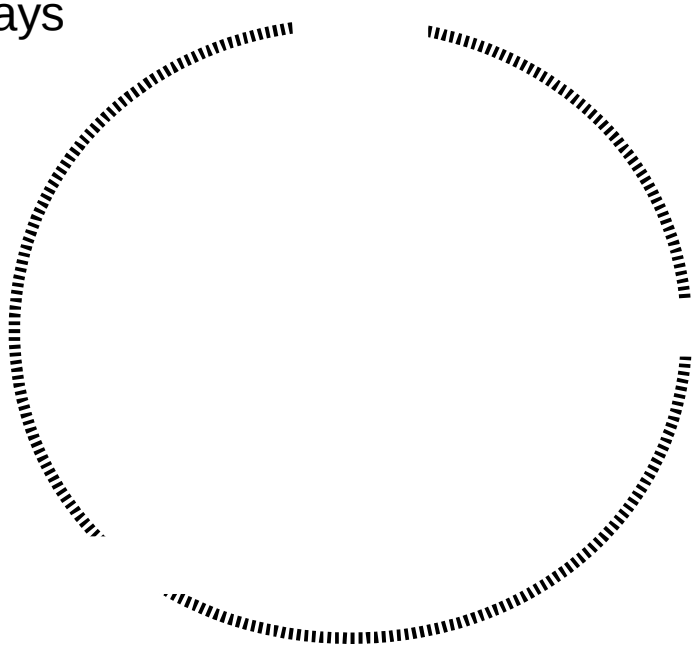


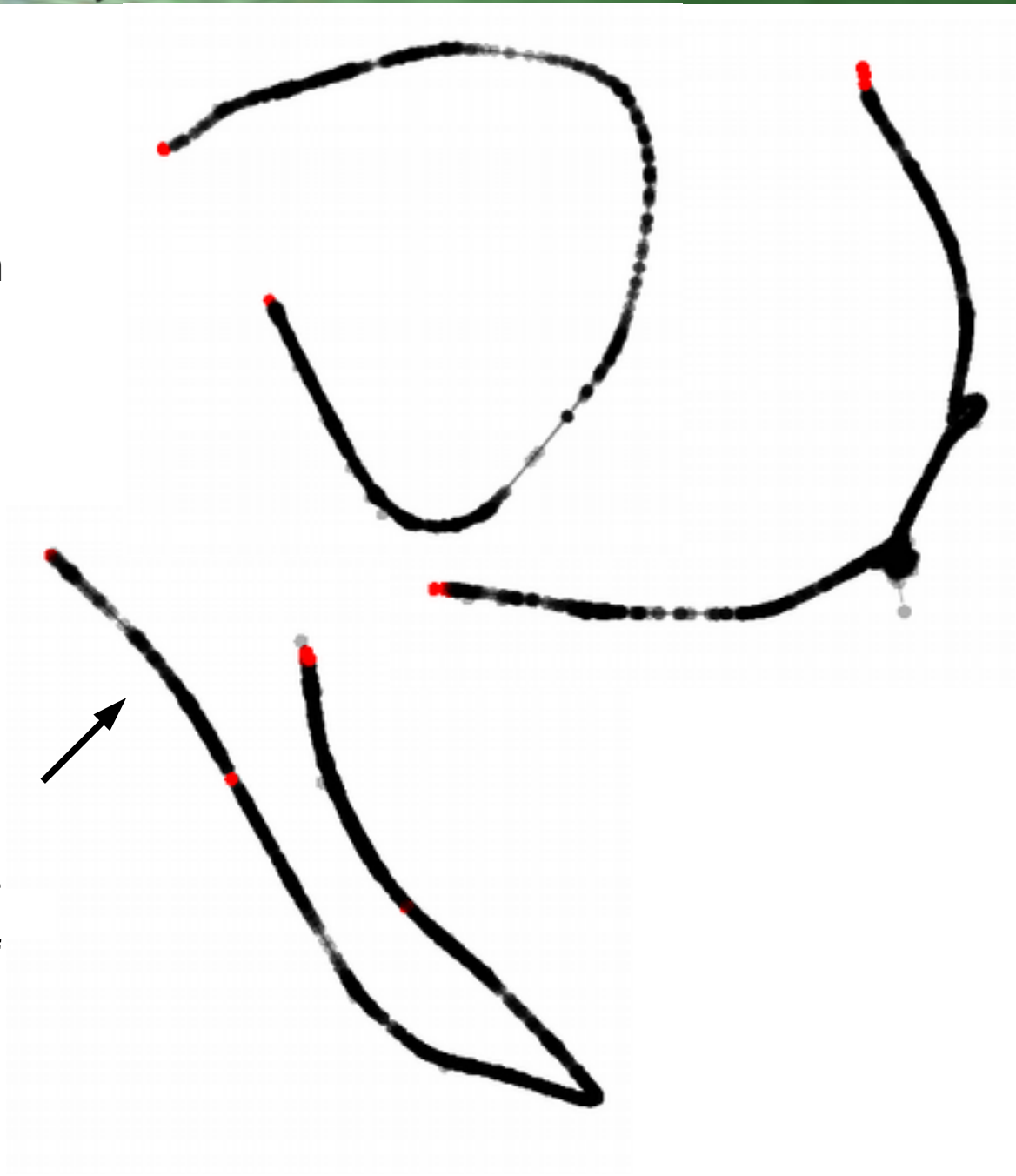
Cluster fragmentation

Modularity corresponds to the fraction of edges that fall within communities minus the corresponding value in the random graph

Disadvantage of modularity:
Long elements splitting

Depending on initial condition, Graph can be splitted in multiple ways



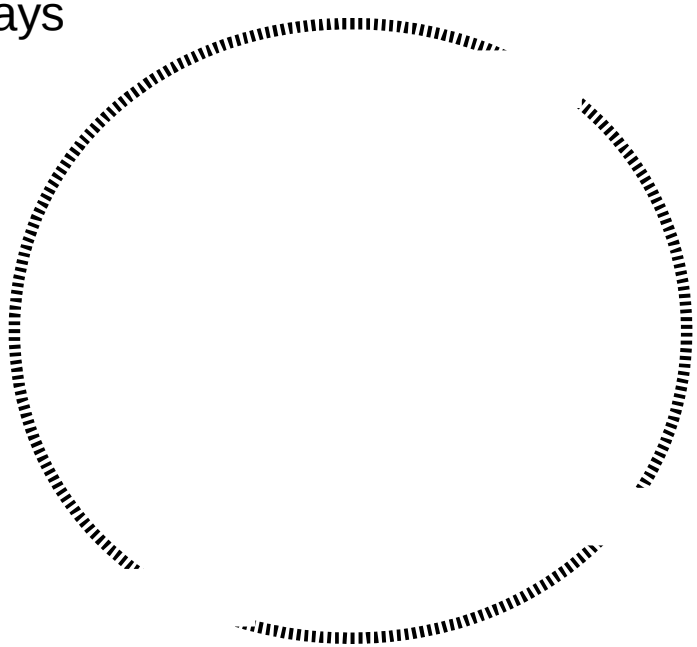


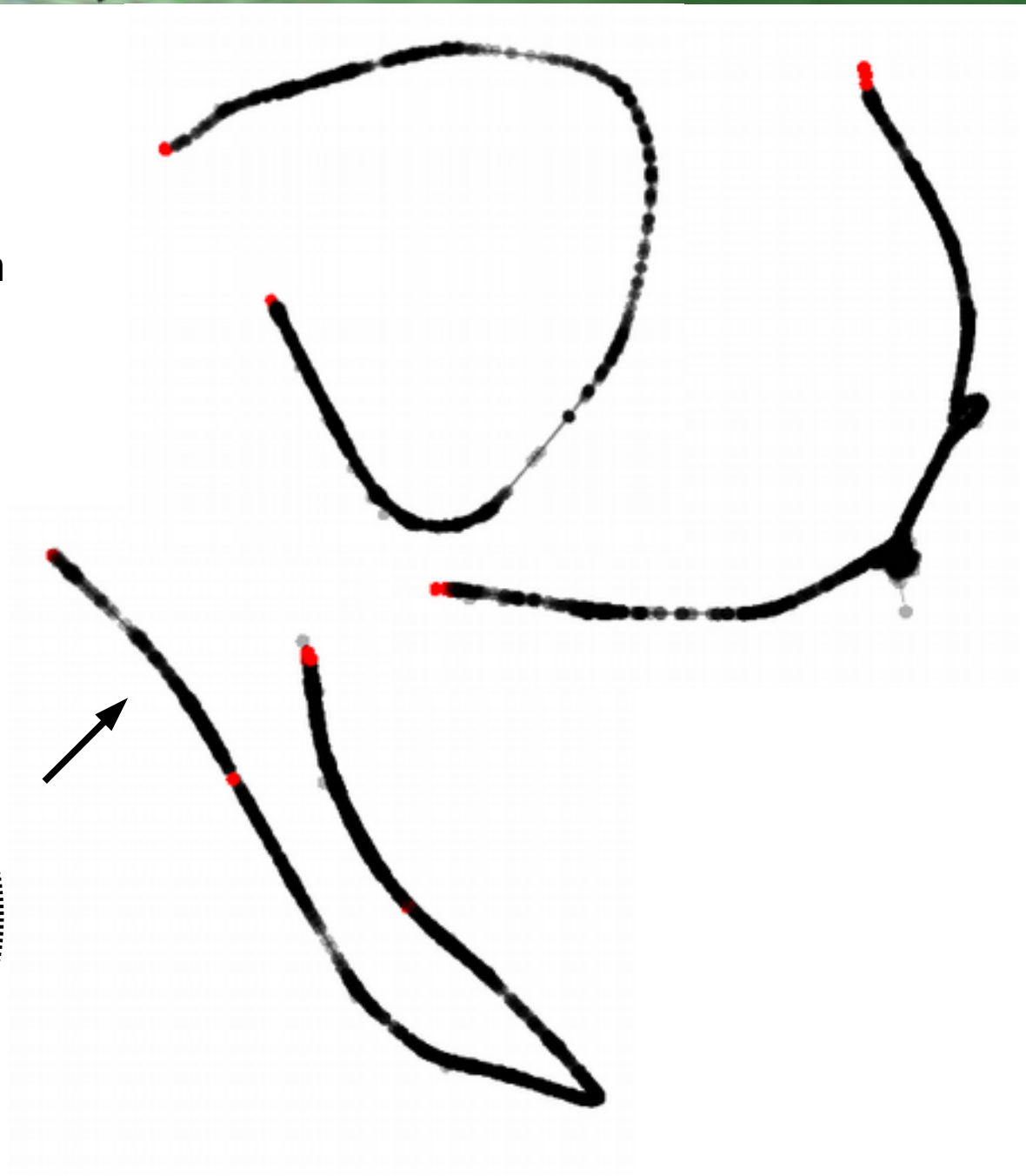
Cluster fragmentation

Modularity corresponds to the fraction of edges that fall within communities minus the corresponding value in the random graph

Disadvantage of modularity:
Long elements splitting

Depending on initial condition, Graph can be splitted in multiple ways

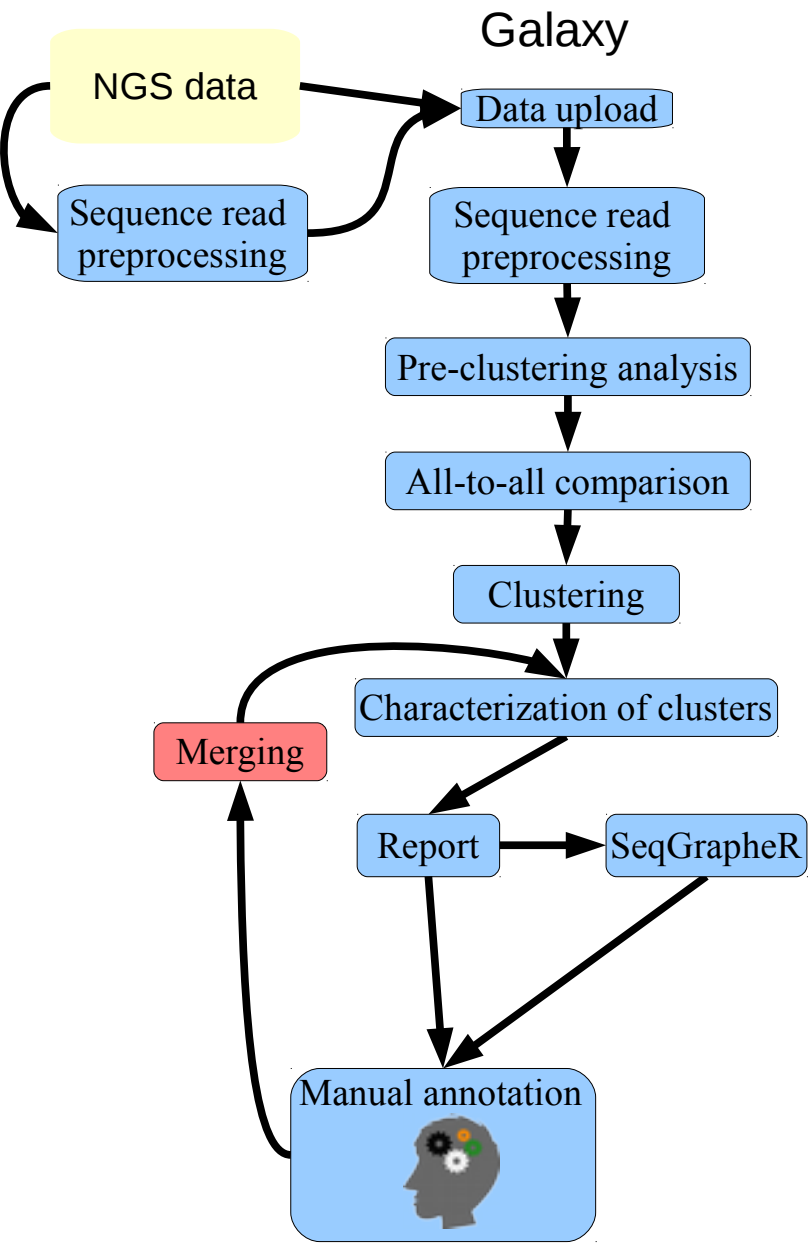




RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Cluster merger

How to identify clusters suitable for merging:

- **Manual exploration of clusters** (annotation, similarities between clusters, pair reads)

- Criterion for merging based on the **paired reads** presence.:

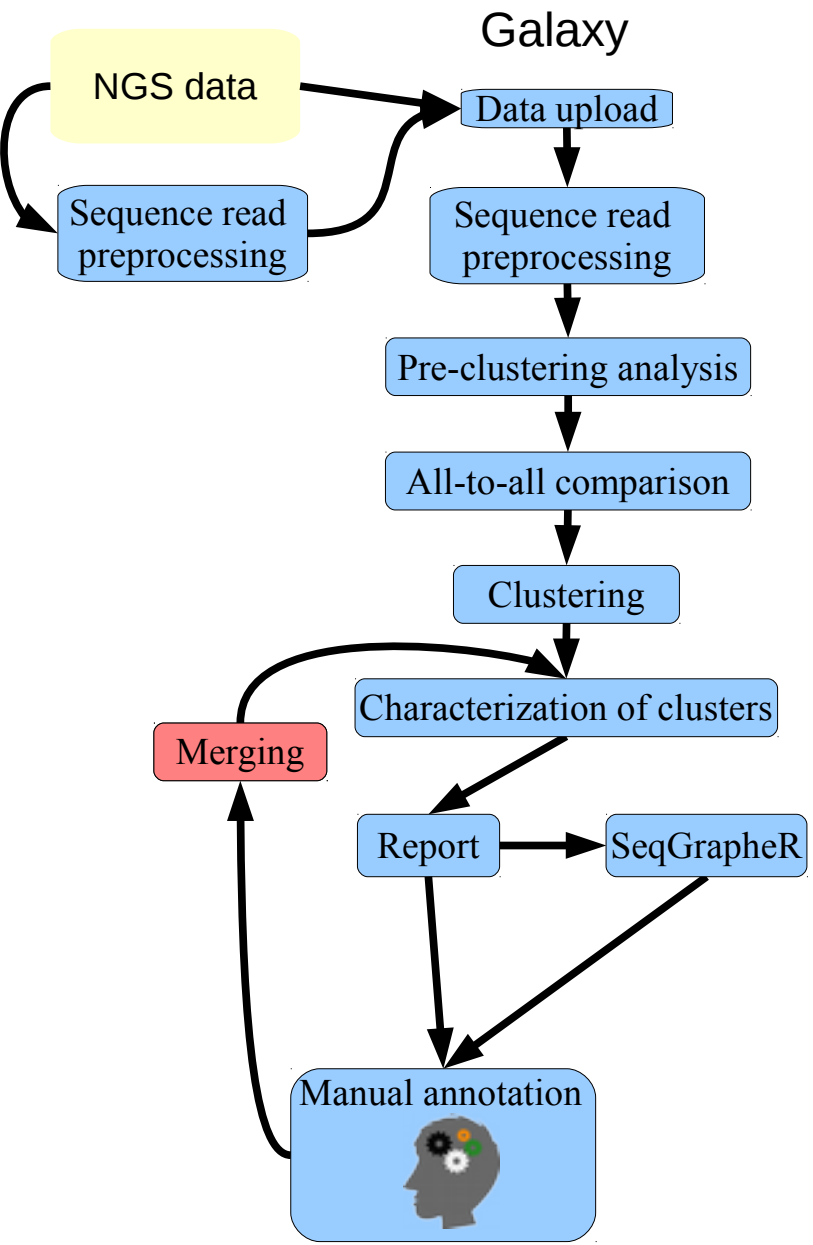
$$k_{x,y} = \frac{2W}{n_x + n_y}$$

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

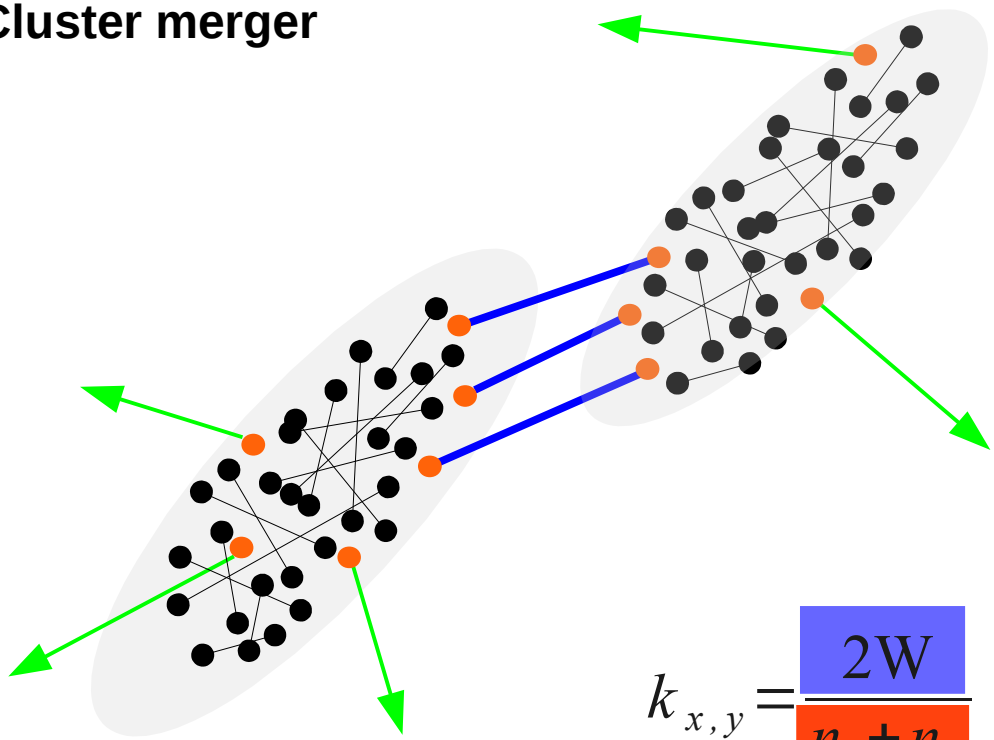
RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Cluster merger



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

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

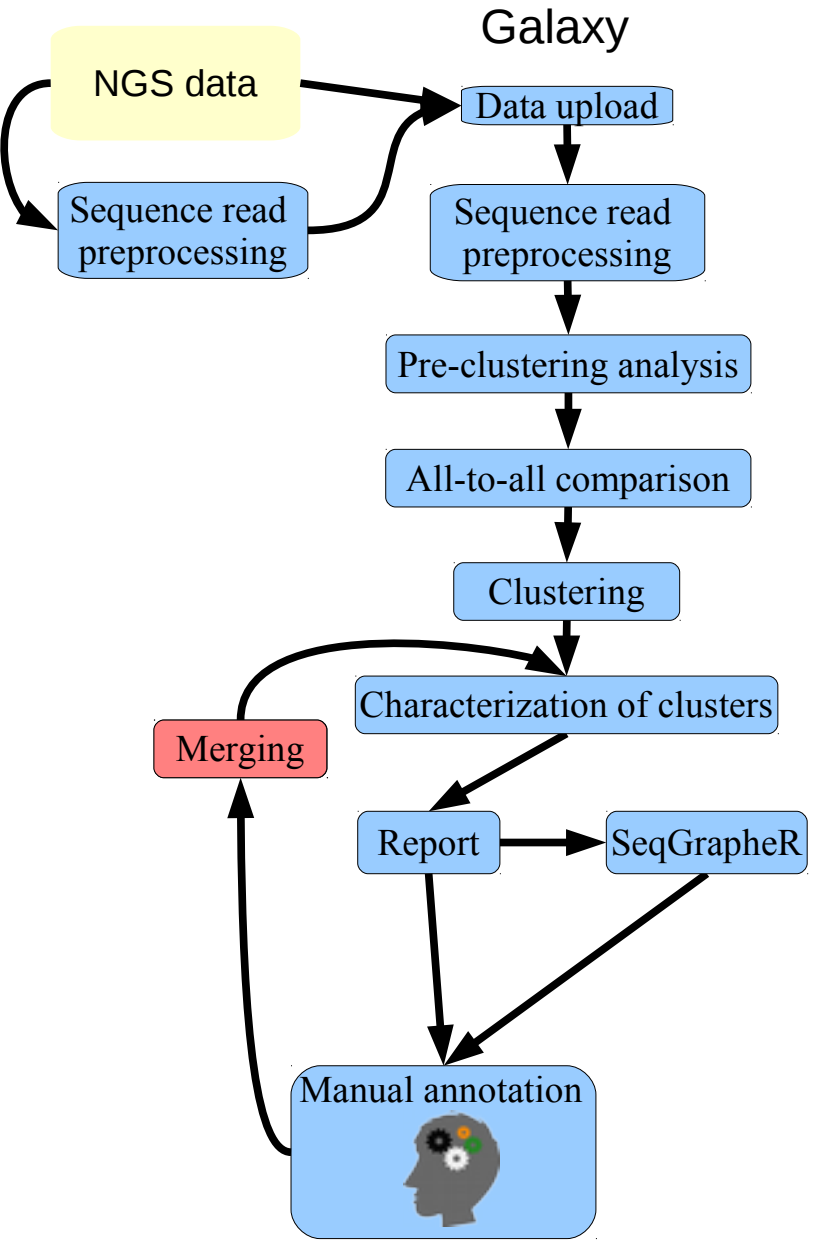
full connection: $k_{x,y} = 1$

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

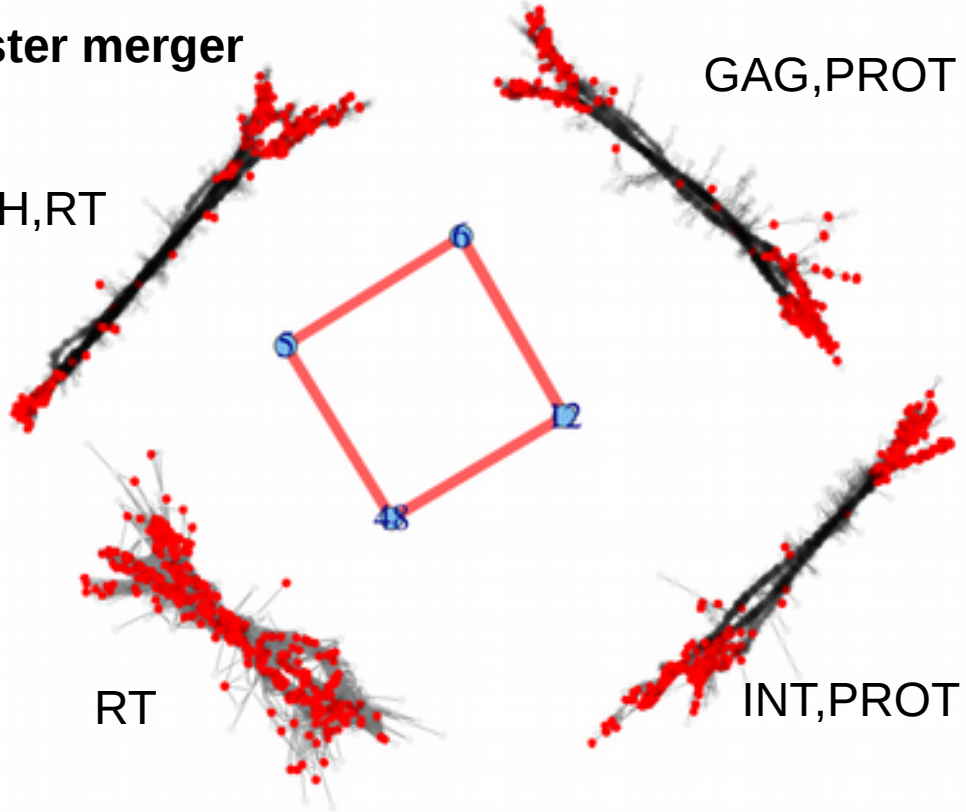
RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Cluster merger

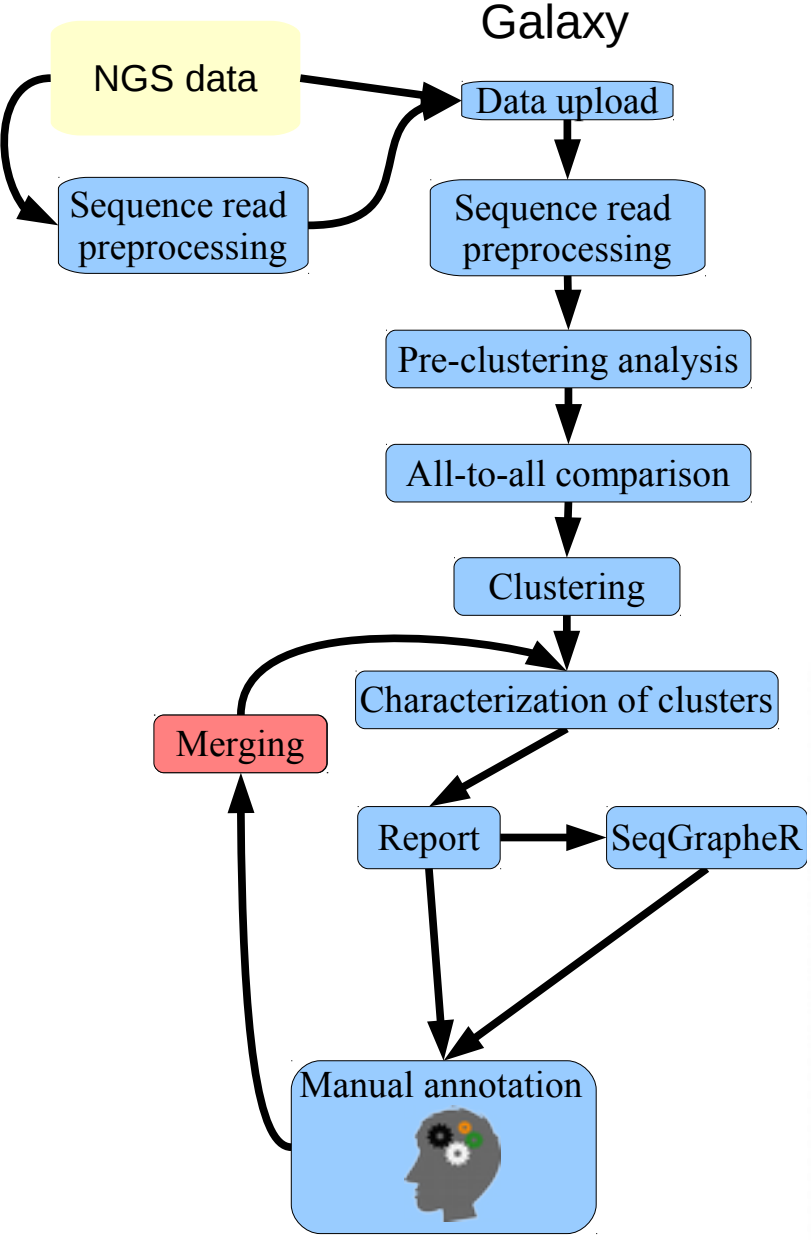


RepeatExplorer

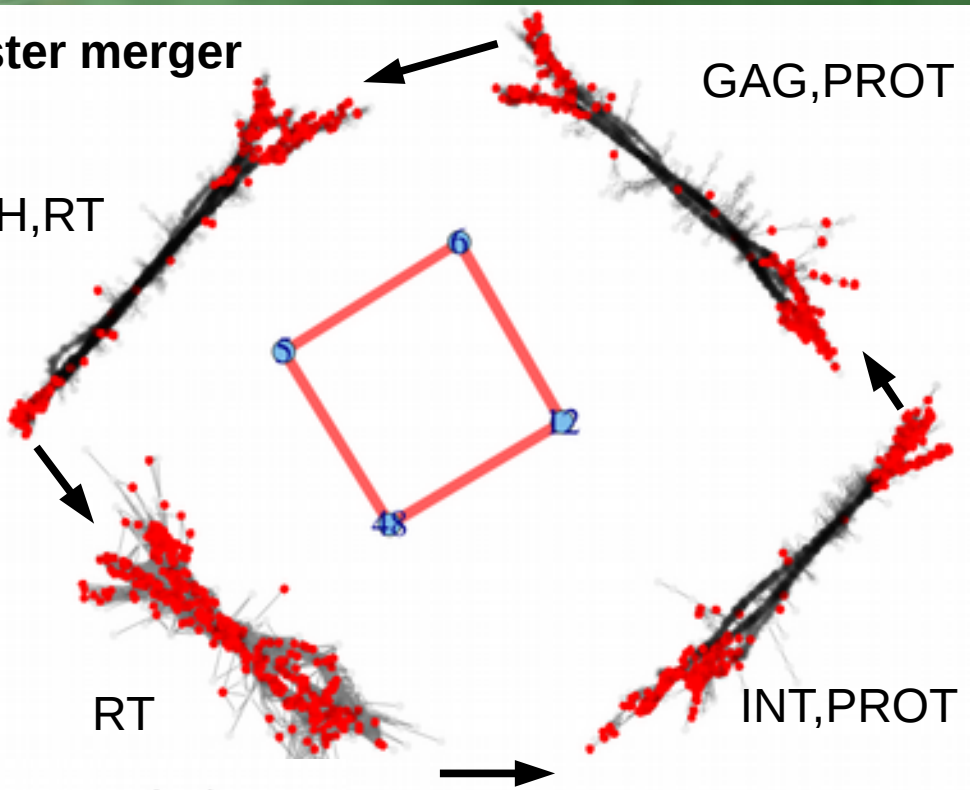
Discover repeats in your next generation sequencing data

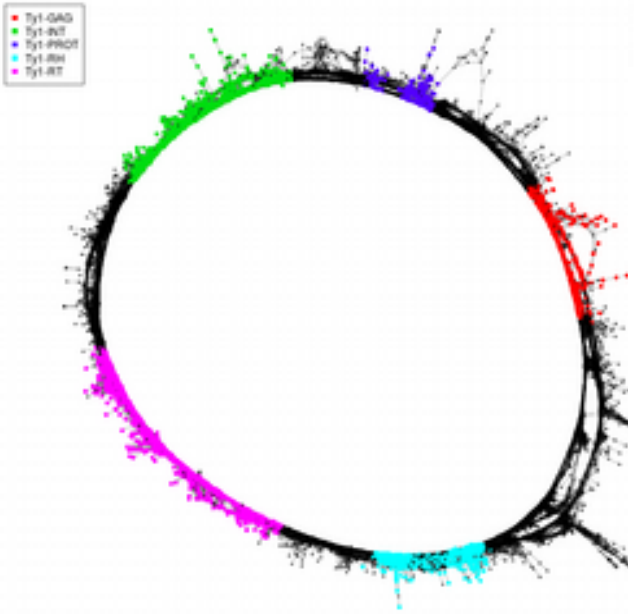
Analysis work-flow

Galaxy



Cluster merger





RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

Galaxy

NGS data

Data upload

Sequence read preprocessing

Sequence read preprocessing

Pre-clustering analysis

All-to-all comparison

Clustering

Characterization of clusters

Merging

Report

SeqGrapheR

Manual annotation

Cluster merger

GAG,PROT

RH,RT

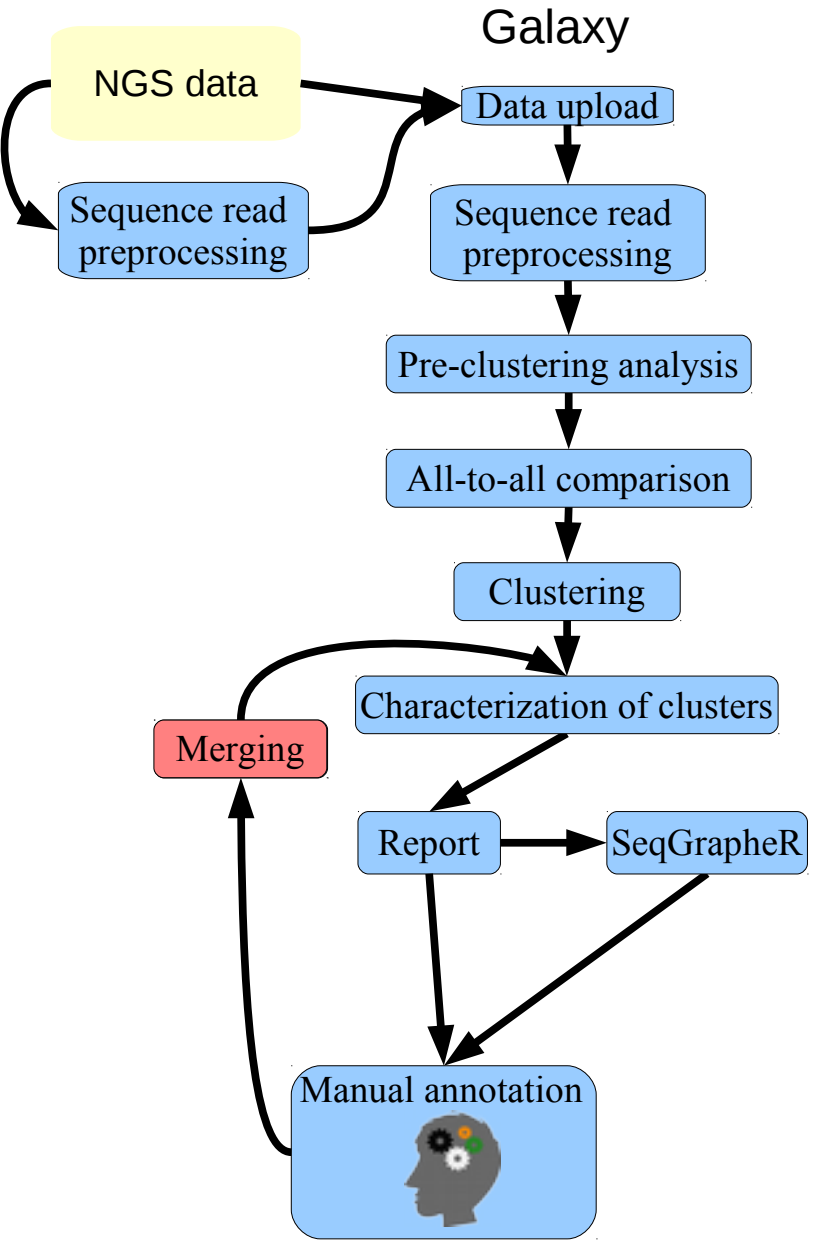
INT,PROT

Clusters without protein domain

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Cluster merger

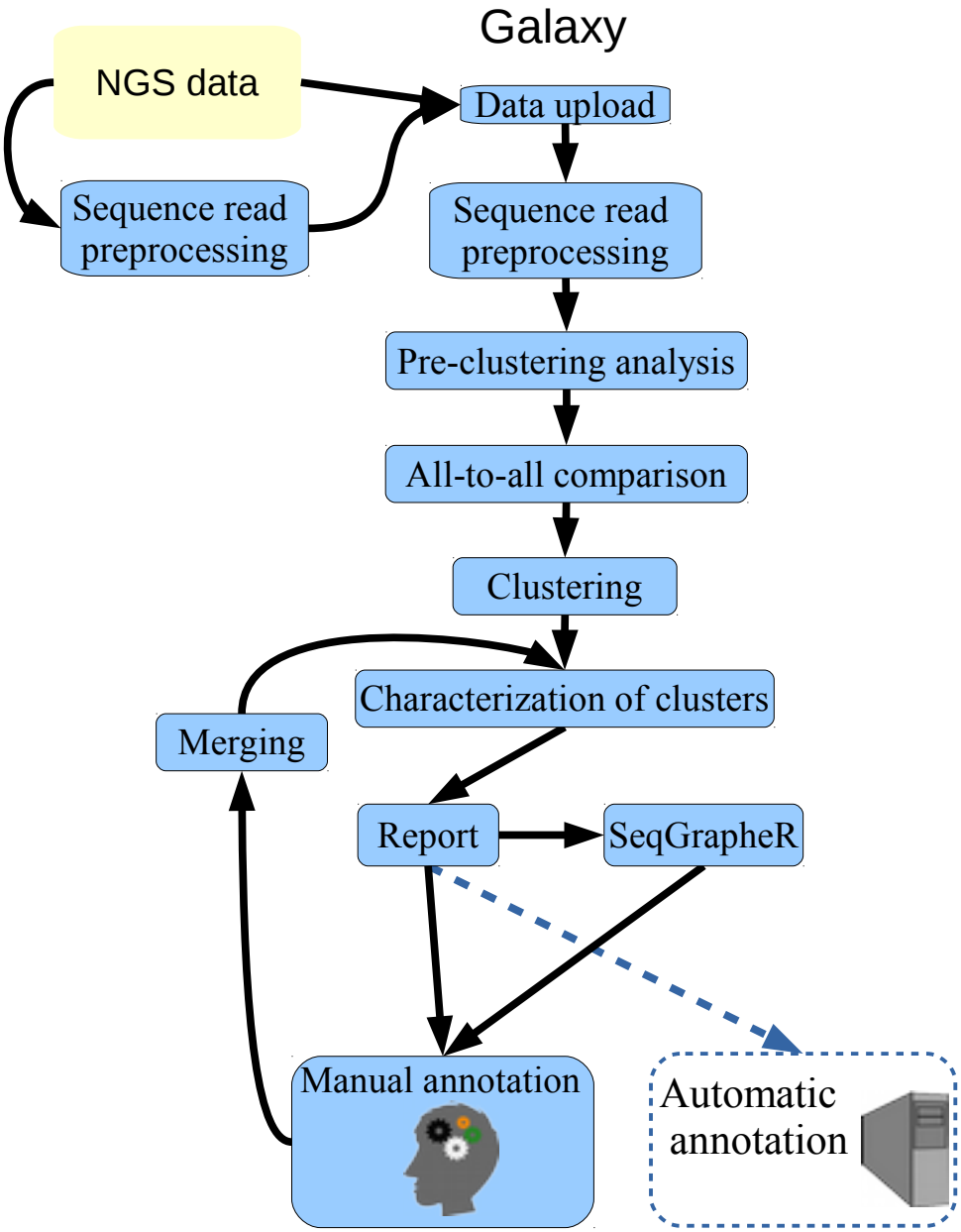
Merges definition:

```
9 12 15 16 17 32 33 39 40 44 48 50
19 27 28 34 41 43 45 47 53
4 10 20 36 49 51 54
14 25 26 37 38
5 7 13 31
18 22 23 29
2 42
6 21
11 55
35 52
```

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

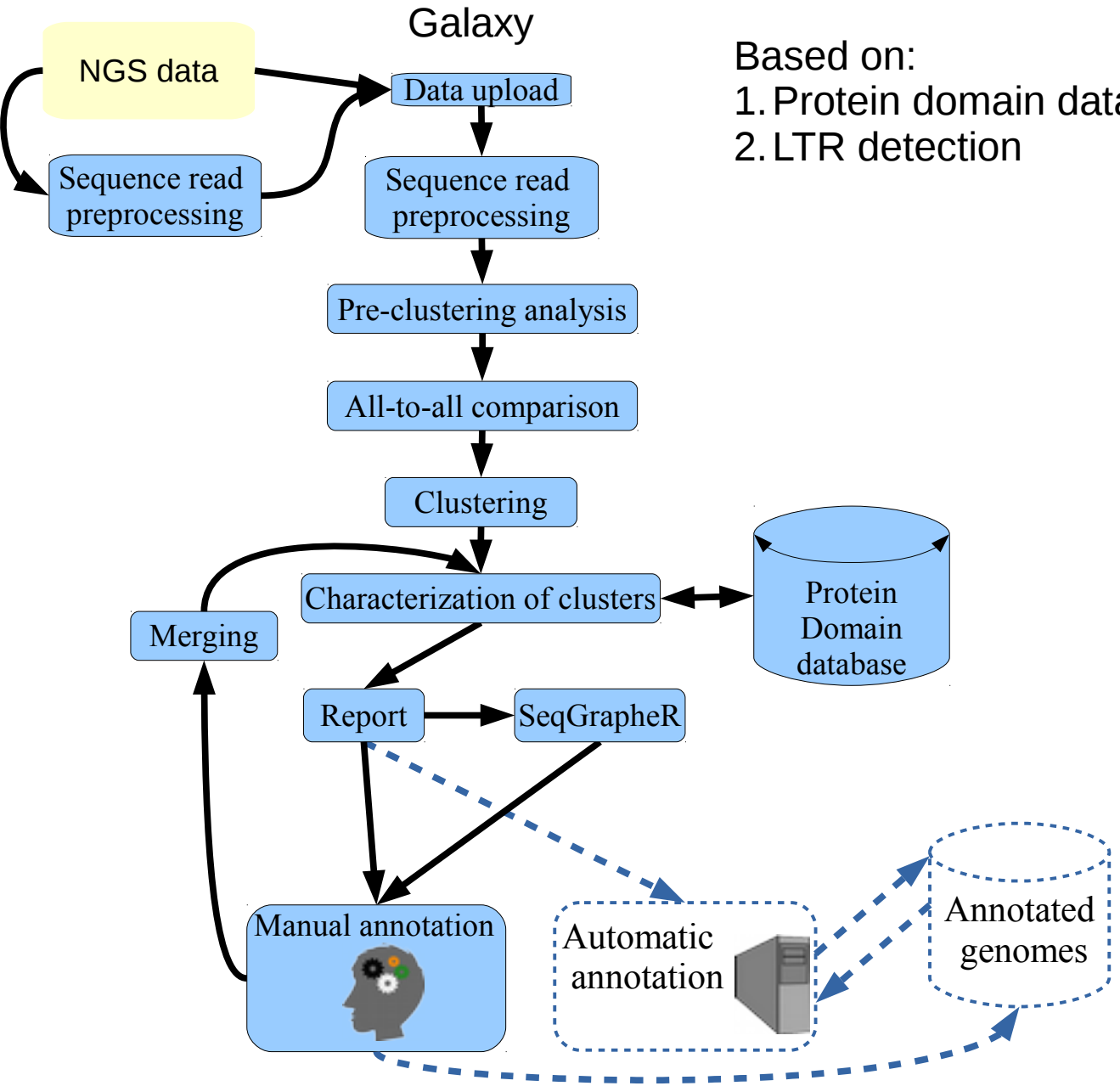


Automatic annotation

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Automatic annotation

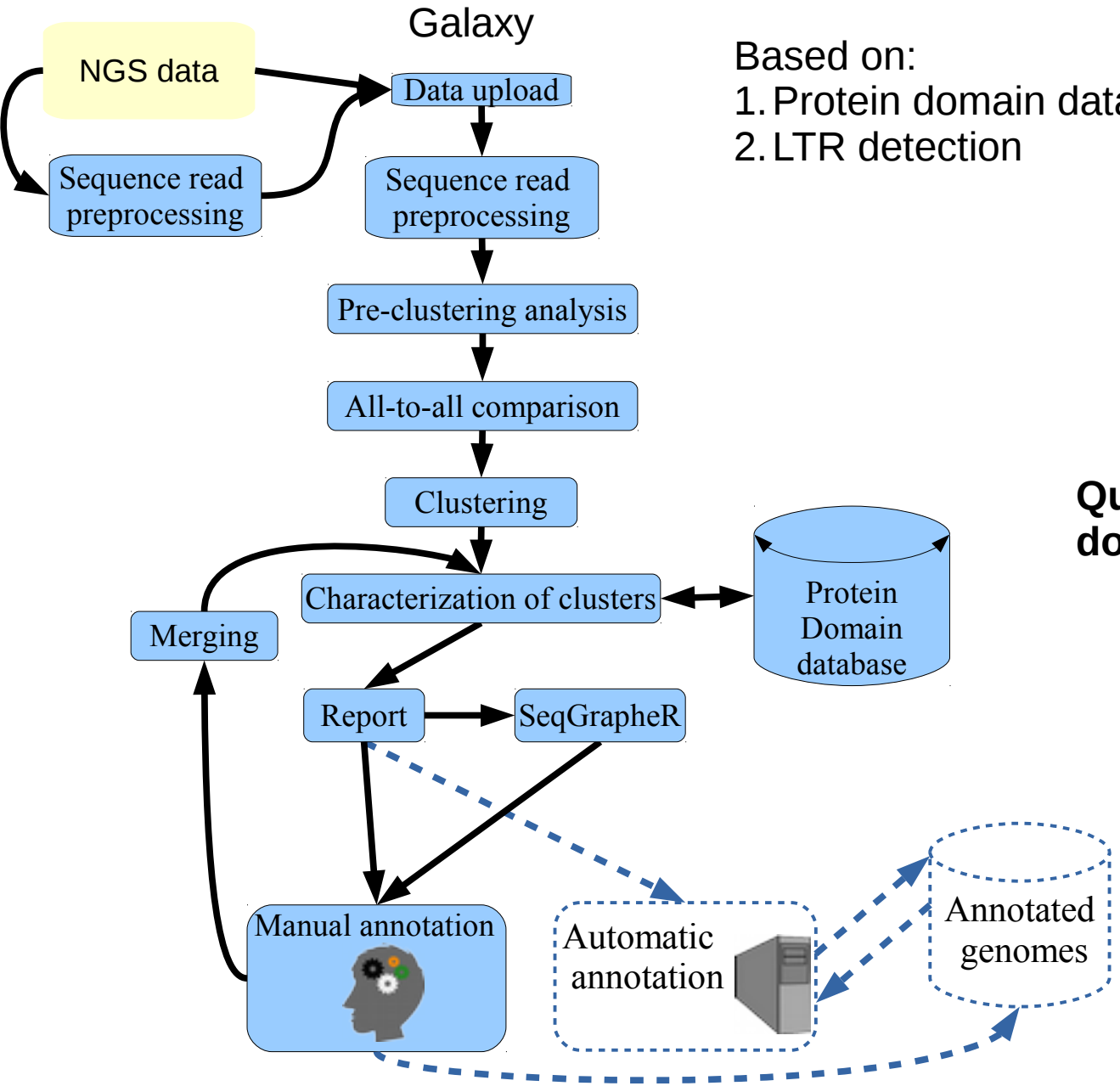
Based on:

1. Protein domain database
2. LTR detection

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Automatic annotation

Based on:

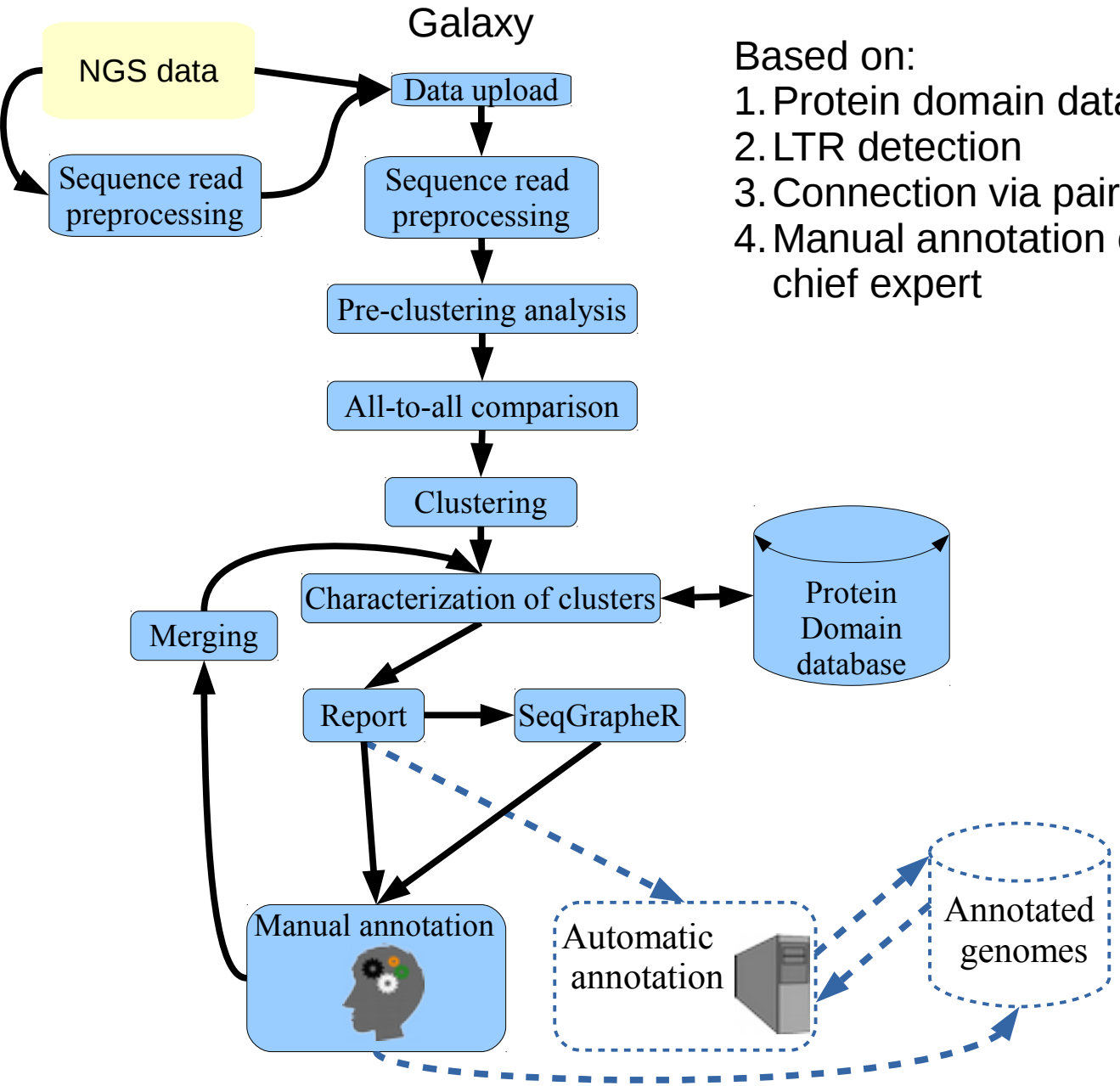
1. Protein domain database
2. LTR detection

Quality and size of protein domain database is essential.

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Automatic annotation

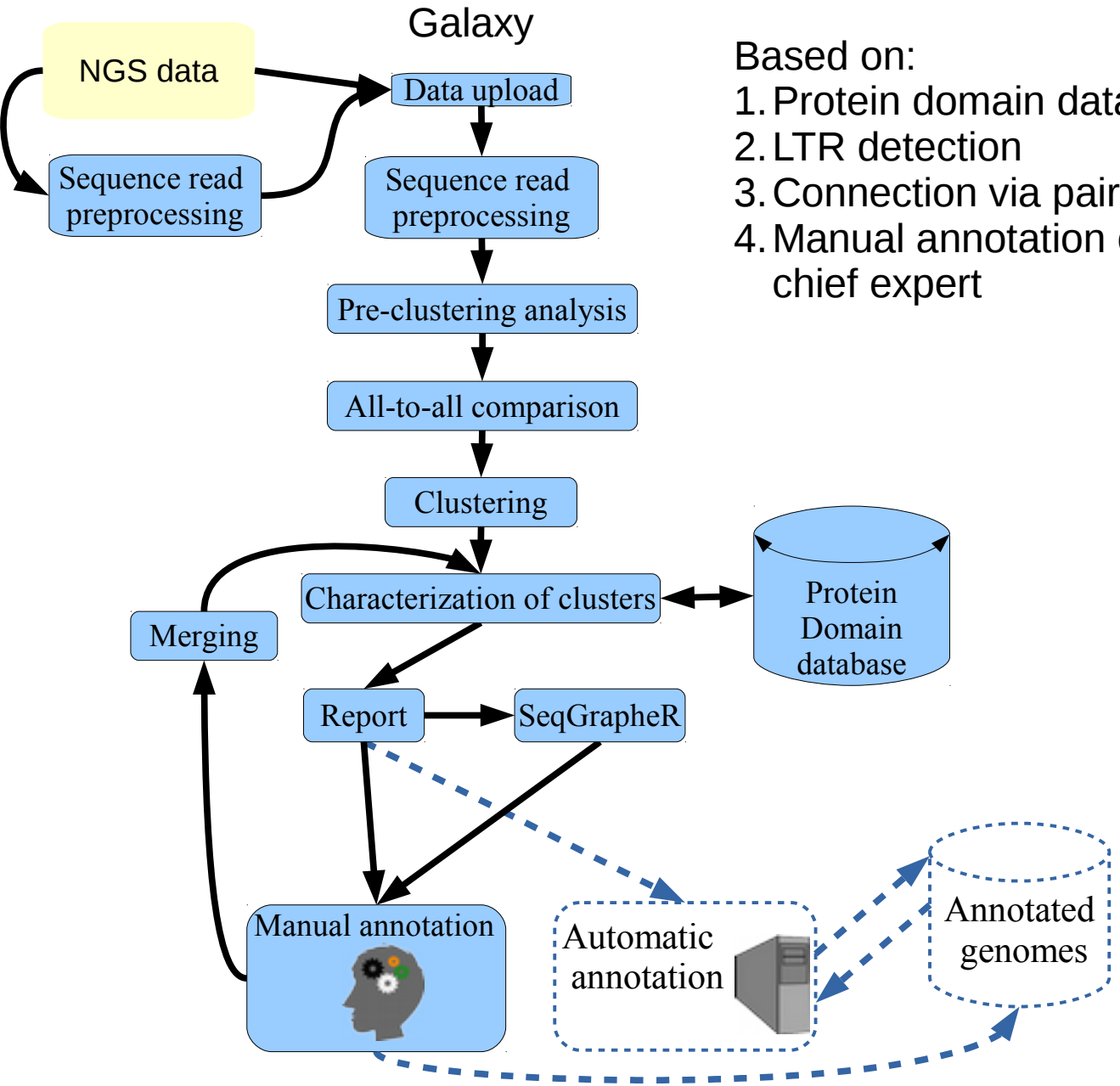
Based on:

1. Protein domain database
2. LTR detection
3. Connection via paired reads
4. Manual annotation of more than 1,600 clusters by chief expert

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Automatic annotation

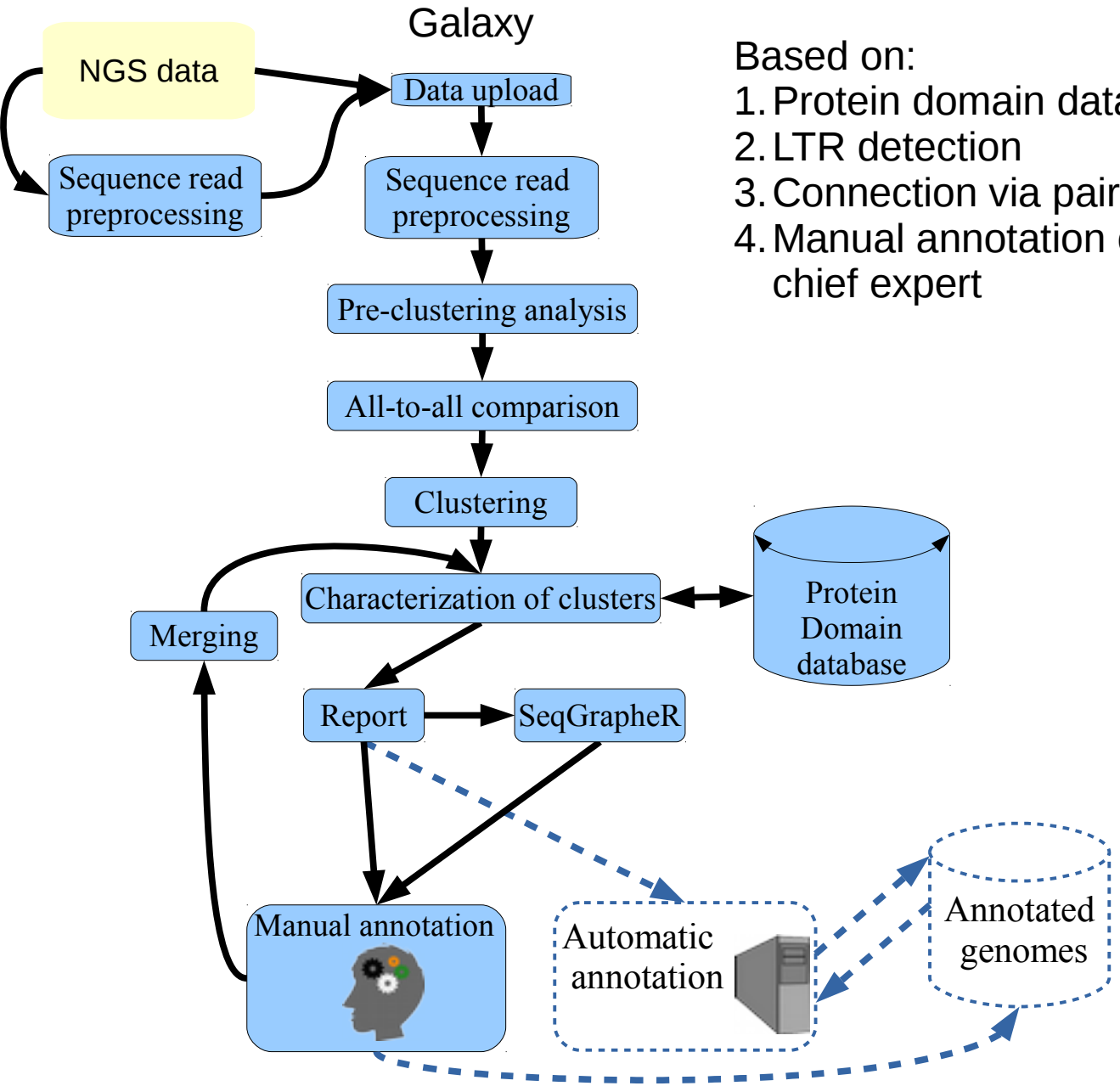
Based on:

1. Protein domain database
2. LTR detection
3. Connection via paired reads
4. Manual annotation of more than 1,600 clusters by chief expert

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Automatic annotation

Based on:

1. Protein domain database
2. LTR detection
3. Connection via paired reads
4. Manual annotation of more than 1,600 clusters by chief expert

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow

Galaxy

NGS data

Data upload

Sequence read preprocessing

Sequence read preprocessing

Pre-clustering analysis

All-to-all comparison

Clustering

Characterization of clusters

Protein Domain database

Merging

Report

SeqGrapheR

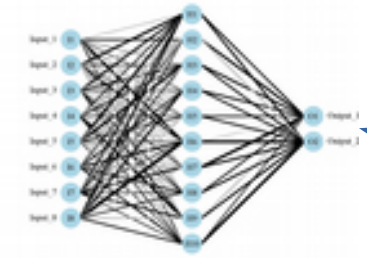
Manual annotation

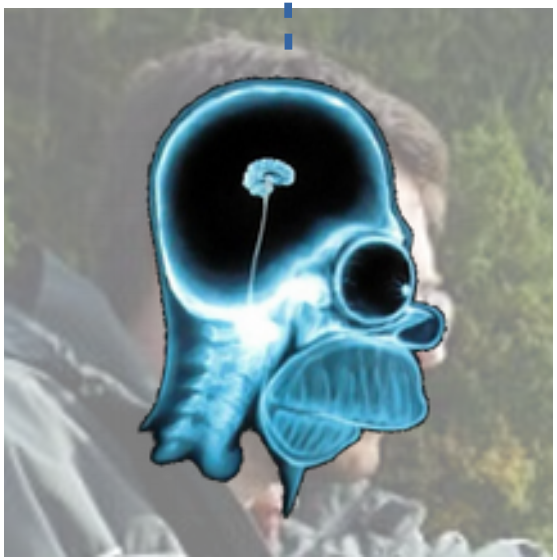
Automatic annotation

Annotated genomes

Automatic annotation

artificial neural network

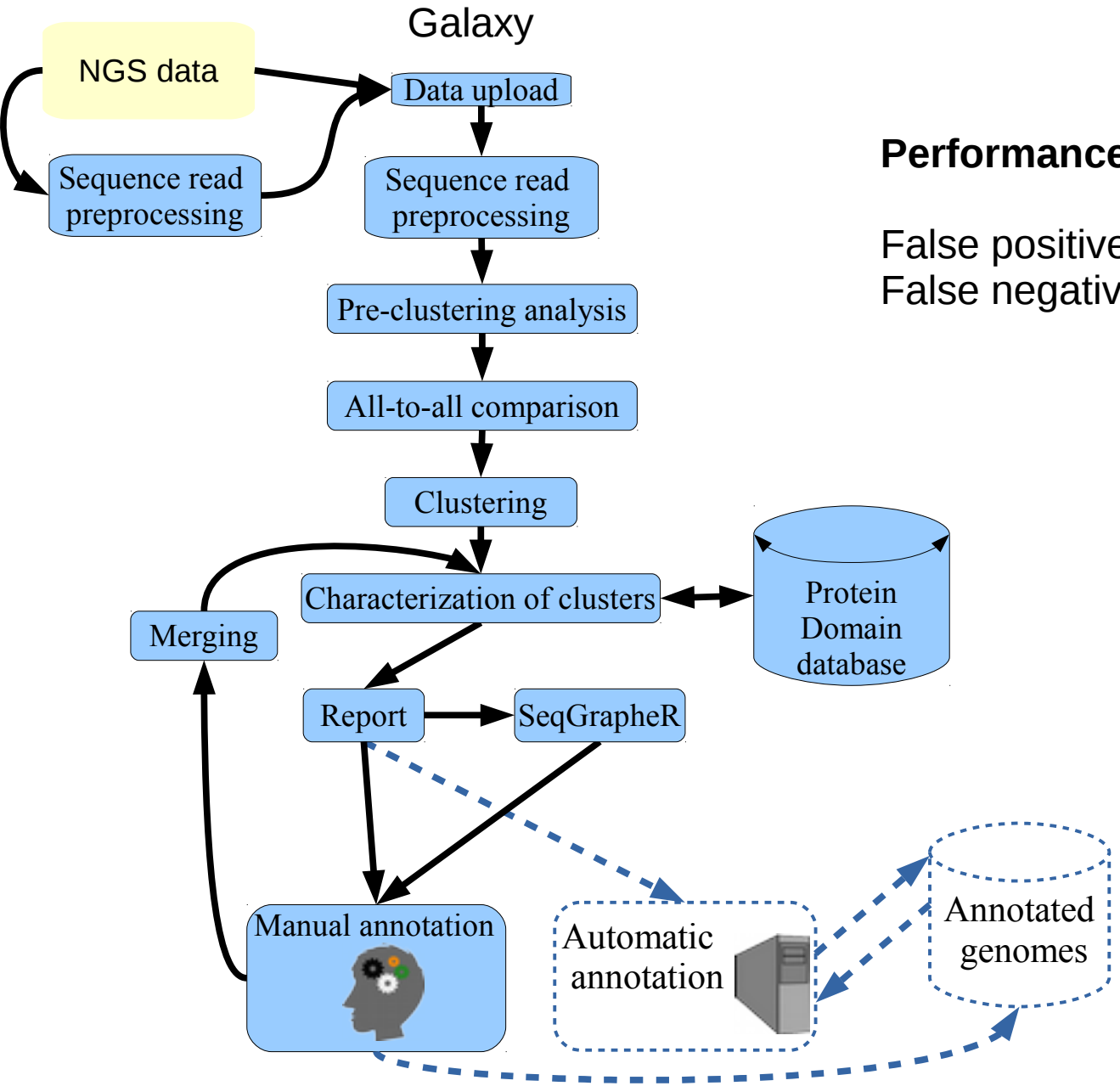




RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Automatic annotation

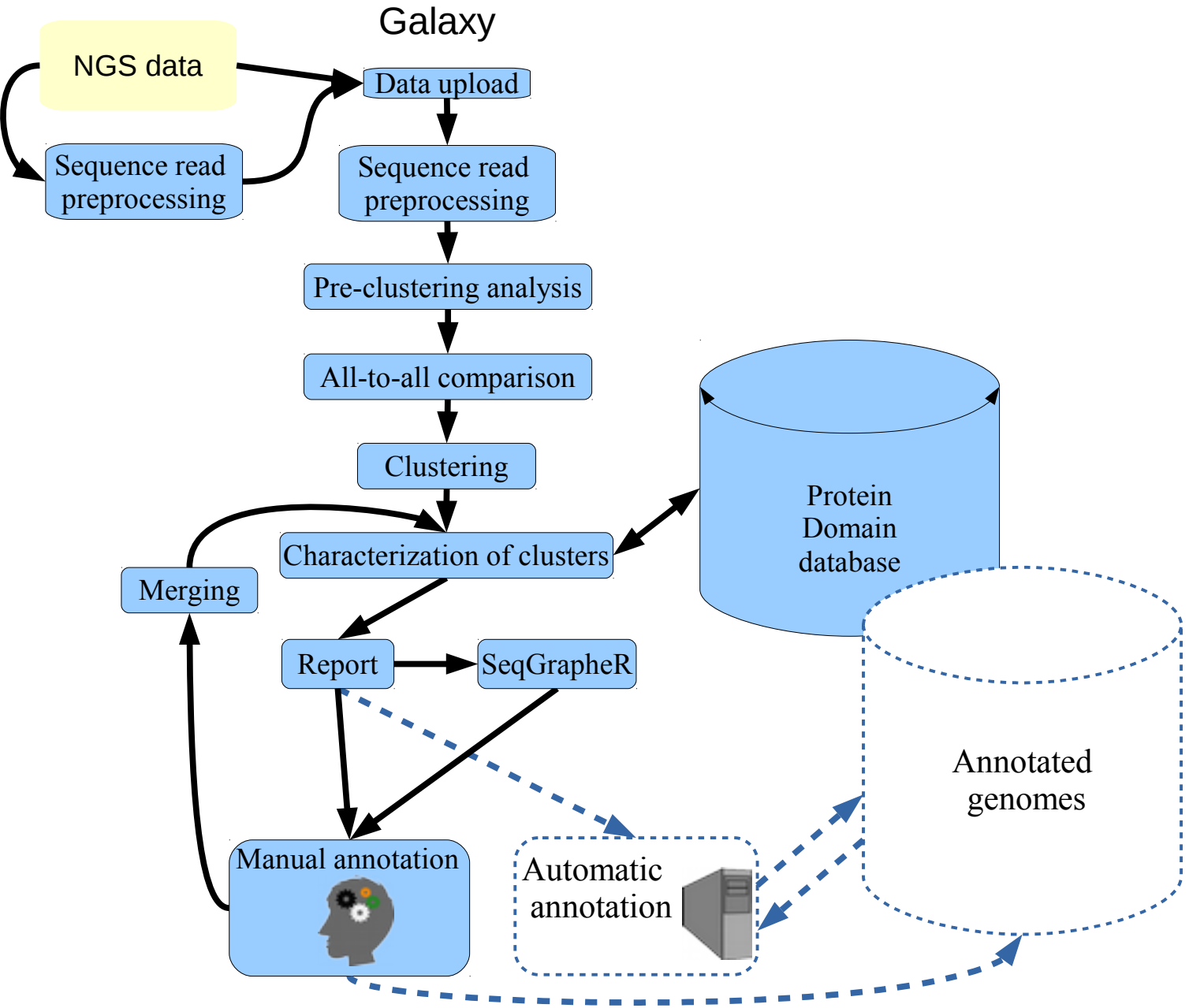
Performance:

False positive rate ~> “0%”
False negative ~> 1-5%

RepeatExplorer

Discover repeats in your next generation sequencing data

Analysis work-flow



Automatic annotation

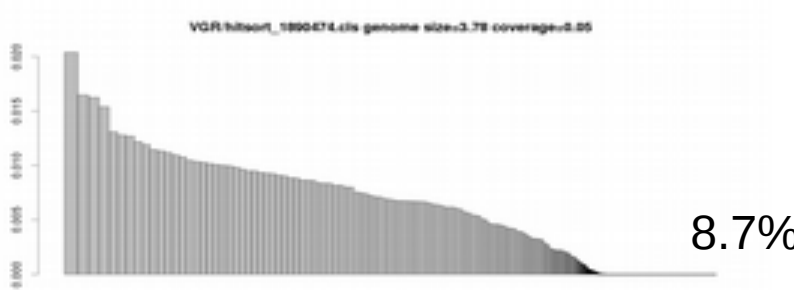
Reproducibility of sequence clustering

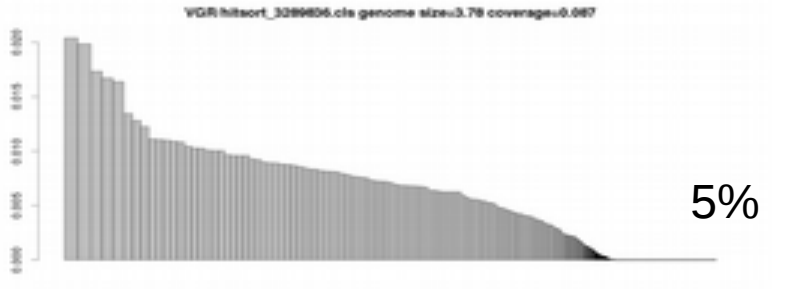
$$Coverage = \frac{Read\ Length \cdot Number\ of\ Reads}{Genome\ size}$$

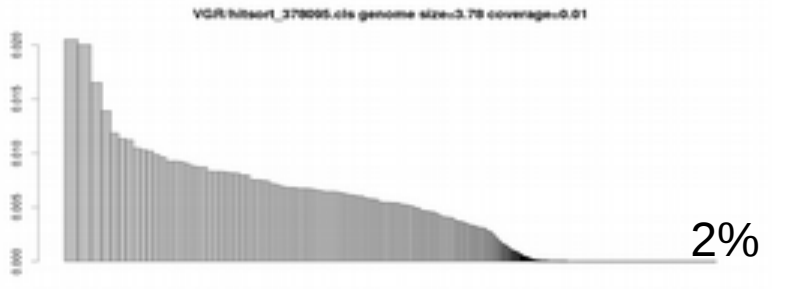
Effect of coverage on:

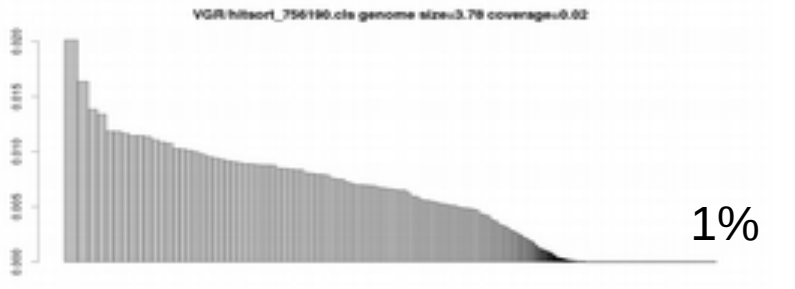
- cluster stability
- estimation of genomic proportions of repeats

Effect of different genome coverage
100 nt reads
V.grandiflora, 1C = 3.87pg



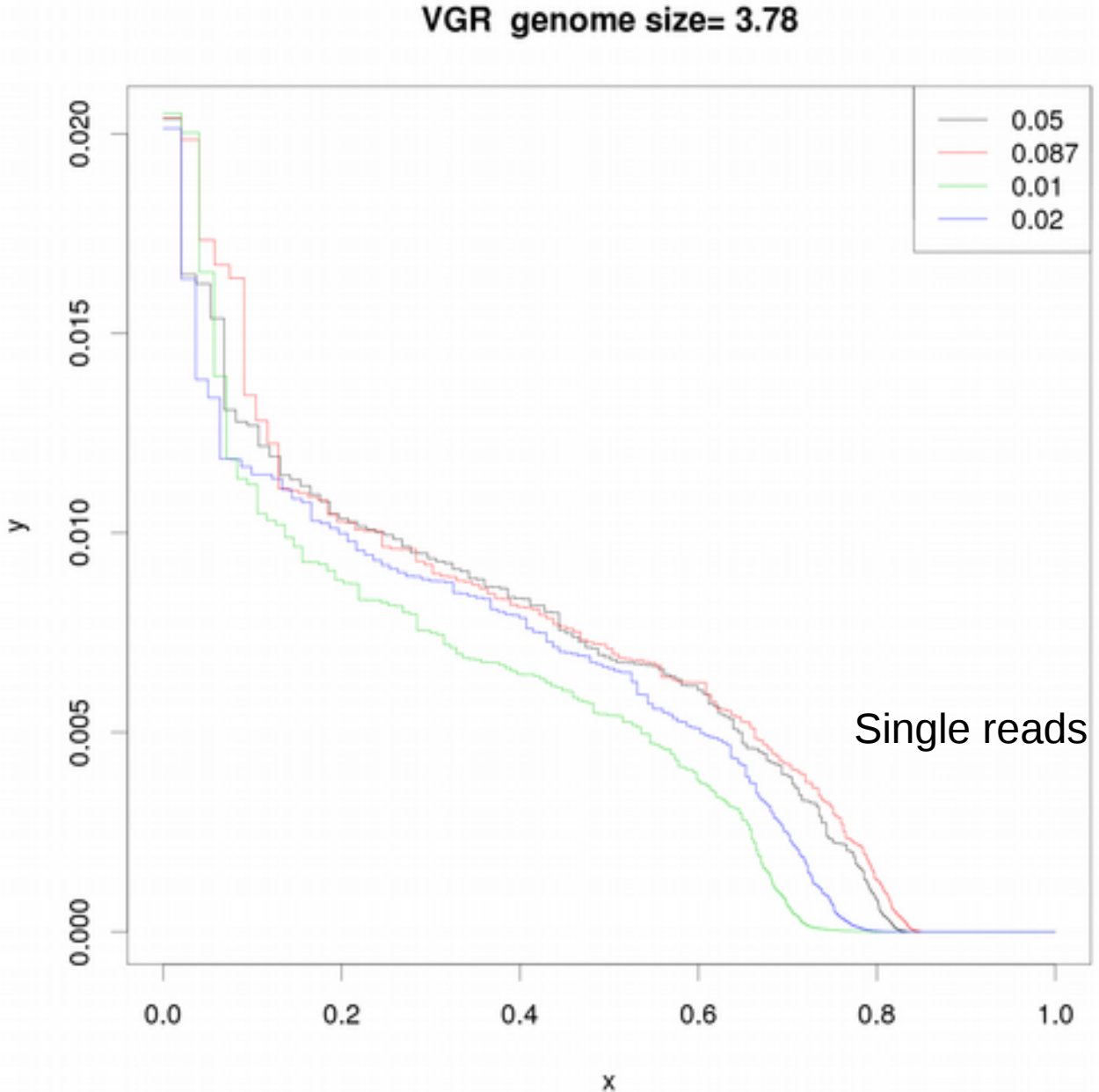


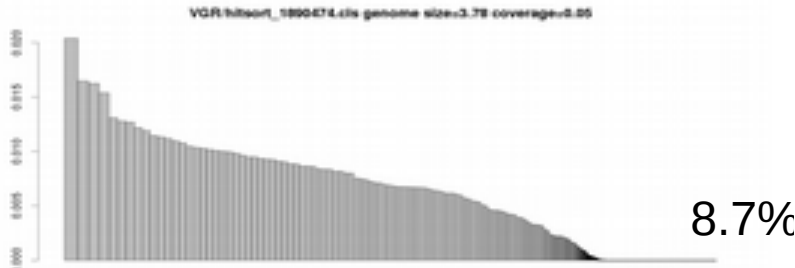


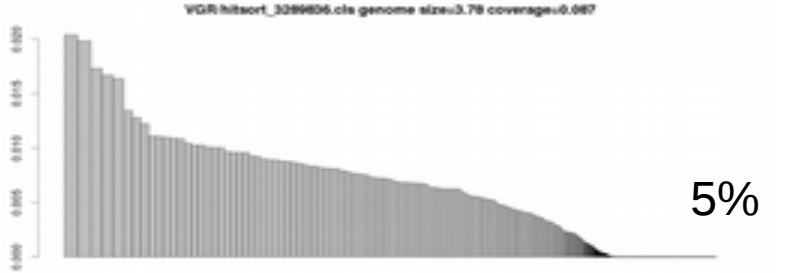


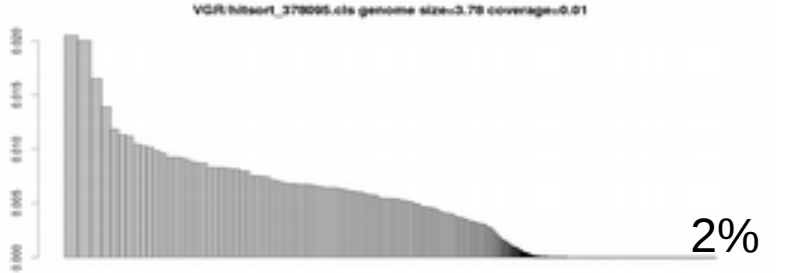
RepeatExplorer

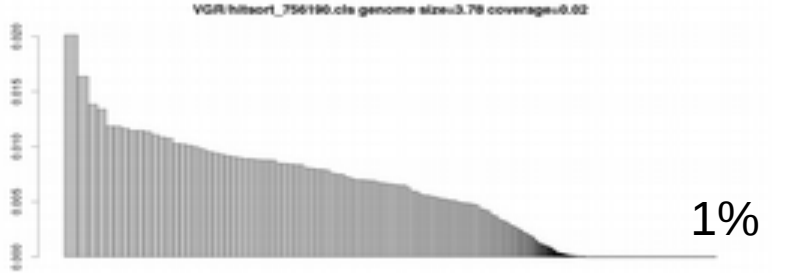
Discover repeats in your next generation sequencing data

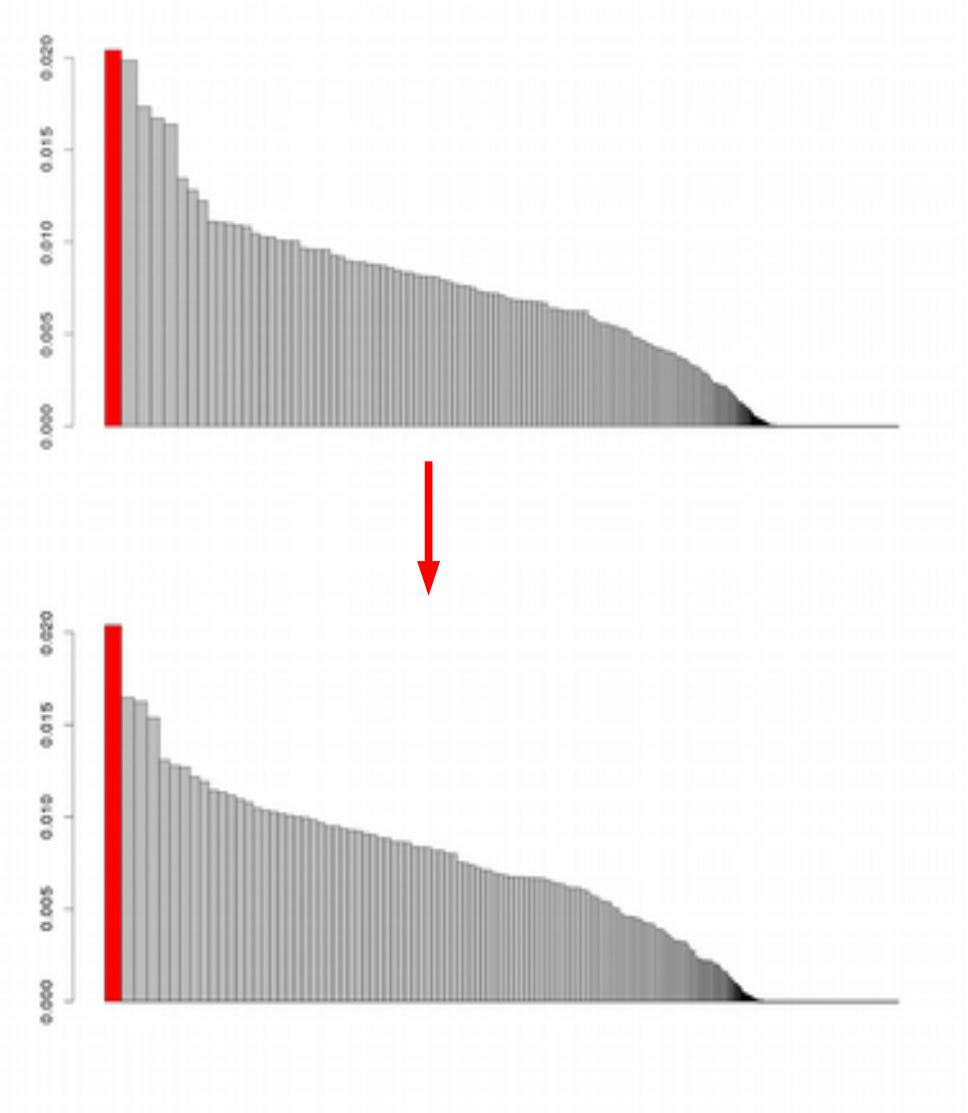






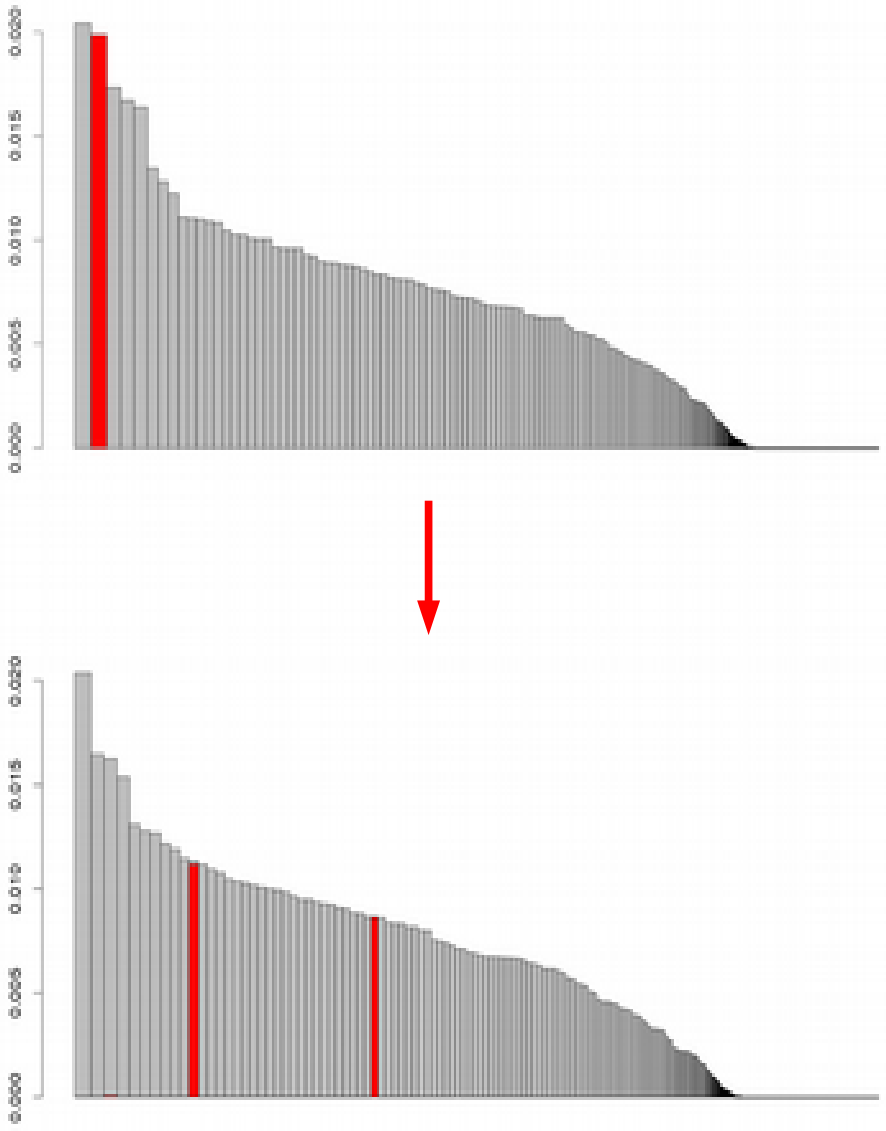






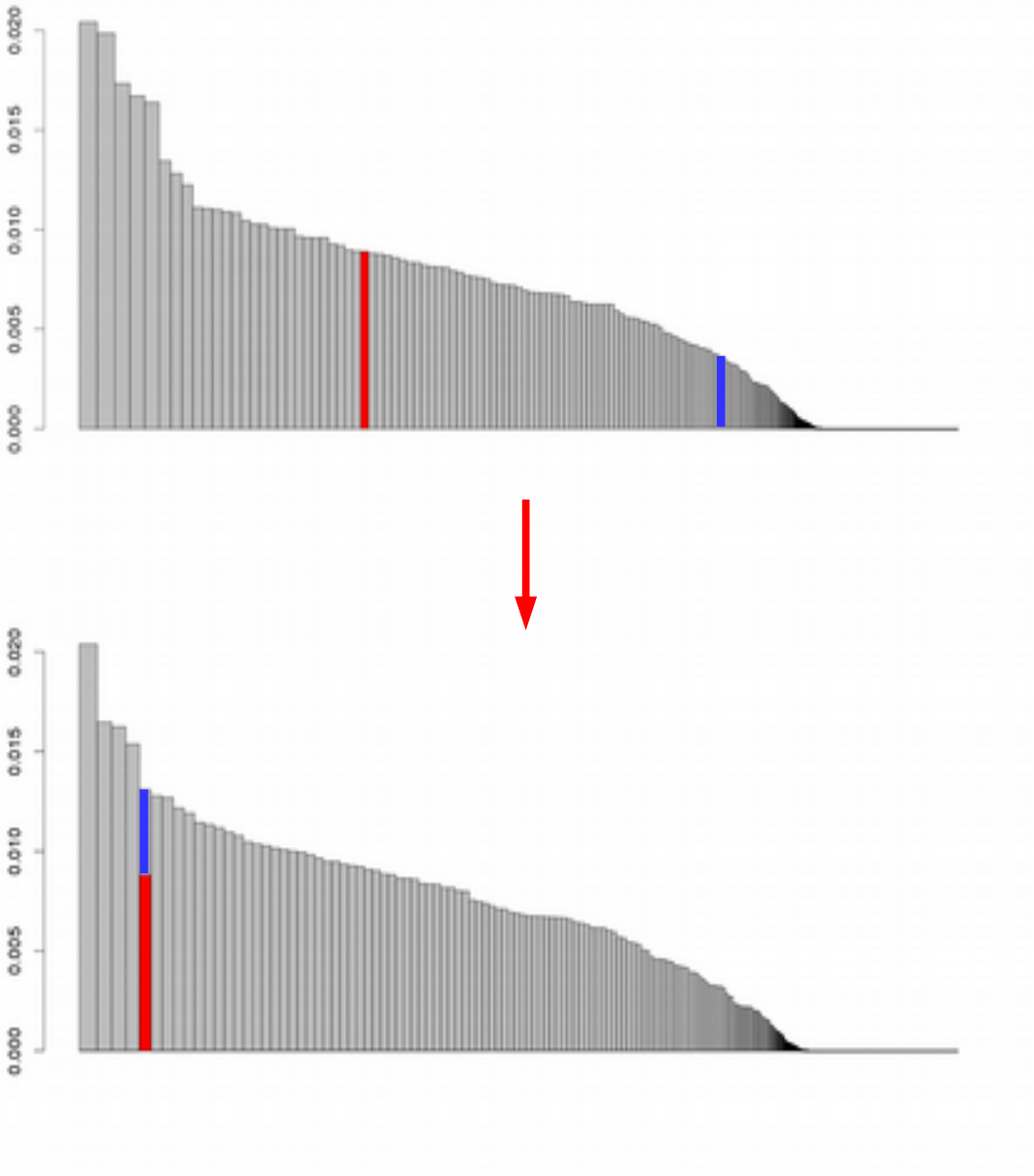
Effect of coverage on cluster stability

1 → 1



Effect of coverage on cluster stability

1 → **2**



Effect of coverage on cluster stability

2 → 1

first 50 largest cluster

Effect of coverage on cluster stability

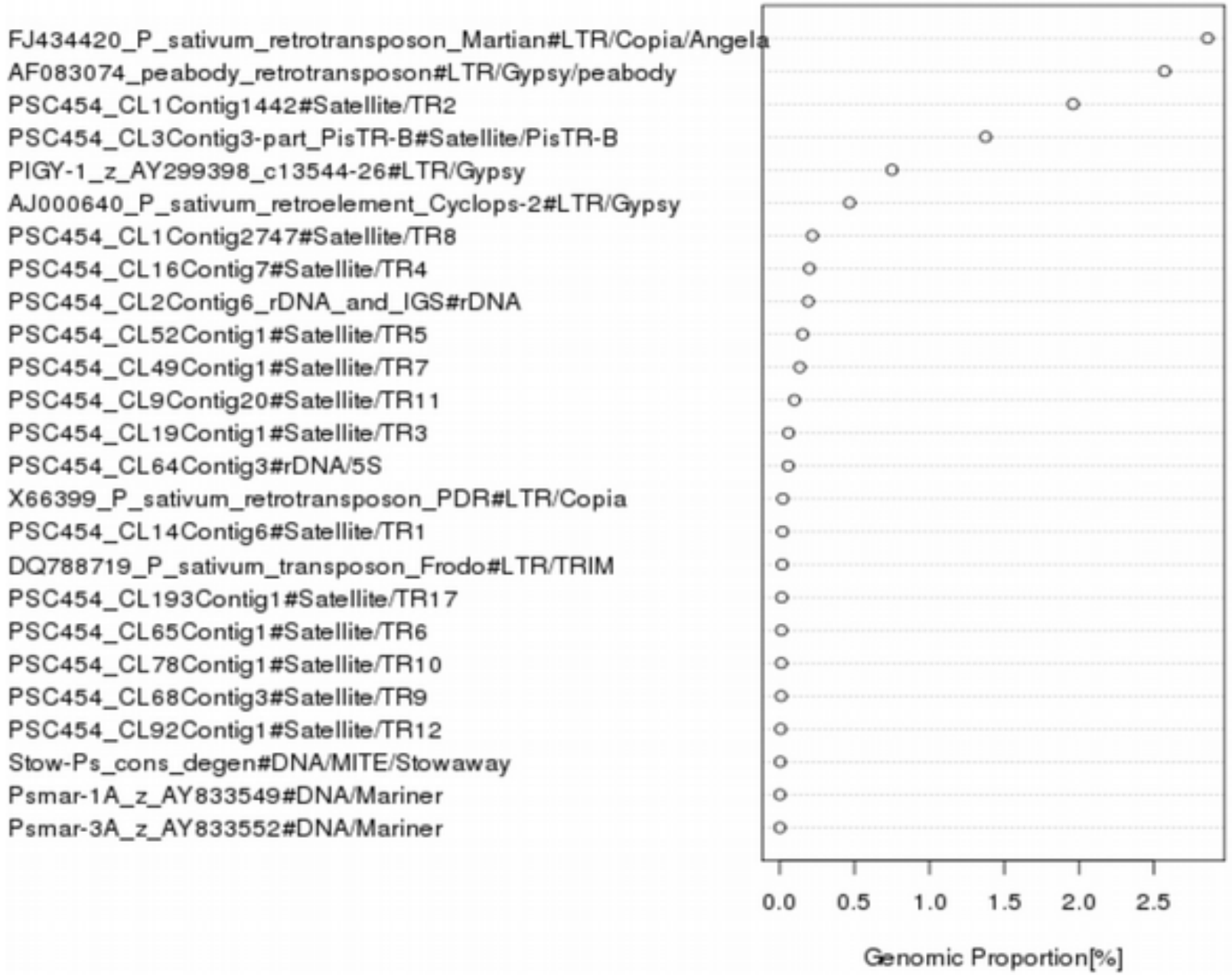
coverage

Type of transition	5%	2%	1%
1 → 1	41	40	25
1 → 2	5	8	18
2 → 1	4	2	11

RepeatExplorer

Discover repeats in your next generation sequencing data

Estimated genomic proportions of repetitive elements

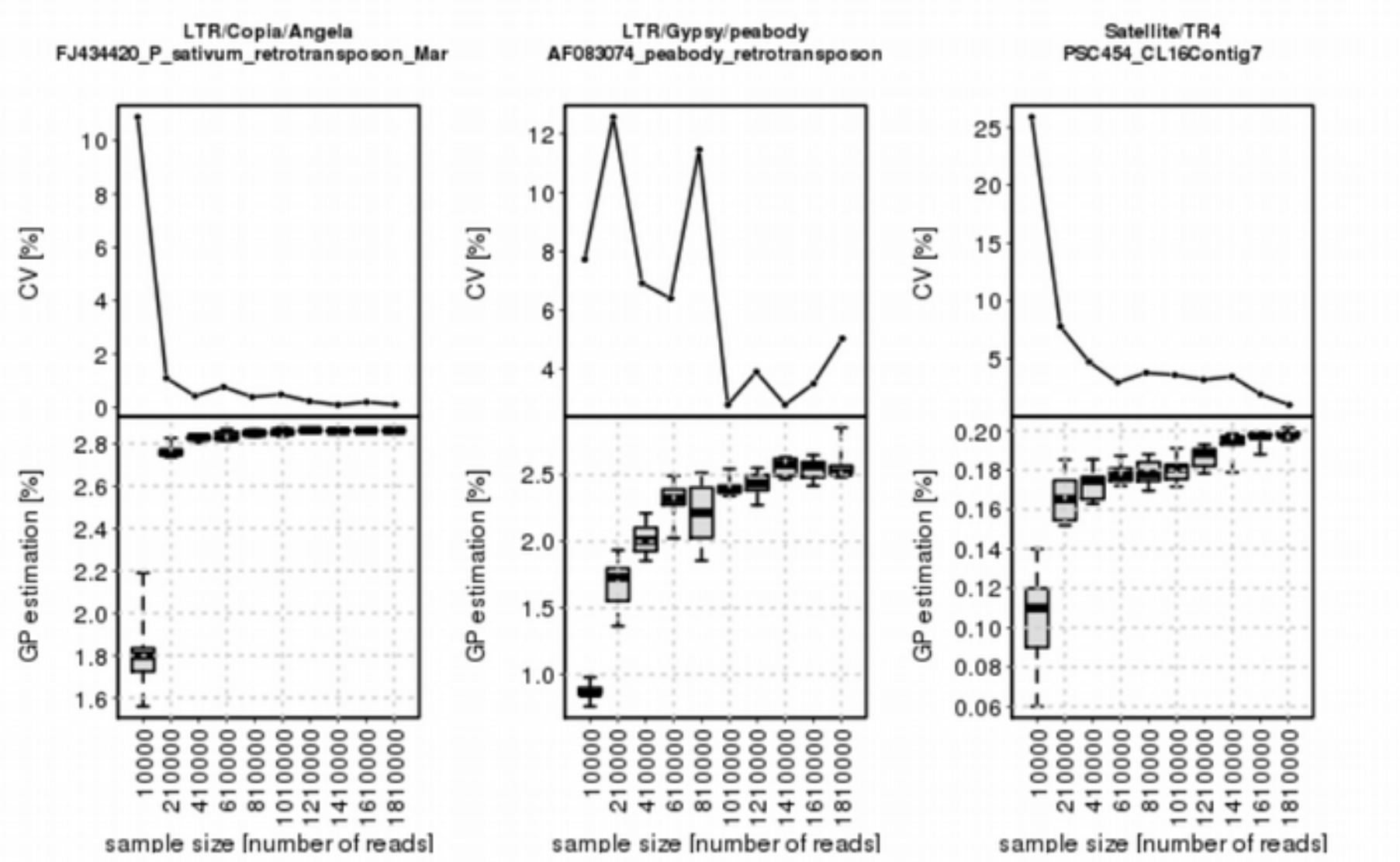


RepeatExplorer

Discover repeats in your next generation sequencing data

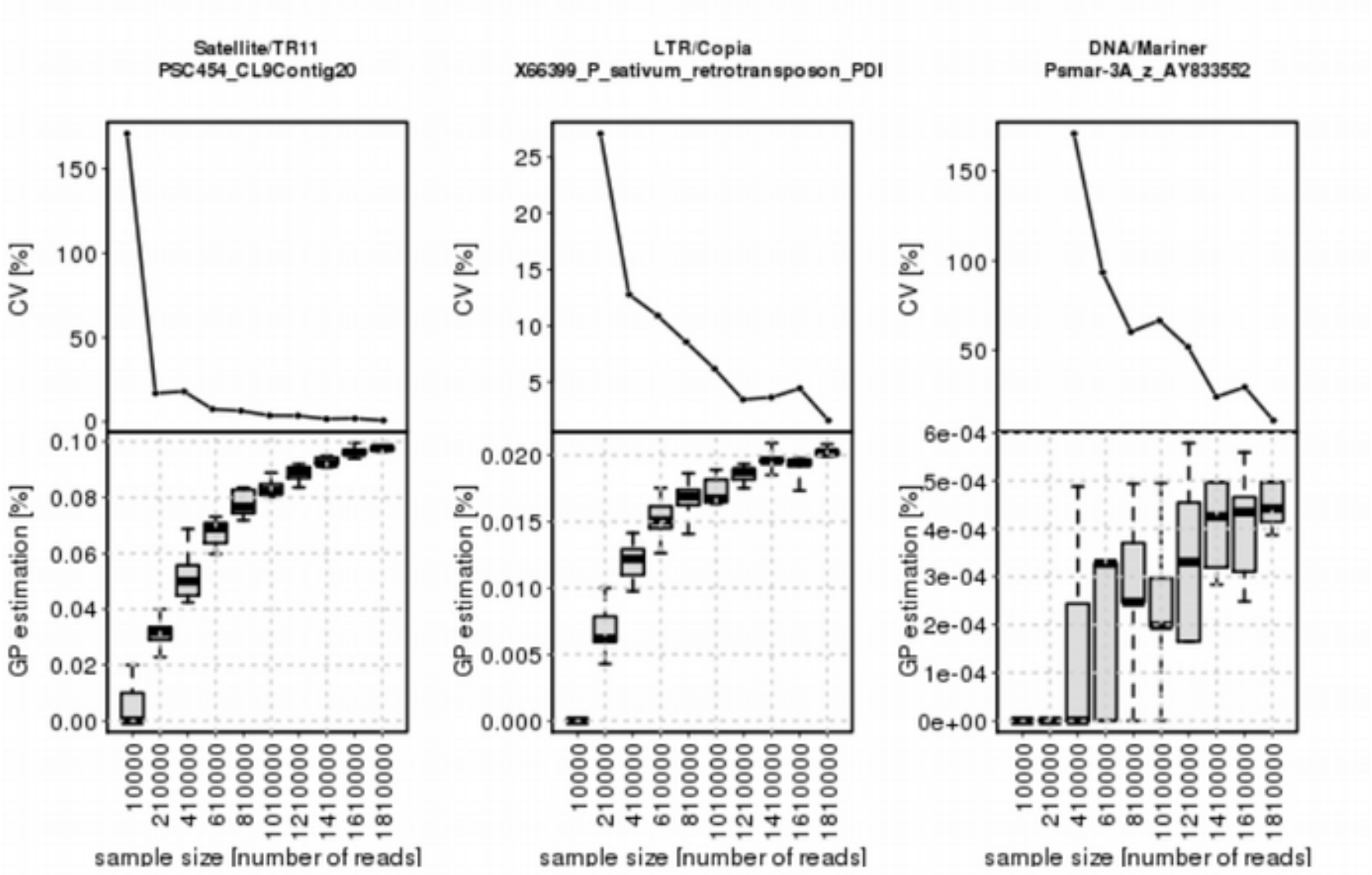
Effect of coverage on estimation of genomic proportions

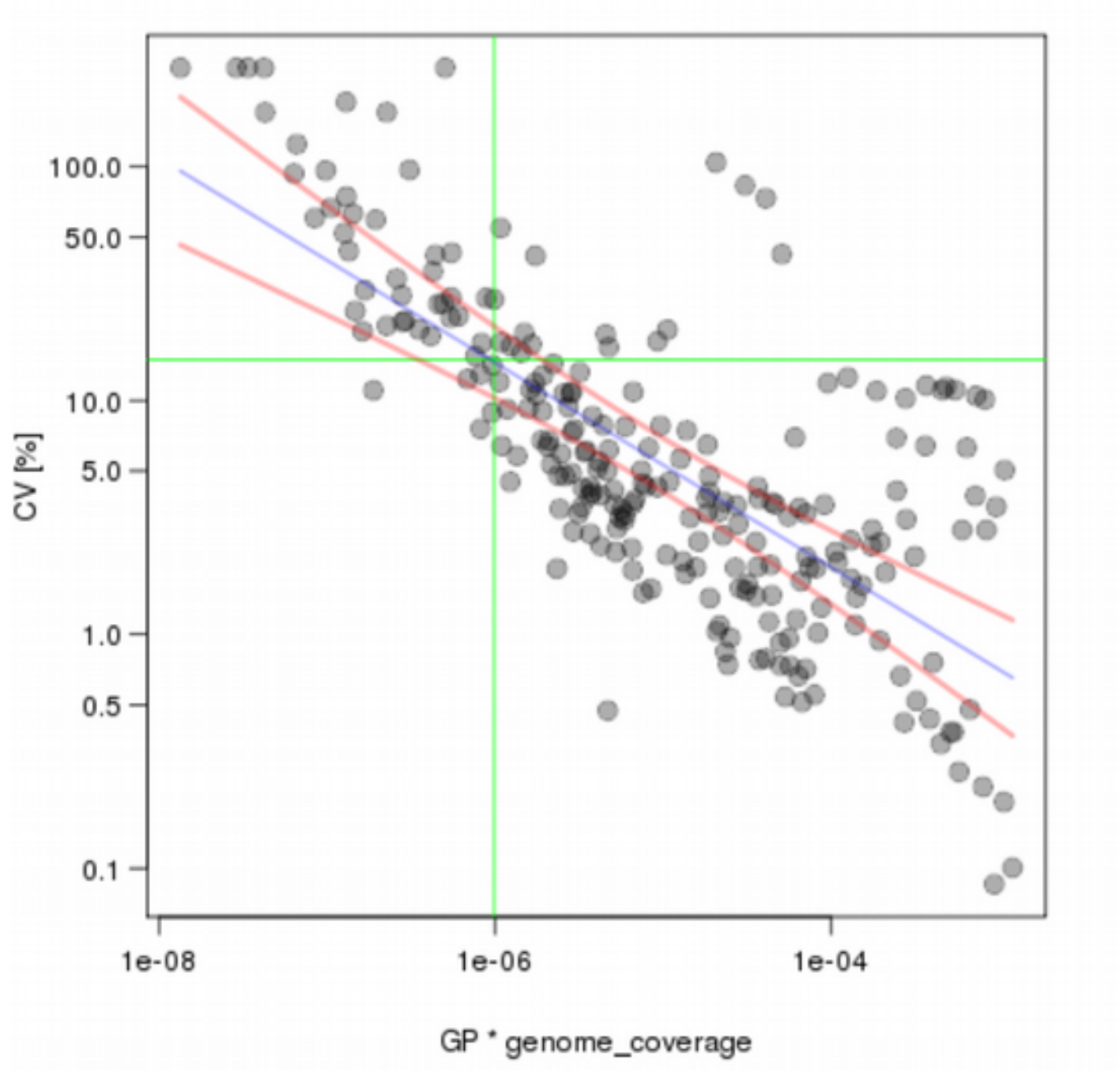
P.sativum repetitive elements



Effect of coverage on estimation of genomic proportions

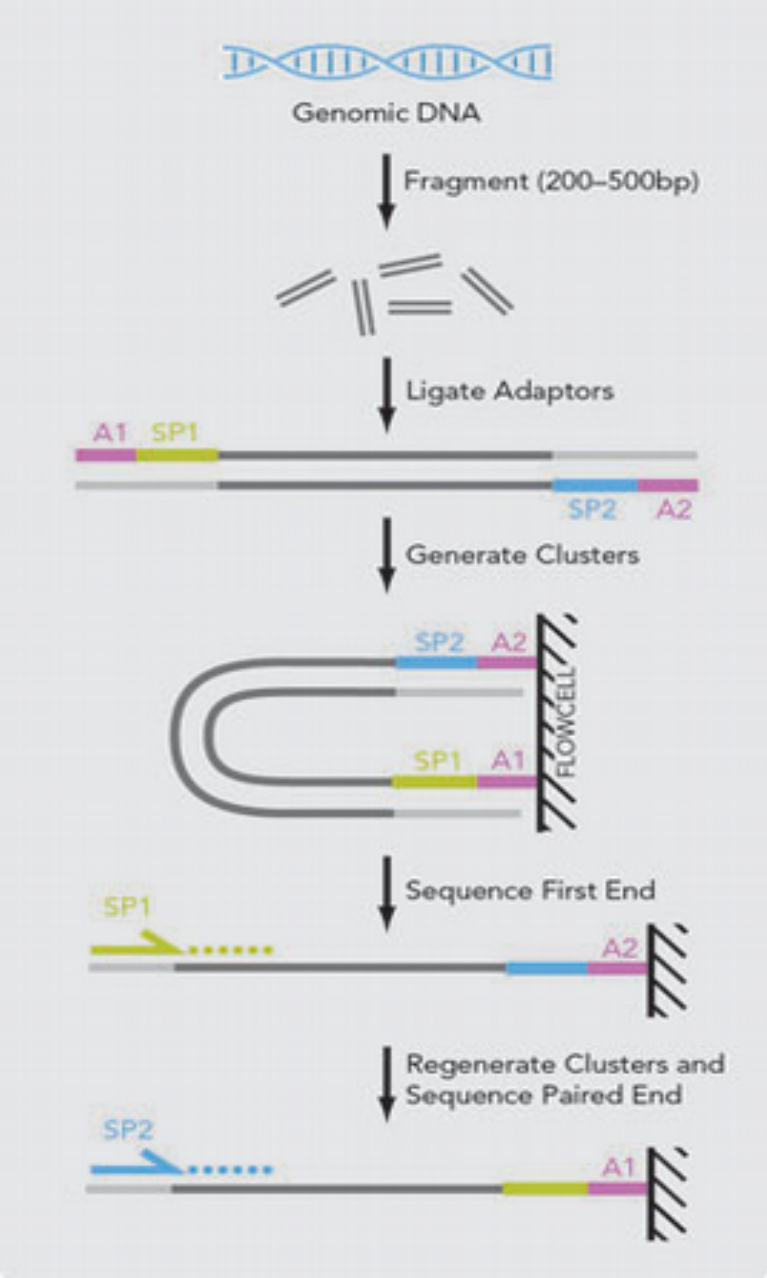
P.sativum repetitive elements





~1% coverage *P.sativum*
CV <15% for all repeats
above 0.01% GP

Illumina Paired-End Sequencing



Length of DNA fragments used for sequencing

Sometime fragments are too short!
Read-through – **presence of adapters** at 3' ends
- clustering slow or impossible
- affect assembly

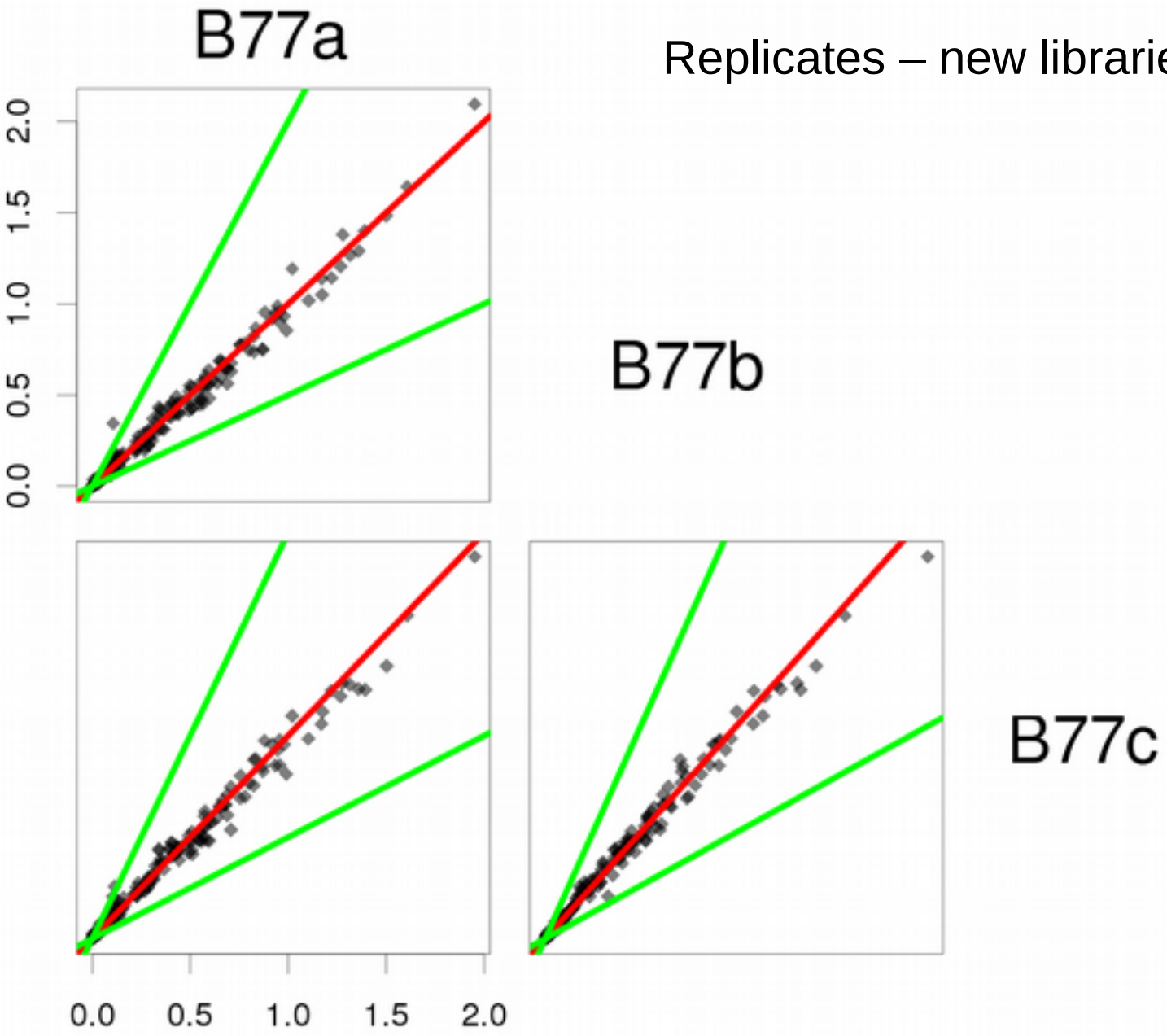
overlapping forward and reverse reads

- impact on quantification
- redundant data

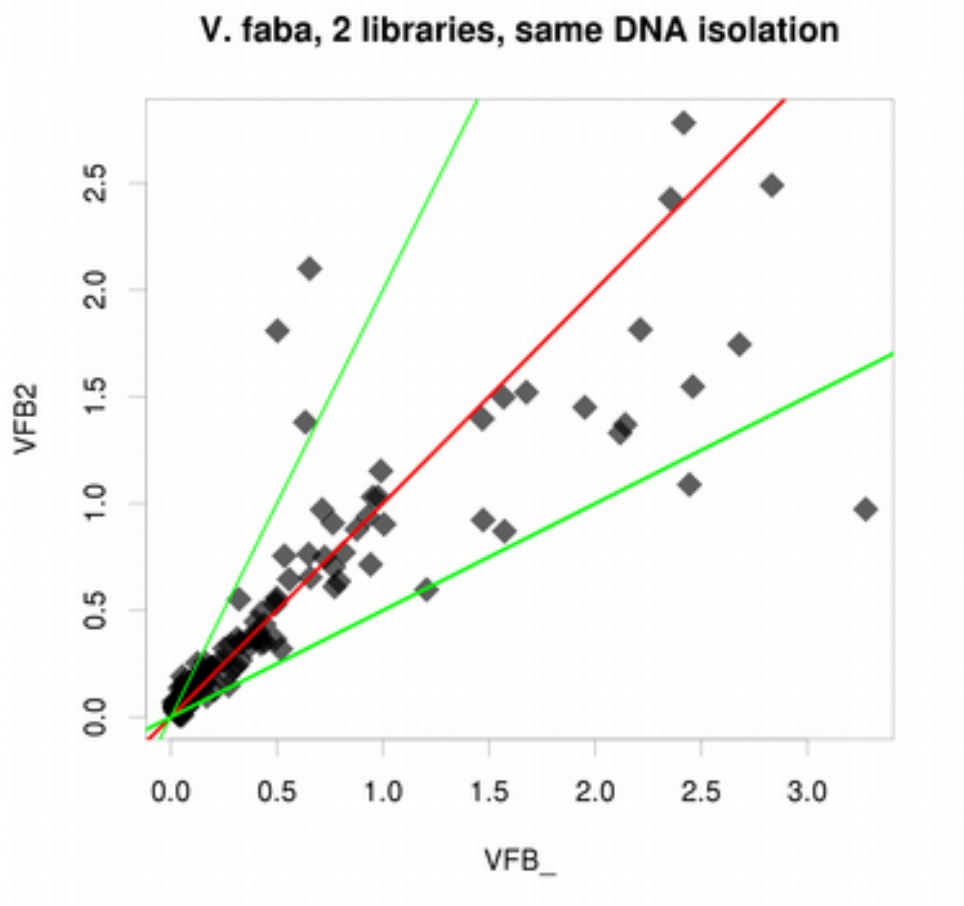
Replicates

- Technical replicates – libraries from the same DNA isolation
- **Biological replicates**

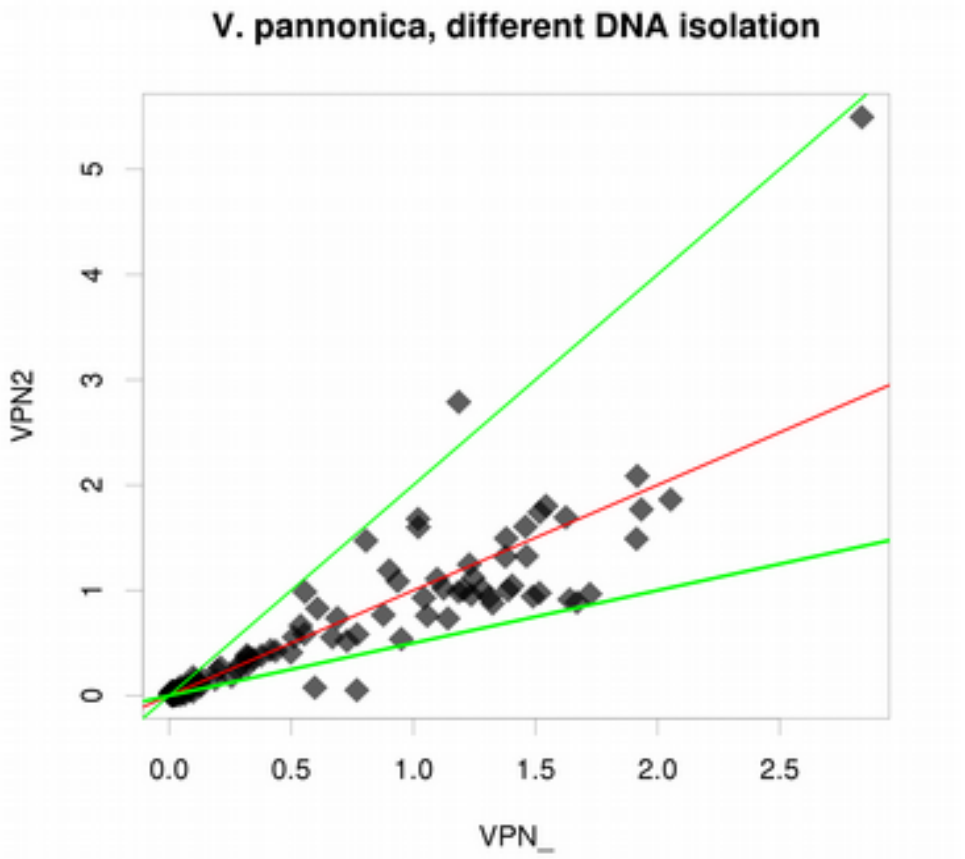
Replicates – new libraries and DNA isolation

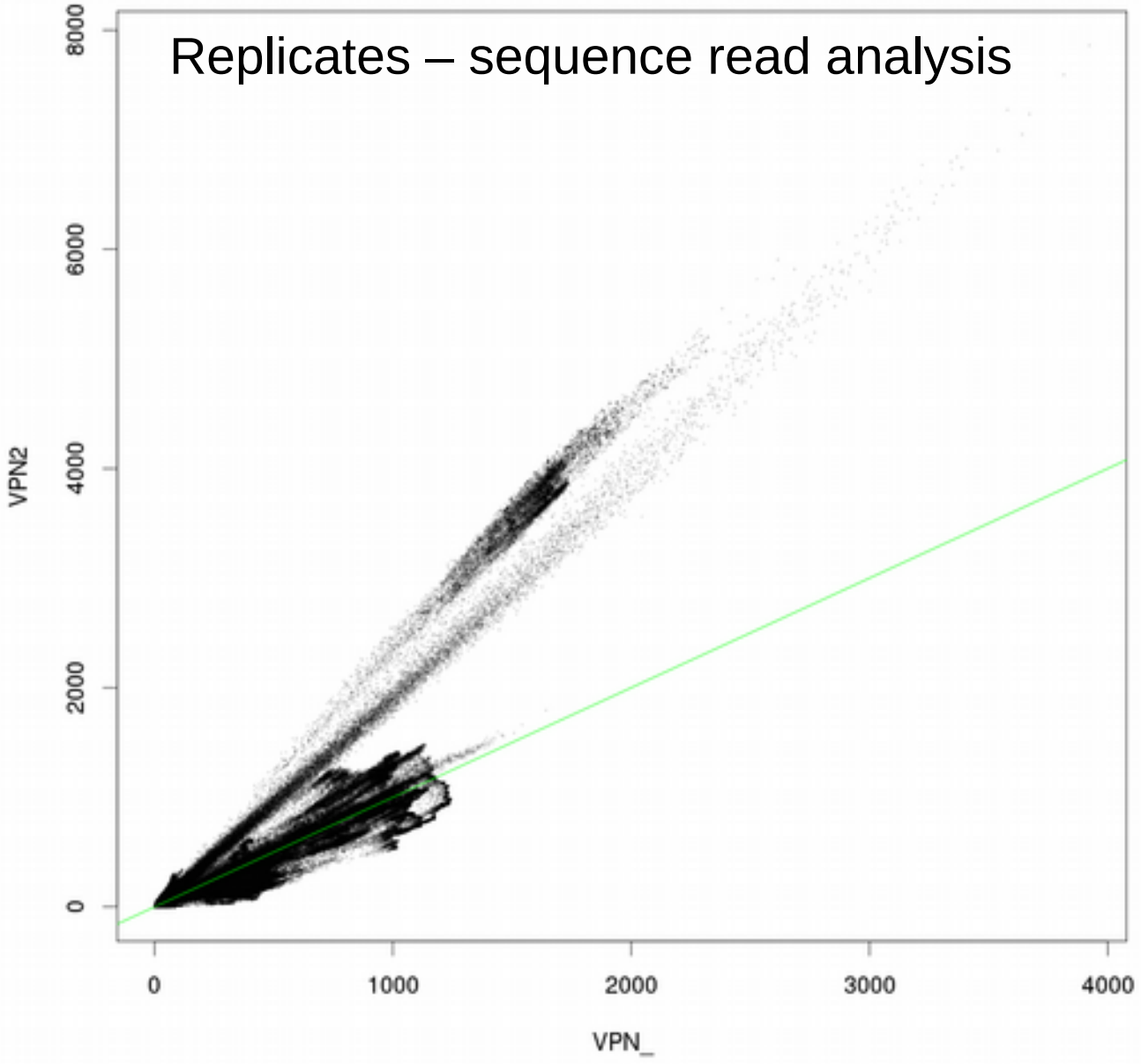


Technical replicates



Biological replicates





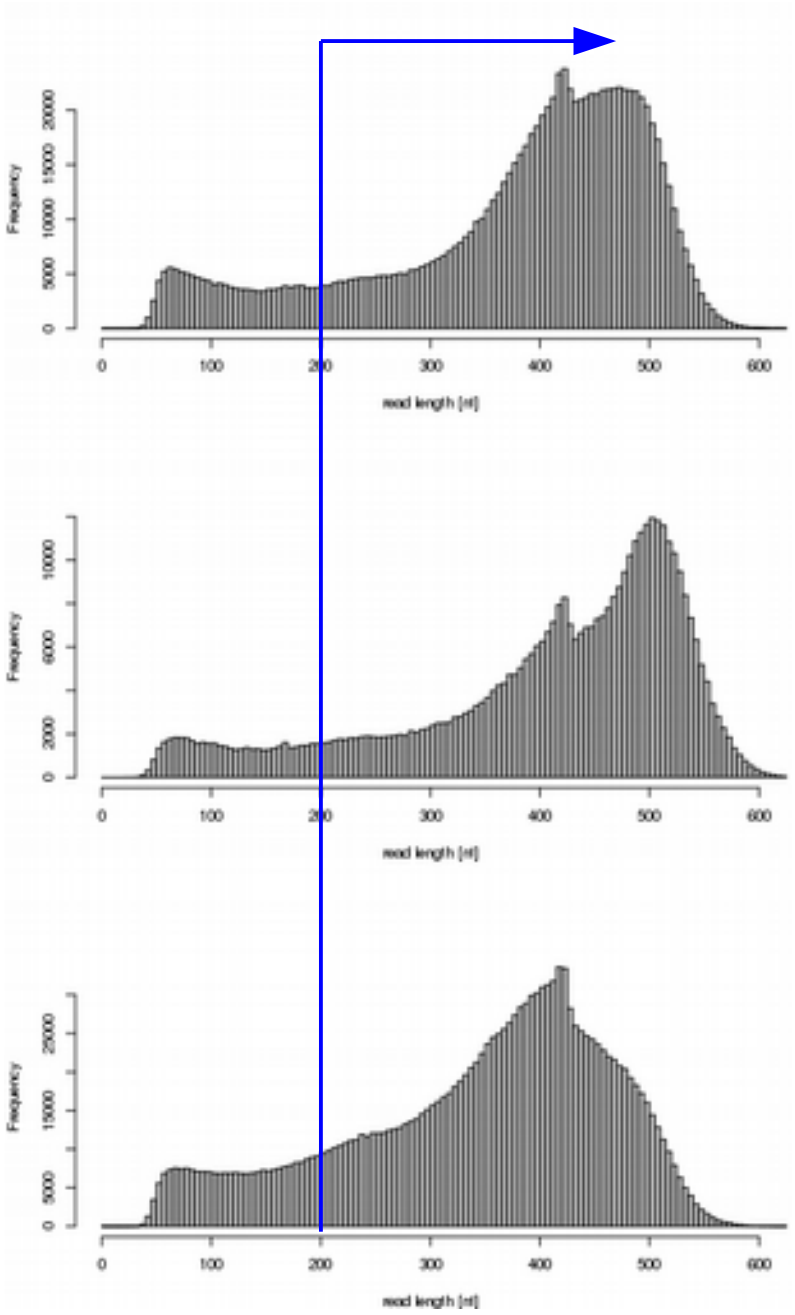
Tandem repeats, sequence copy number bias

Roche 454 reads

- Variable read length
- Trimming to the same length makes evaluation of clustering easier
- Data loss
- But there is a sequence bias

  - tandem repeats

- For assembly – it is better to use full length reads



Thank you for your attention!