Transposable Elements in the Era of Data Science

Hand-On/Demo session: TE detection and annotation with short-reads data

Anna-Sophie Fiston-Lavier (ISEM-U. Montpellier), Romain Guyot (IRD)

May 14th

Introduction

Short read data

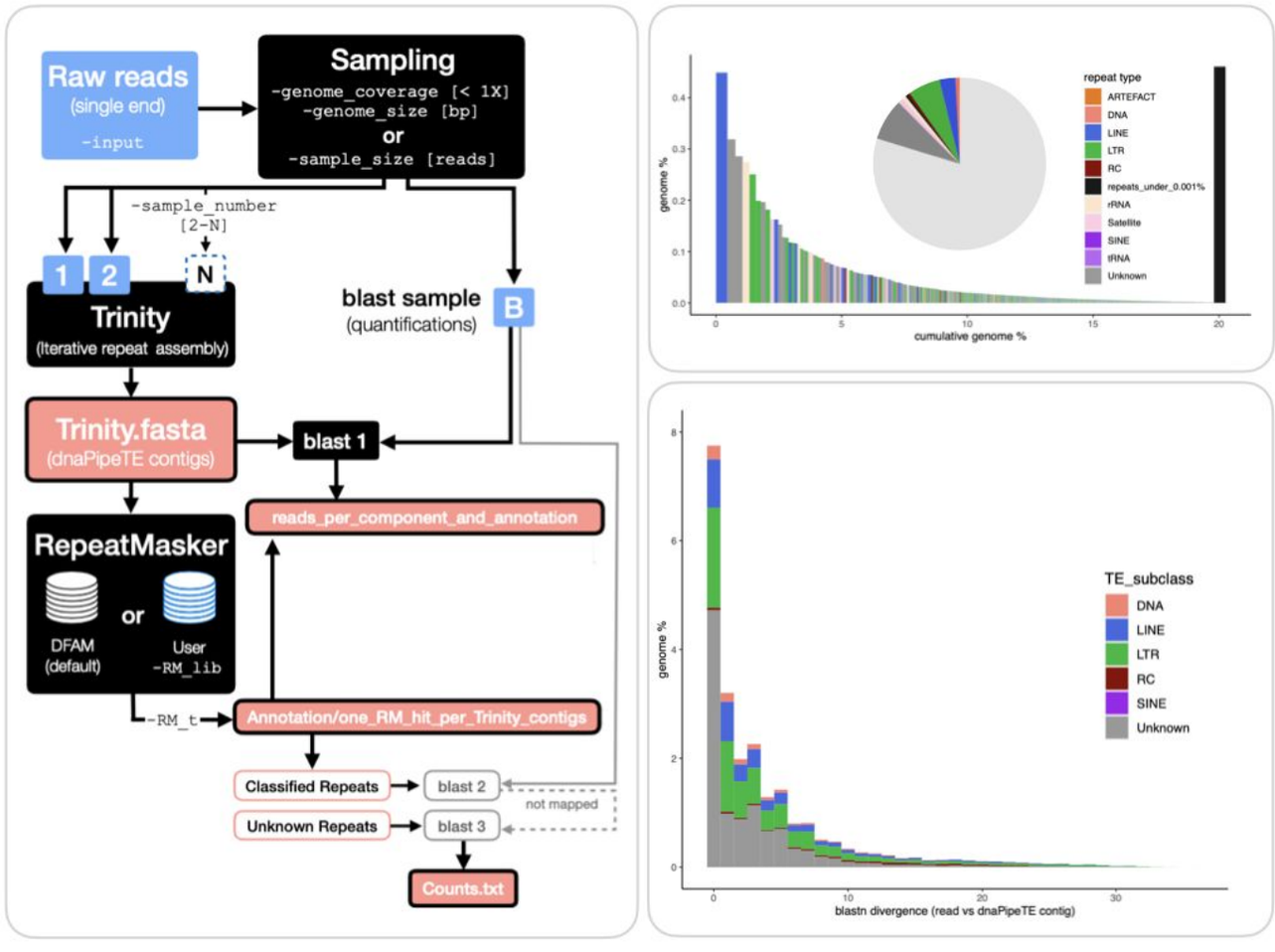
Why detect TEs in short reads? -> To study TE abundance, variation, insertion events, or activity

General Approaches:

1_ Mapping to a reference genome with TE annotation (TIPS_finder, TEtools, McClintock..)

2_ Mapping to a TE database

3_ *De novo* (dnaPipeTE, RepeatExplorer)



https://github.com/clemgoub/dnaPipeTE

Introduction

Short read data

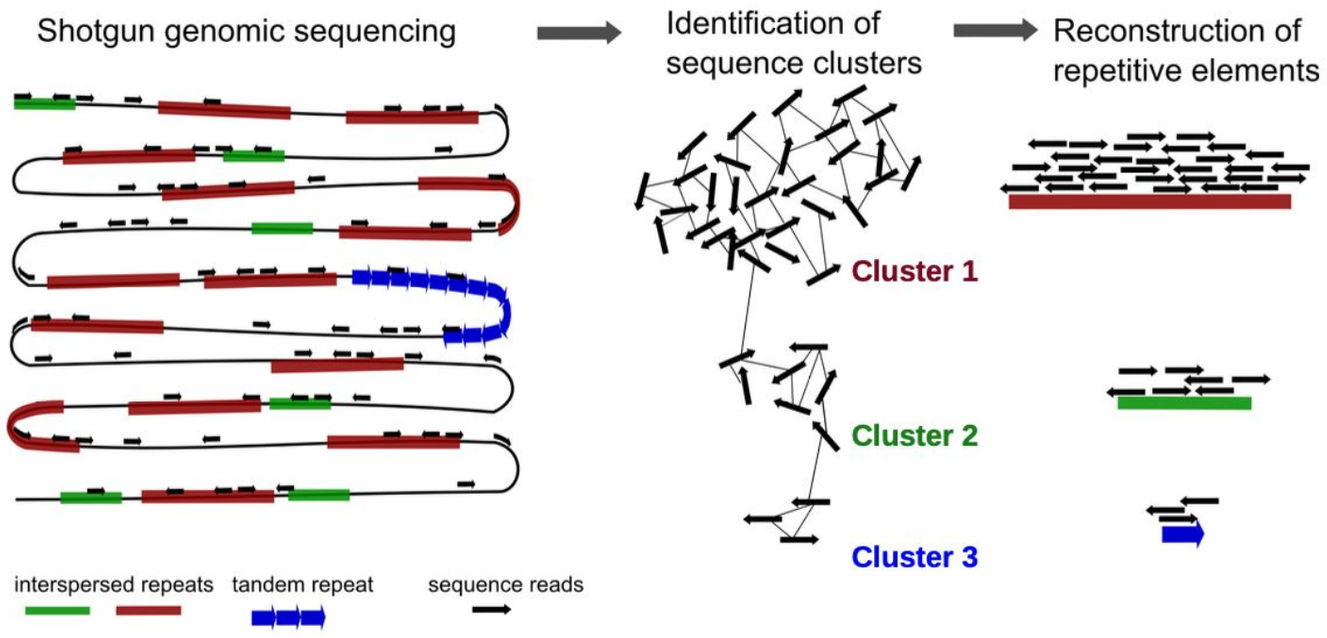
Why detect TEs in short reads? -> To study TE abundance, variation, insertion events, or activity

General Approaches:

1_ Mapping to a reference genome with TE annotation (TIPS_finder, TEtools, McClintock..)

2_ Mapping to a TE database

3_ *De novo* **RepeatExplorer2**



Low coverage sequencing
CLUSTER = a set of frequently overlapping reads = REPEAT FAMILY

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

Introduction

Short read data

Why detect TEs in short reads? ->To study TE abundance, variation, insertion events, or activity

General Approaches:

1_ Mapping to a reference genome with TE annotation (TIPS_finder, TEtools, McClintock..)

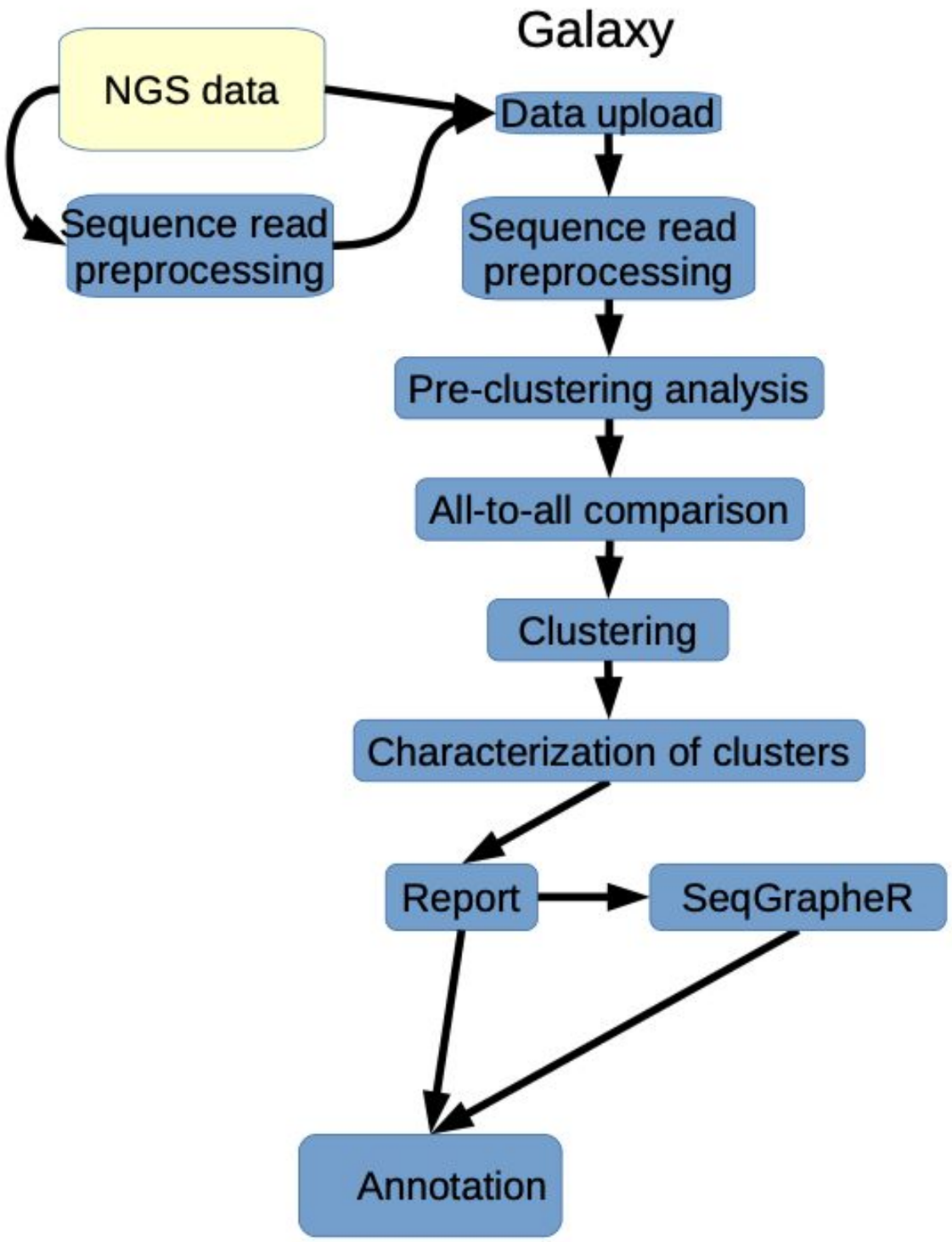
2_ Mapping to a TE database

3_ *De novo*: RepeatExplorer2 Graph based clustering

Sequence reads: 100 bp PE

100,000 to 10 M reads

55 bp overlap and 90% similarities



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

Introduction

Short read data

Why detect TEs in short reads? ->To study TE abundance, variation, insertion events, or activity

General Approaches:

1_ Mapping to a reference genome with TE annotation (TIPS_finder, TEtools, McClintock..)

2_ Mapping to a TE database

3_ *De novo*: RepeatExplorer2 Graph based clustering

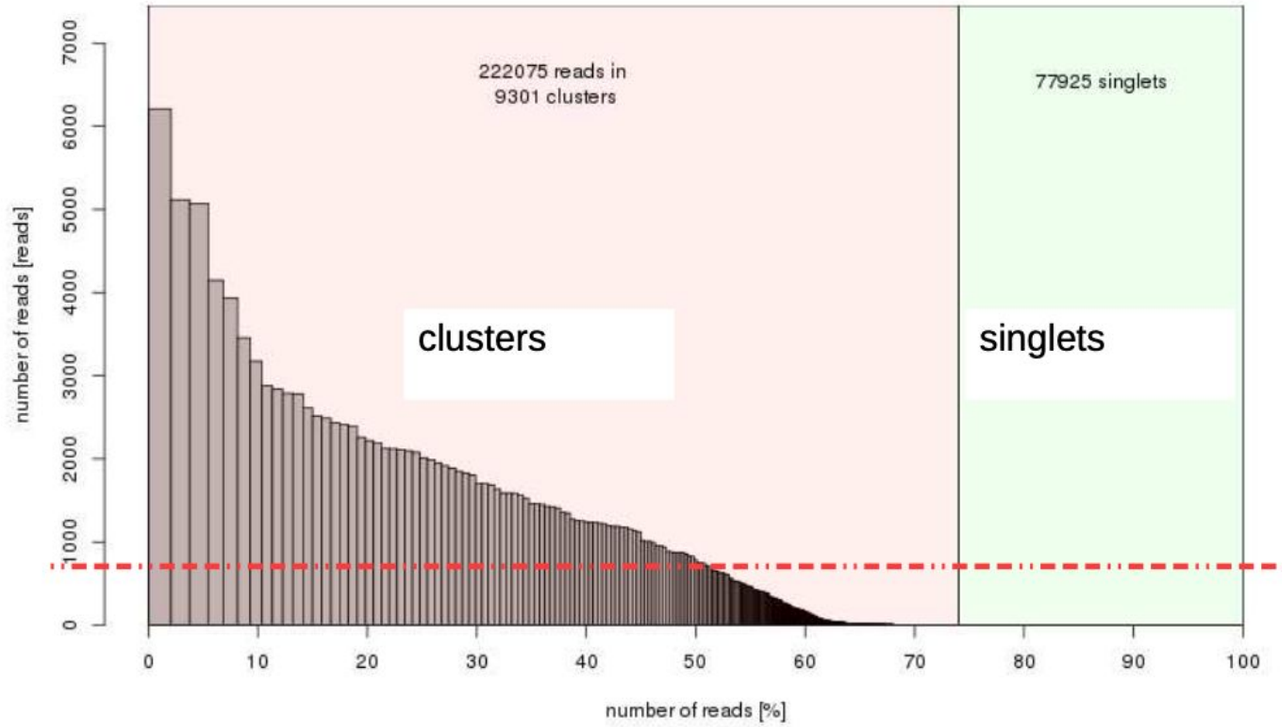
Sequence reads: 100 bp PE

100,000 to 10 M reads

55 bp overlap and 90% similarities

Characterization of cluster

RepeatMasker search (Repbase, custom db)



Typical cluster size distribution

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

Introduction

Short read data

Why detect TEs in short reads? ->To study TE abundance, variation, insertion events, or activity

General Approaches:

1_ Mapping to a reference genome with TE annotation (TIPS_finder, TEtools, McClintock..)

2_ Mapping to a TE database

3_ *De novo*: RepeatExplorer2 Graph based clustering

Sequence reads: 100 bp PE

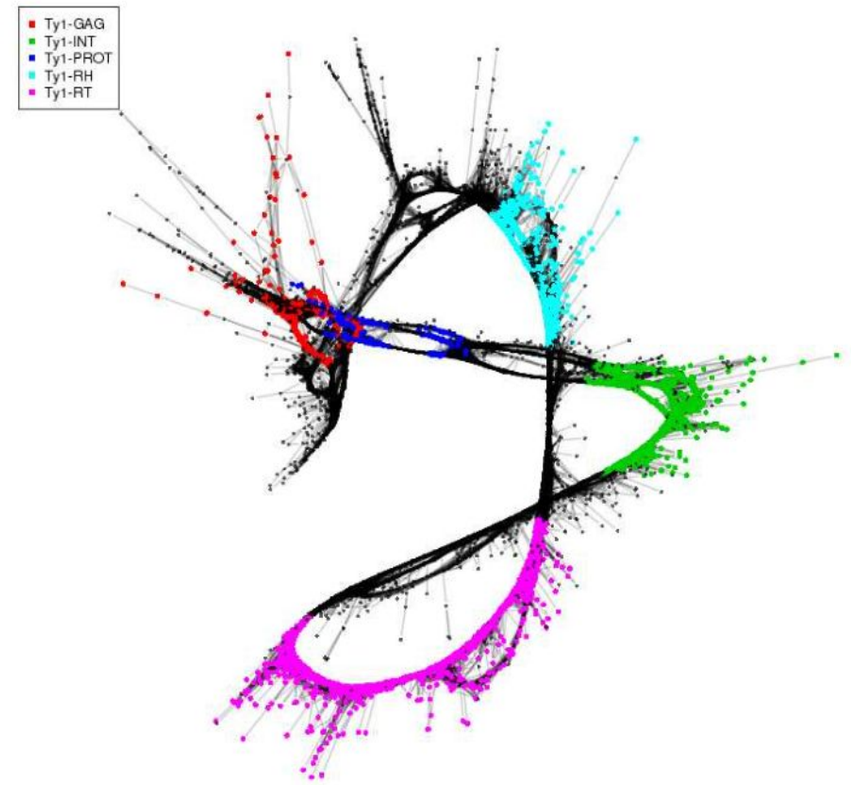
100,000 to 10 M reads

55 bp overlap and 90% similarities

Characterization of cluster

RepeatMasker search (Repbase, custom db)

Protein domain database + Graph layouts



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

Introduction

Short read data

Why detect TEs in short reads? ->To study TE abundance, variation, insertion events, or activity

General Approaches:

1_ Mapping to a reference genome with TE annotation (TIPS_finder, TEtools, McClintock..)

2_ Mapping to a TE database

3_ *De novo*: RepeatExplorer2 Graph based clustering

Sequence reads: 100 bp PE

100,000 to 10 M reads

55 bp overlap and 90% similarities

Characterization of cluster

RepeatMasker search (Repbase, custom db)

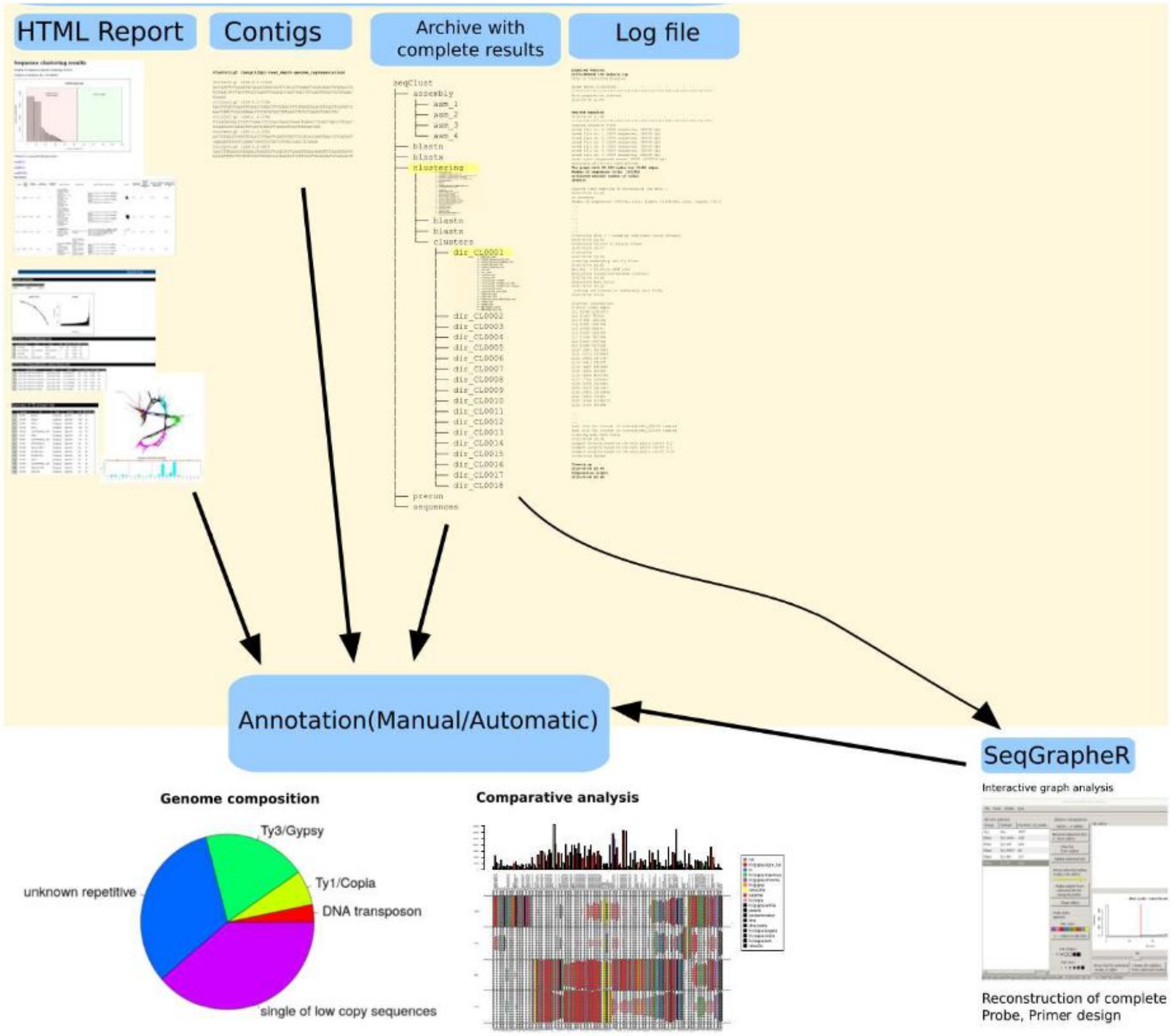
Output: Html report, contigs, log,…

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, 15.3%**) 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 (1110 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

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

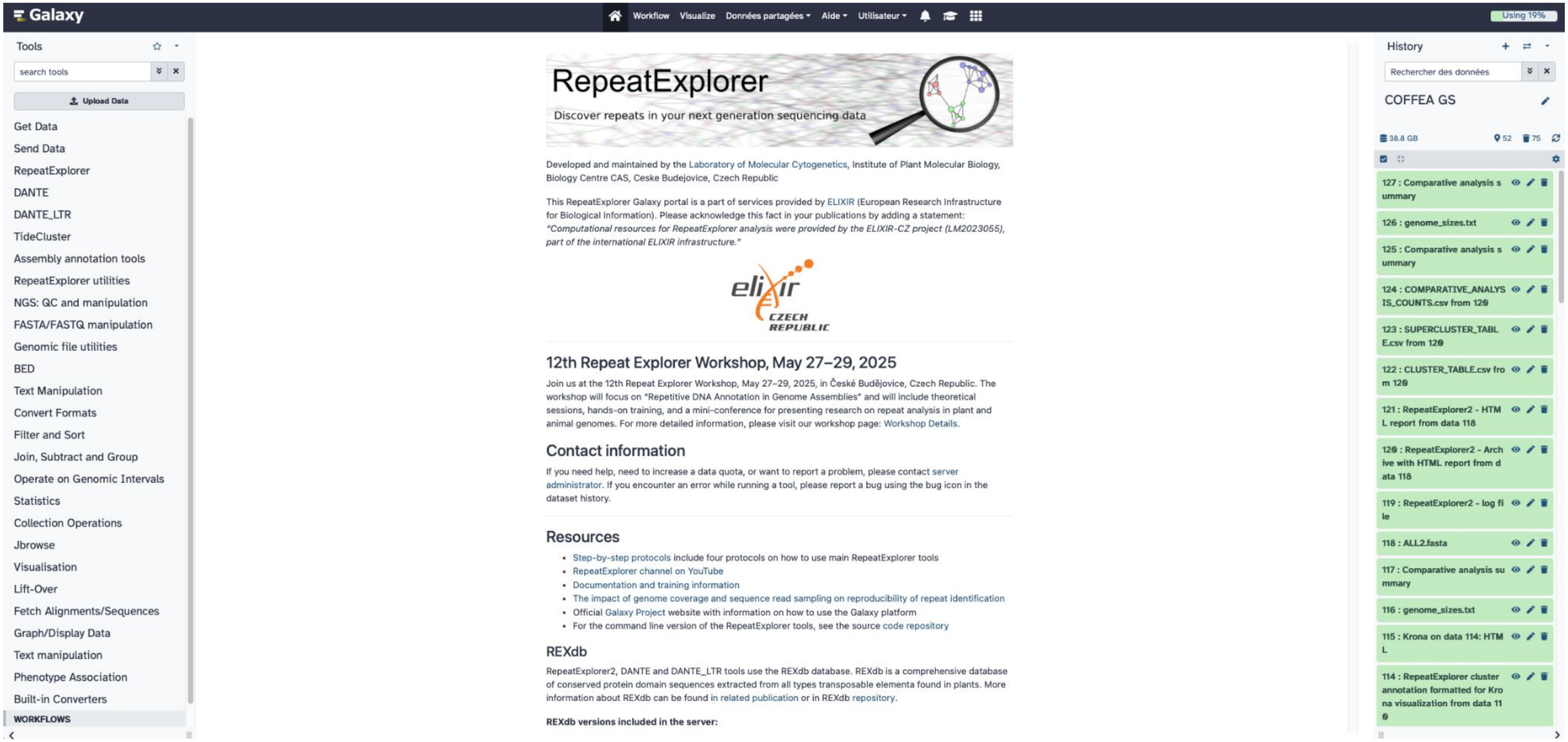
Introduction



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

Introduction

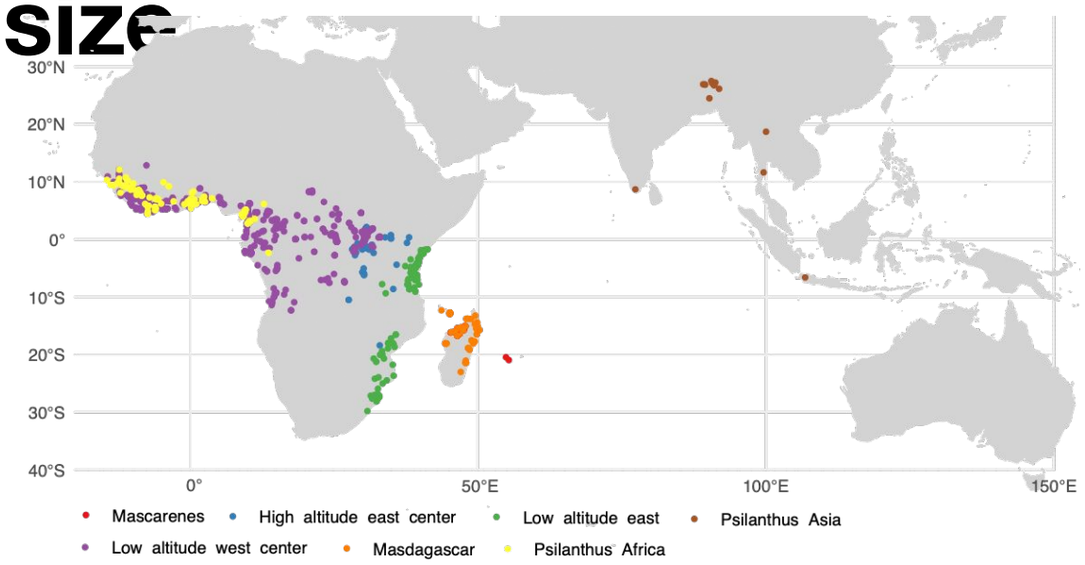
Short read data

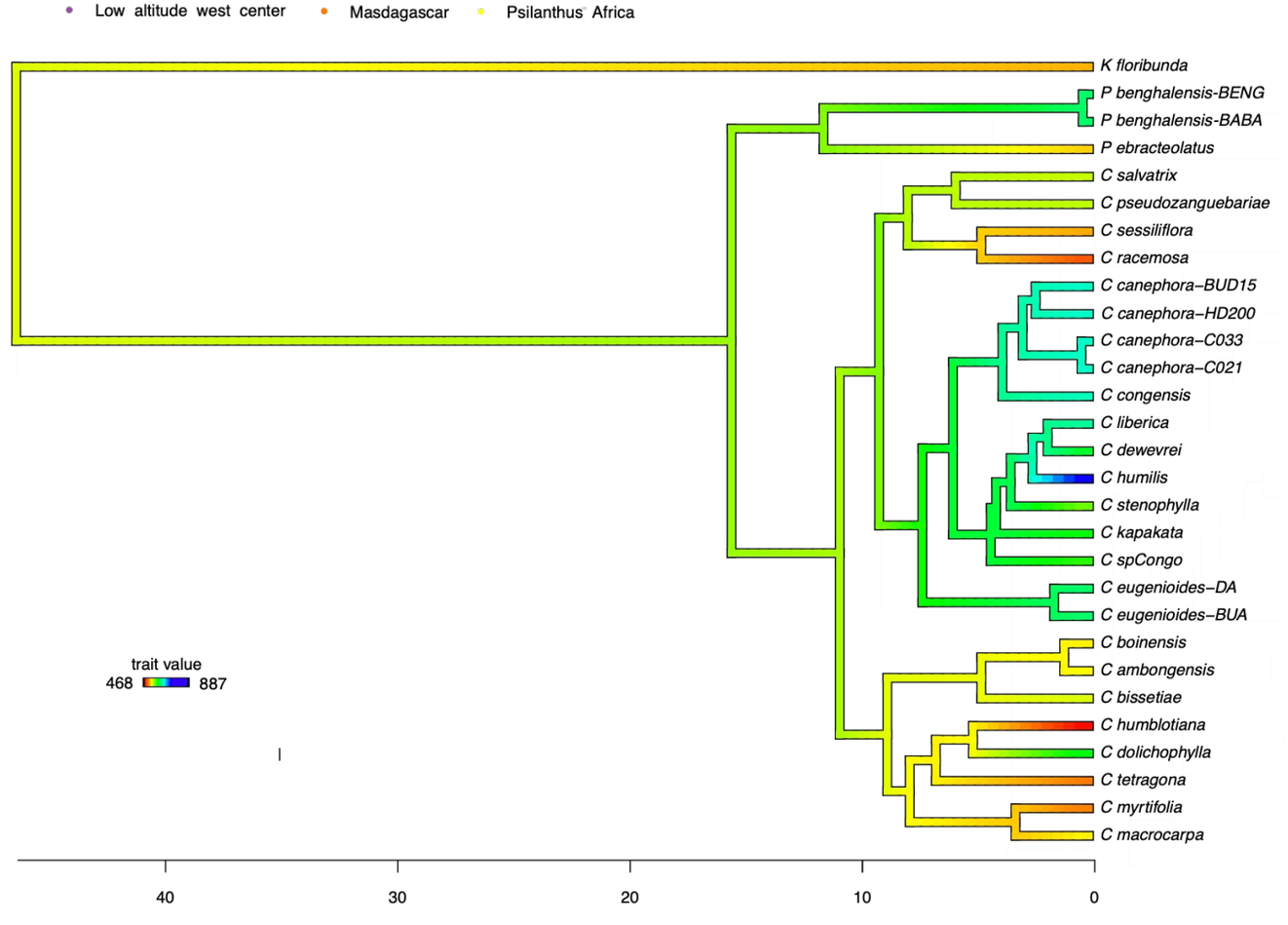


https://repeatexplorer-elixir.cerit-sc.cz/galaxy/

Results

Short read data: Application to *Coffea* genome size





https://www.wildcoffeedb.org

Results

Short read data: Application to *Coffea* genome size

RepeatExplorer2

34 sequencing archives with 100,00 random PE

_step1 In the case if your samples are clean and random: concatenate all samples into one archive (ALL.fasta) and upload it to Galaxy

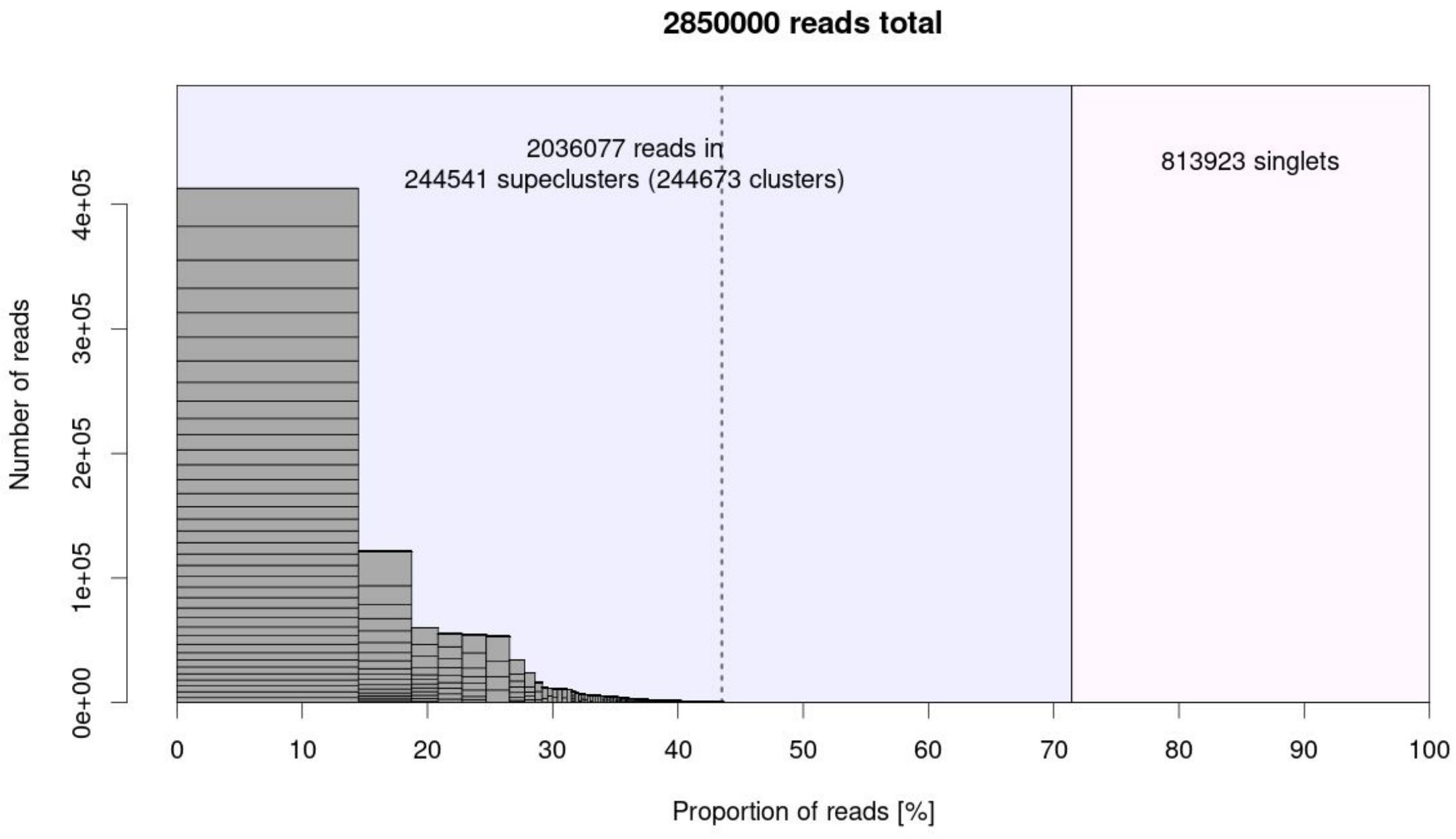
_step2 In ‘RepeatExplorer2 clustering’, select the ALL.fasta file, Viridiplantae 4.0 database of REXdb

_step3 Analyse Results (RepeatExplorer2 html report)

_step4 In RepeatExplorer utilities, use Visualization of comparative clustering. Load the results from step3: CLUSTER_TABLE.csv and COMPARATIVE_ANALYSIS_COUNTS.csv. You may normalize comparative analysis with the genome size of species (genome_sizes.txt)

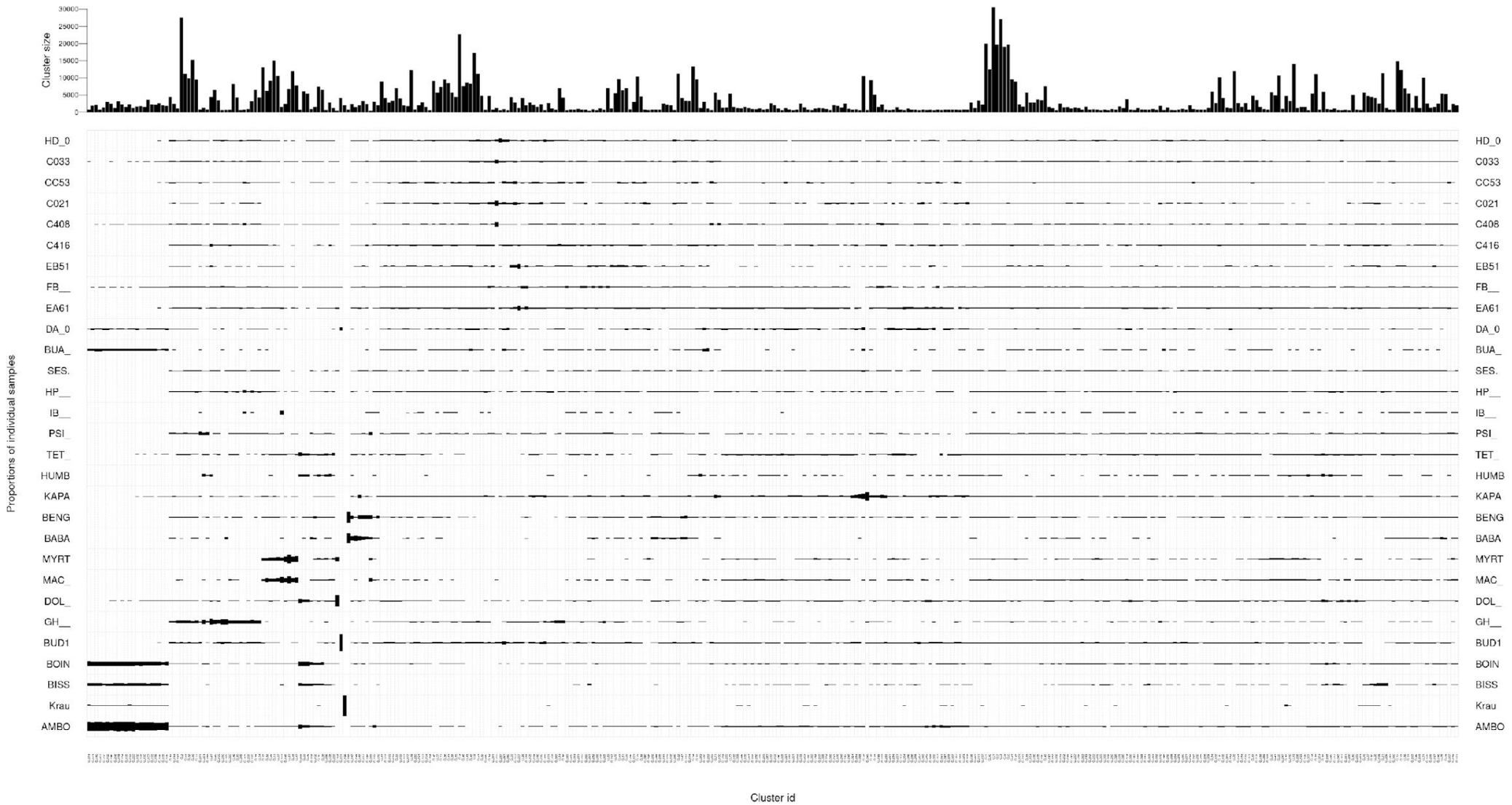
Results

Short read data: Application to *Coffea* genome size



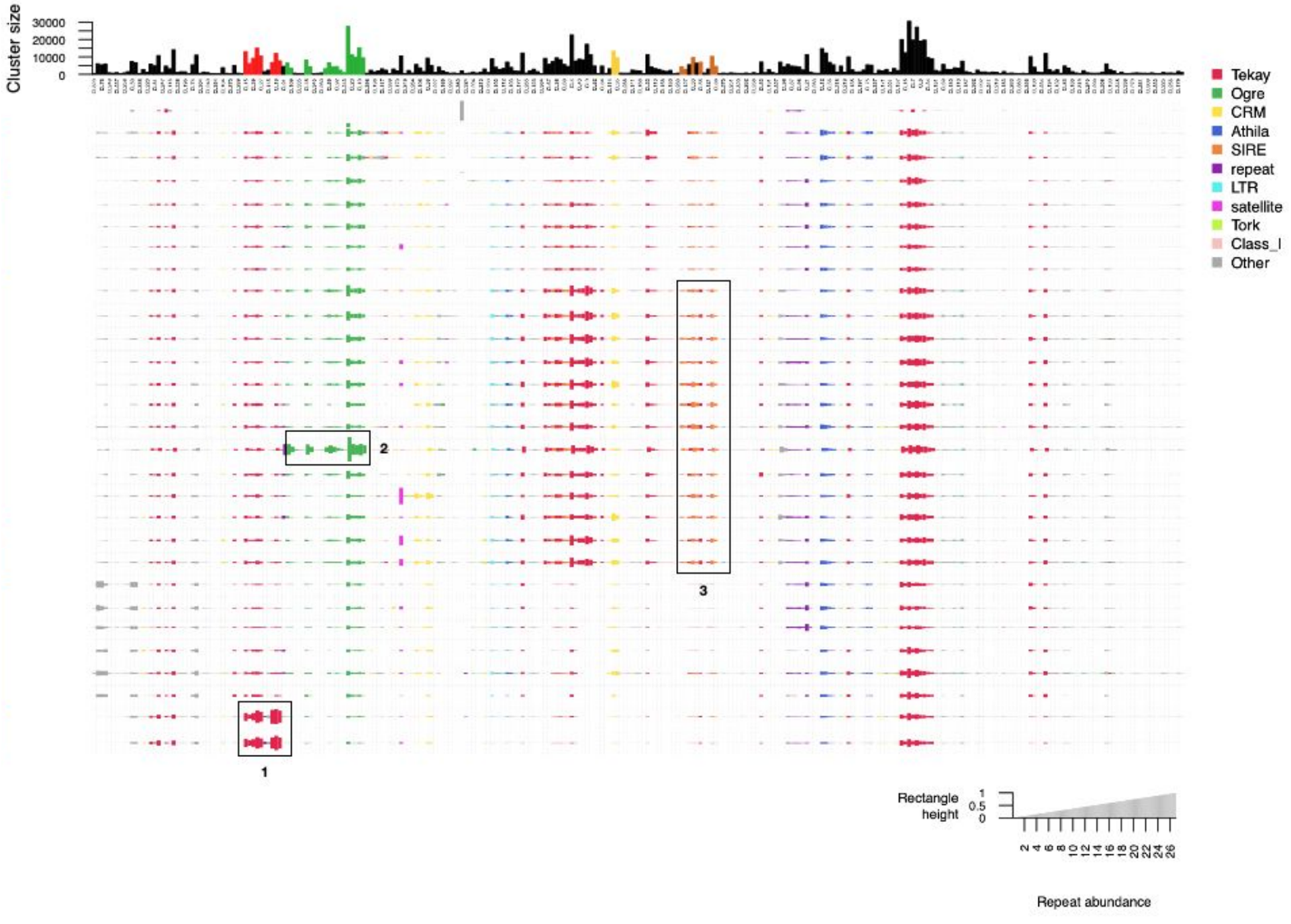
Results

Short read data: Application to *Coffea* genome size



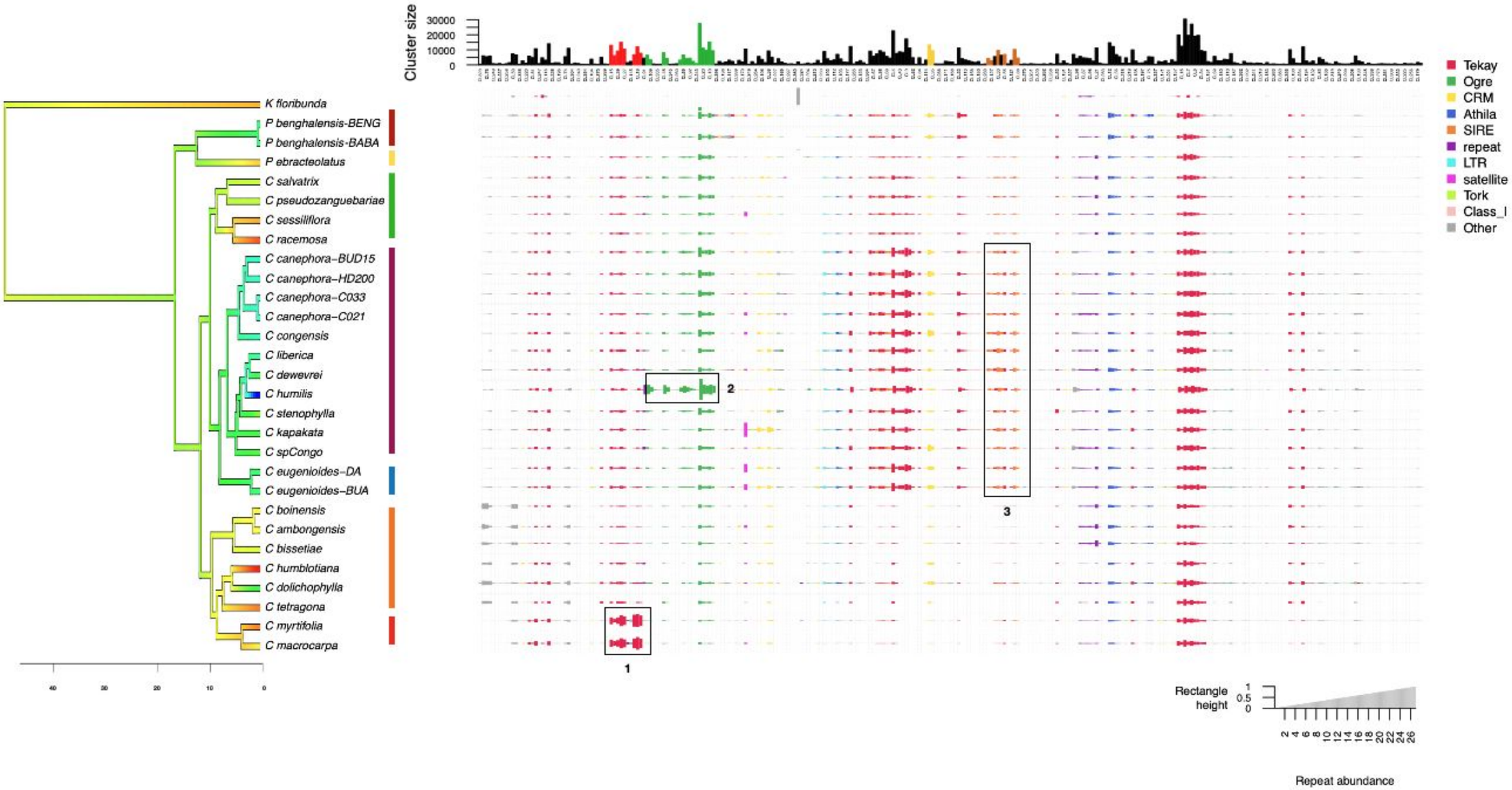
Results

Short read data: Application to *Coffea* genome size



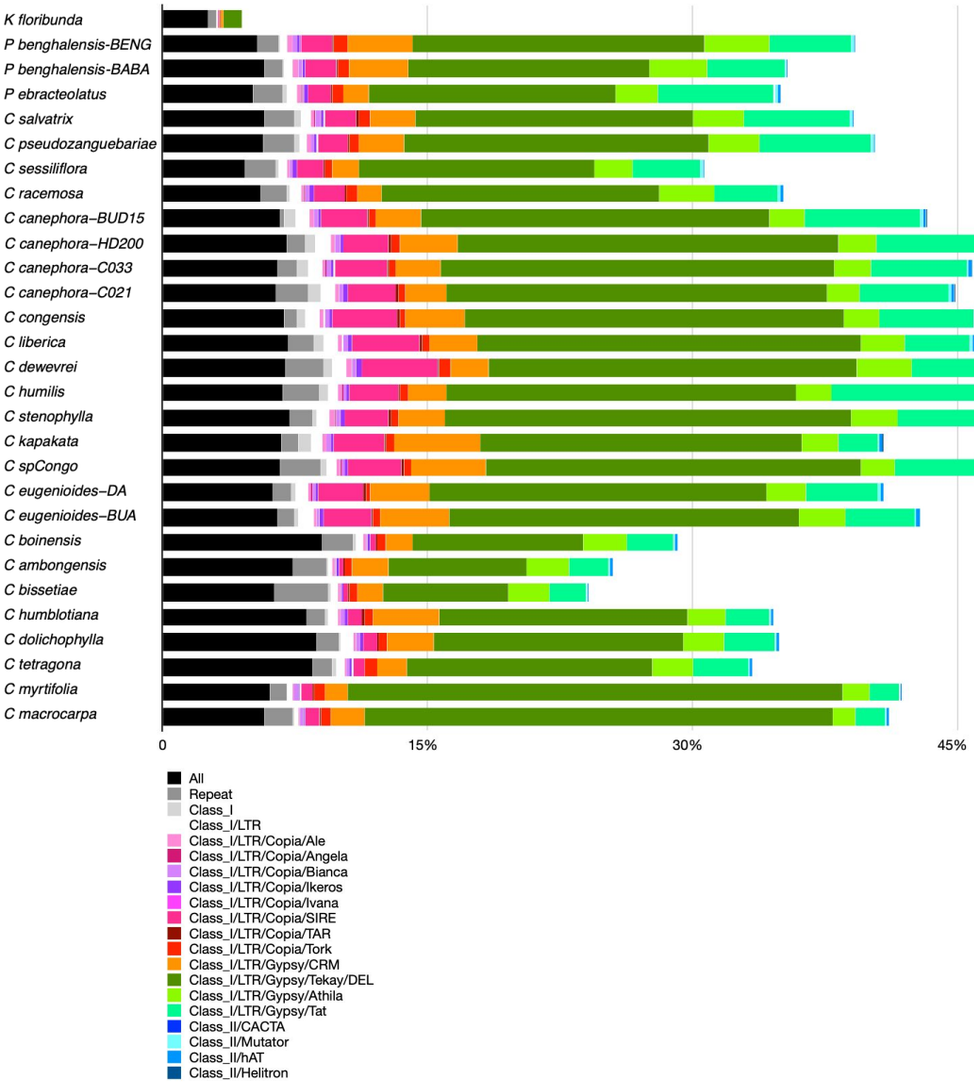
Results

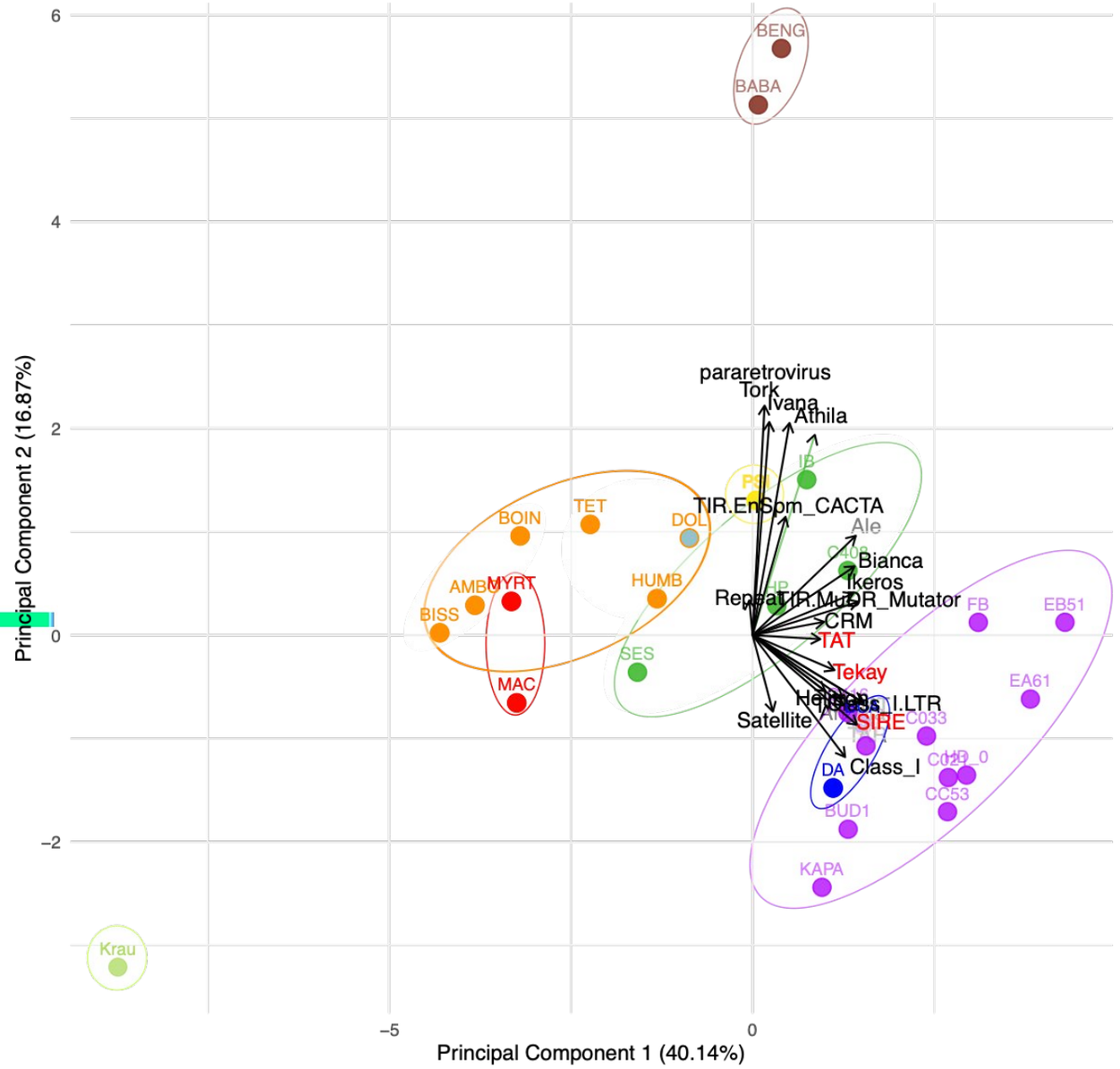
Short read data: Application to *Coffea* genome size



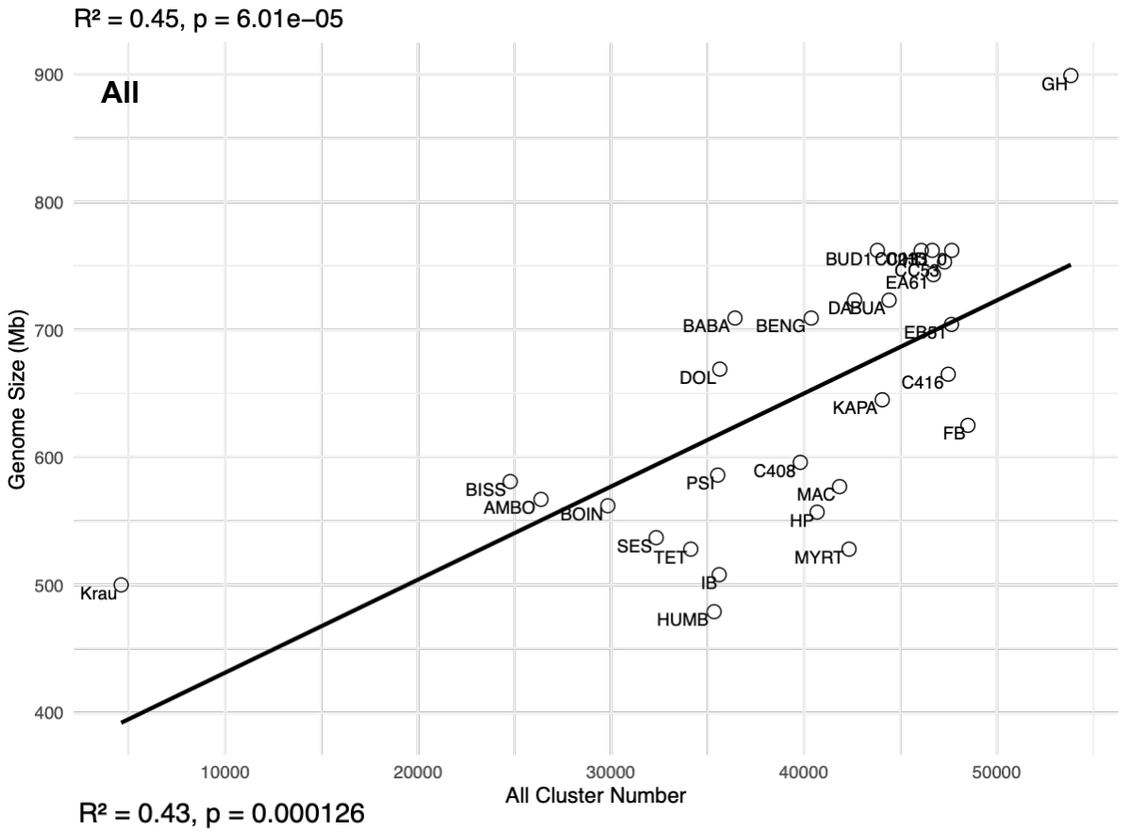
Results

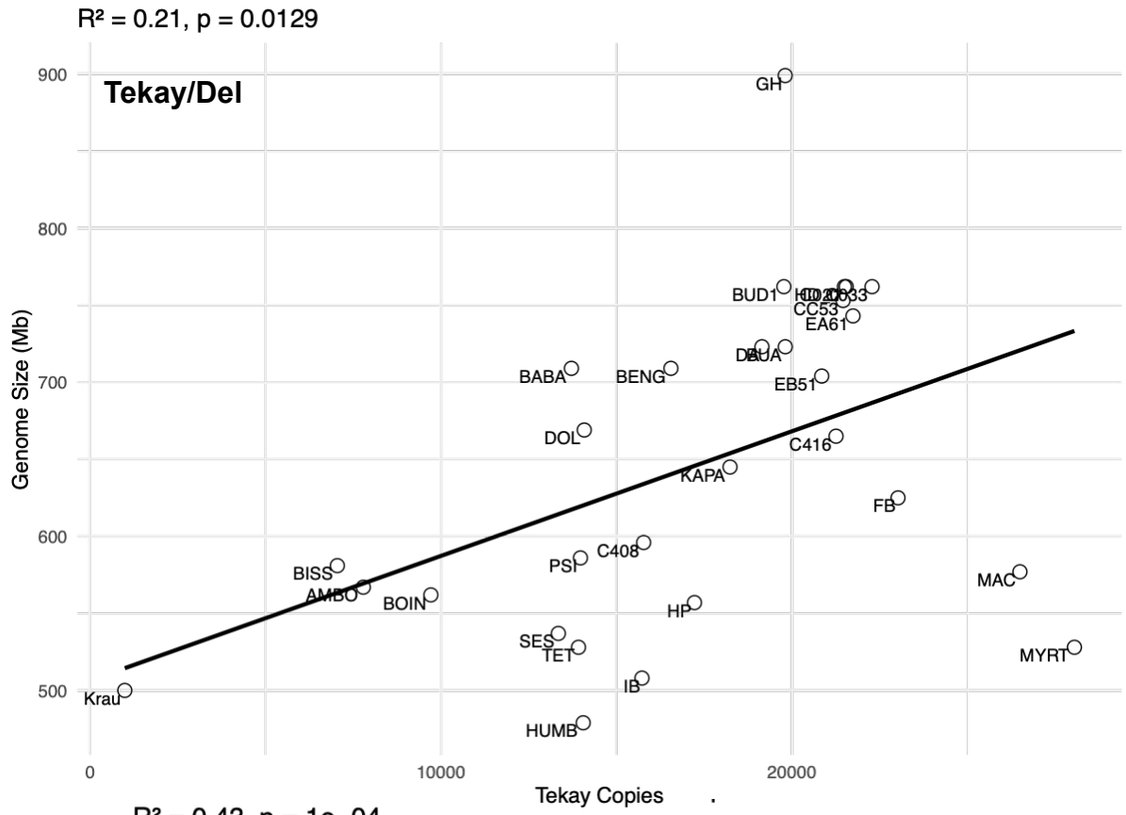
Short read data: Application to *Coffea* genome size

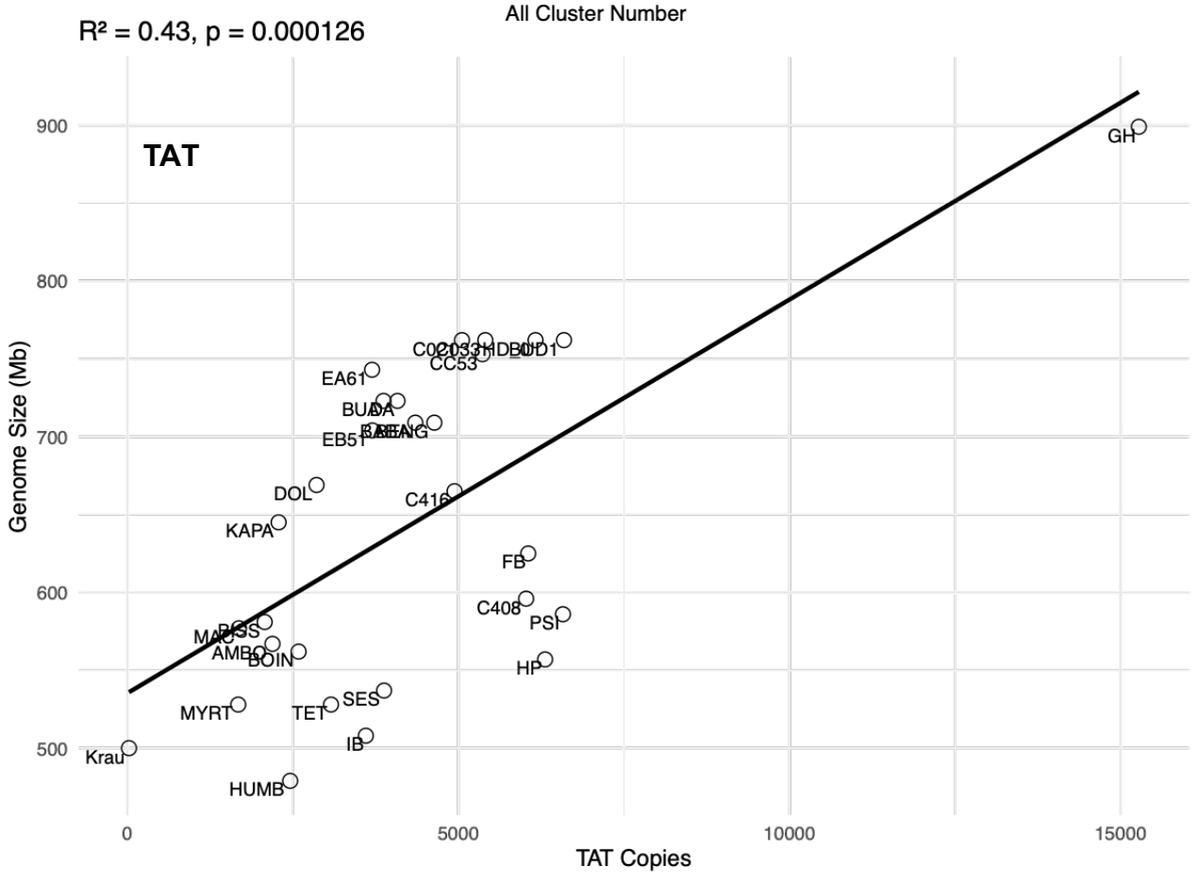


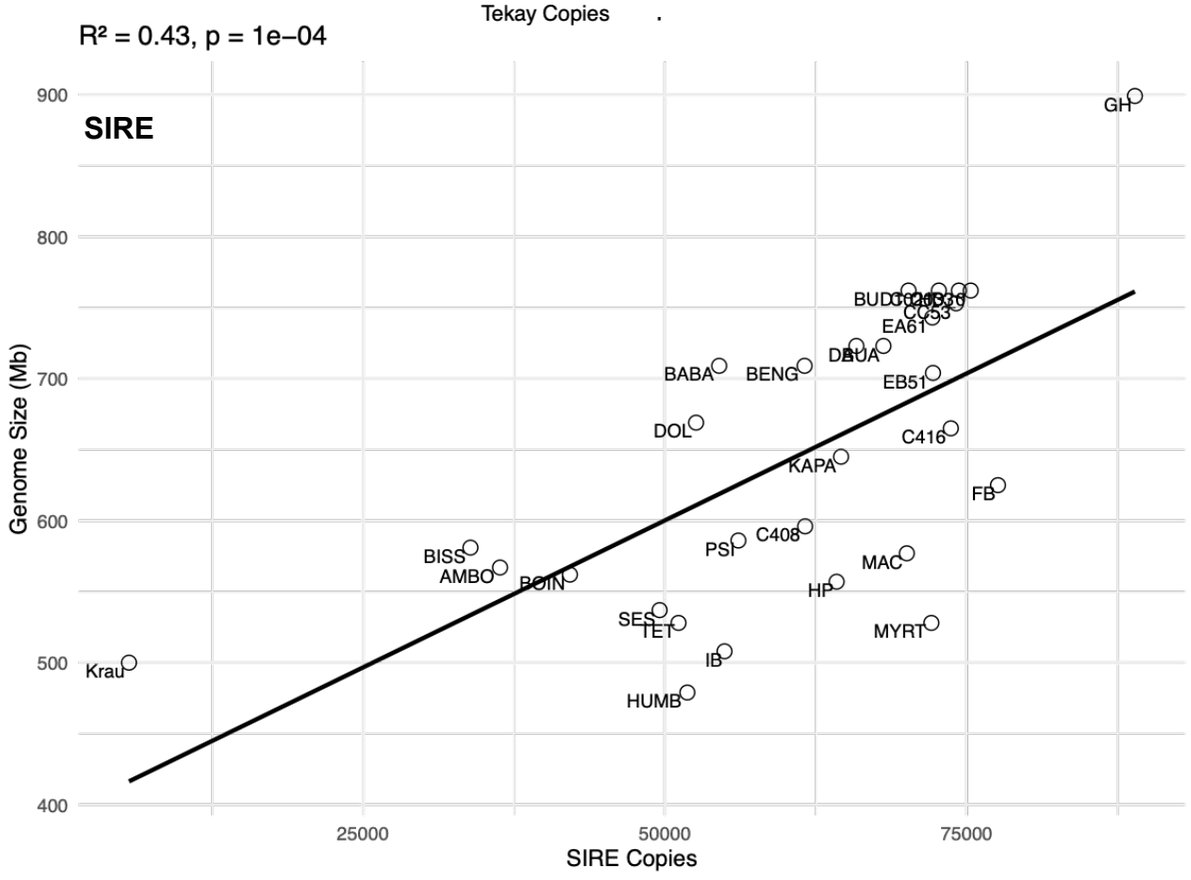


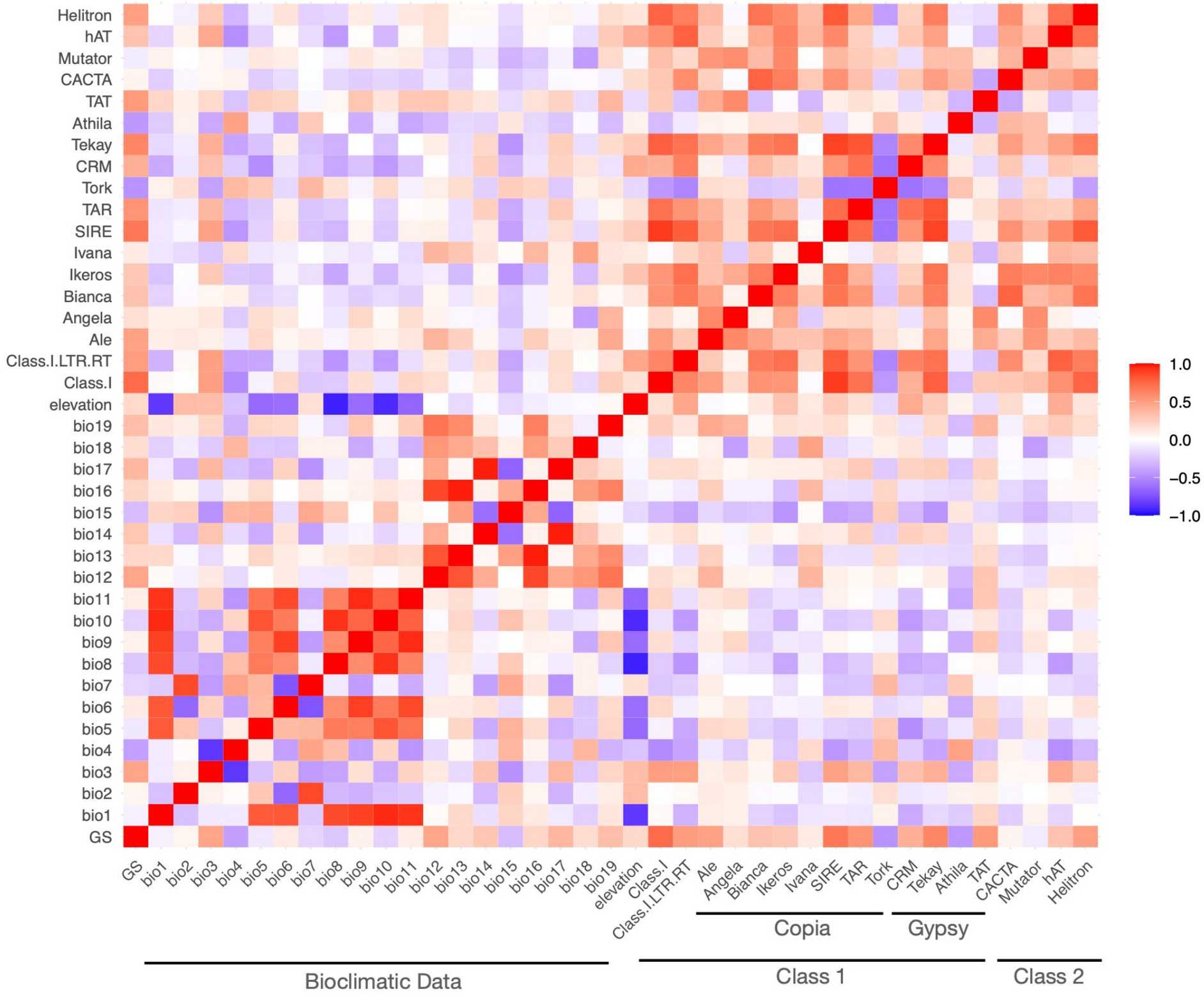
Results


R² = 0.45, p = 6.01e−05
All
Genome Size (Mb)
All Cluster Number
GH
BUD1
CC53
EA61
DABUA
BABA
BENG
EB51
DOL
C416
KAPA
FB
C408
PSI
MAC
HP
BISS
AMBO
BOIN
SES
TET
MYRT
IB
HUMB
Krau


R² = 0.21, p = 0.0129
Tekay/Del
Genome Size (Mb)
Tekay Copies
GH
BUD1
CC53
EA61
BABA
BENG
EB51
DOL
C416
KAPA
FB
C408
PSI
MAC
HP
BISS
AMBO
BOIN
SES
TET
MYRT
IB
HUMB
Krau


R² = 0.43, p = 0.000126
TAT
Genome Size (Mb)
TAT Copies
GH
CC53
EA61
DOL
C416
KAPA
FB
C408
PSI
HP
MYRT
TET
SES
IB
HUMB
Krau


R² = 0.43, p = 1e−04
SIRE
Genome Size (Mb)
SIRE Copies
GH
CC53
EA61
BABA
BENG
EB51
DOL
C416
KAPA
FB
C408
PSI
MAC
HP
BISS
AMBO
BOIN
SES
TET
MYRT
IB
HUMB
Krau


Helitron
hAT
Mutator
CACTA
TAT
Athila
Tekay
CRM
Tork
TAR
SIRE
Ivana
Ikeros
Bianca
Angela
Ale
Class.I.LTR.RT
Class.I
elevation
bio19
bio18
bio17
bio16
bio15
bio14
bio13
bio12
bio11
bio10
bio9
bio8
bio7
bio6
bio5
bio4
bio3
bio2
bio1
GS
1.0
0.5
0.0
−0.5
−1.0
Copia
Gypsy
Bioclimatic Data
Class 1
Class 2

TE detection and genotyping

After library building

Detection of individual TE insertions (de novo & reference insertions)