



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)

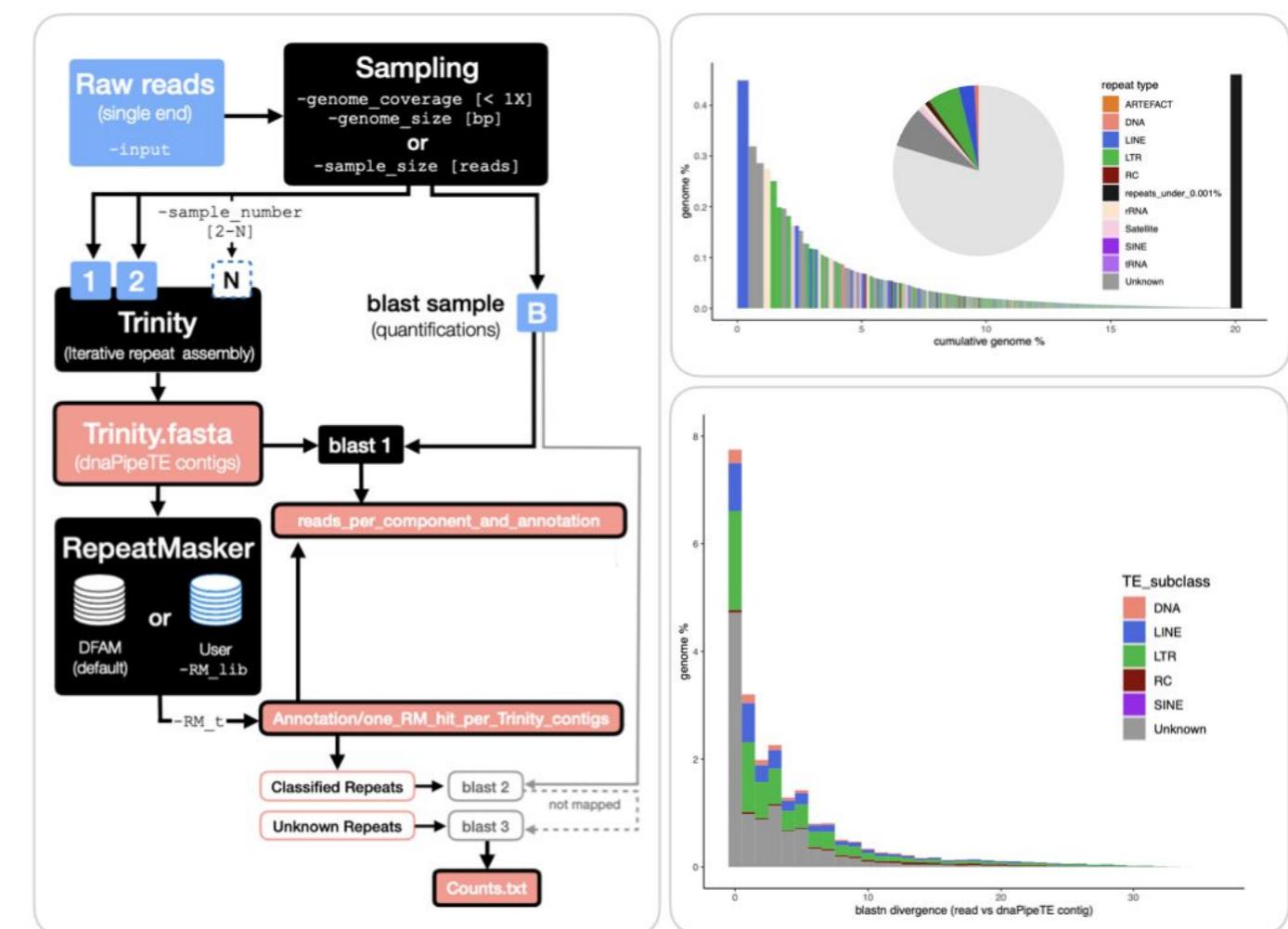
Introduction

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)



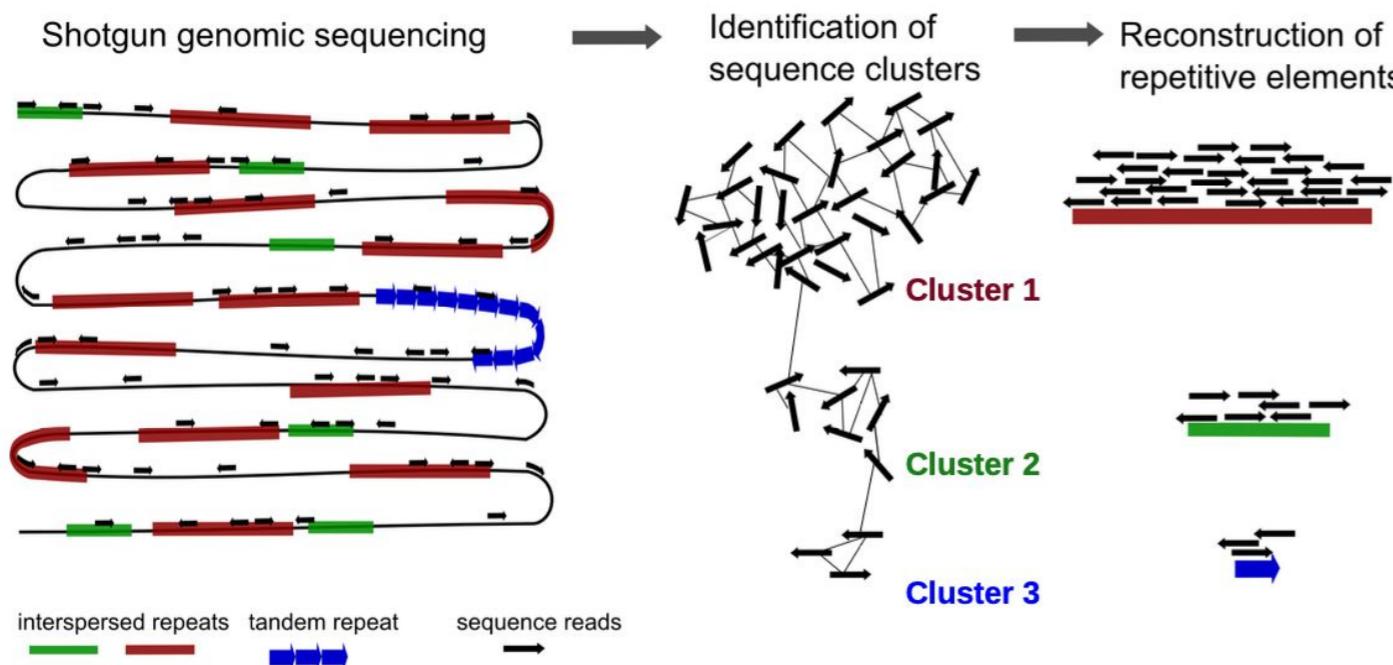
Introduction

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

Introduction

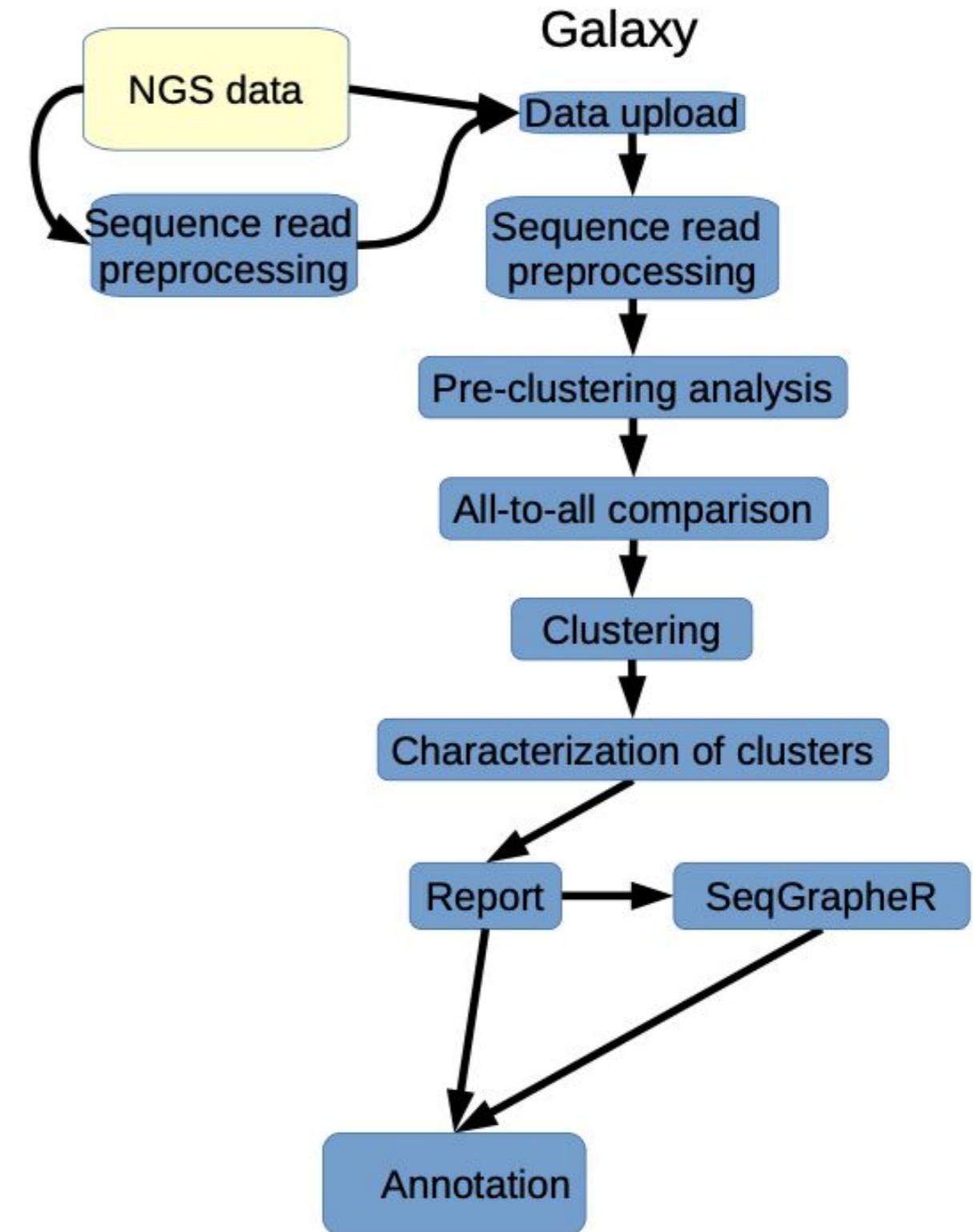
Short read data

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



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

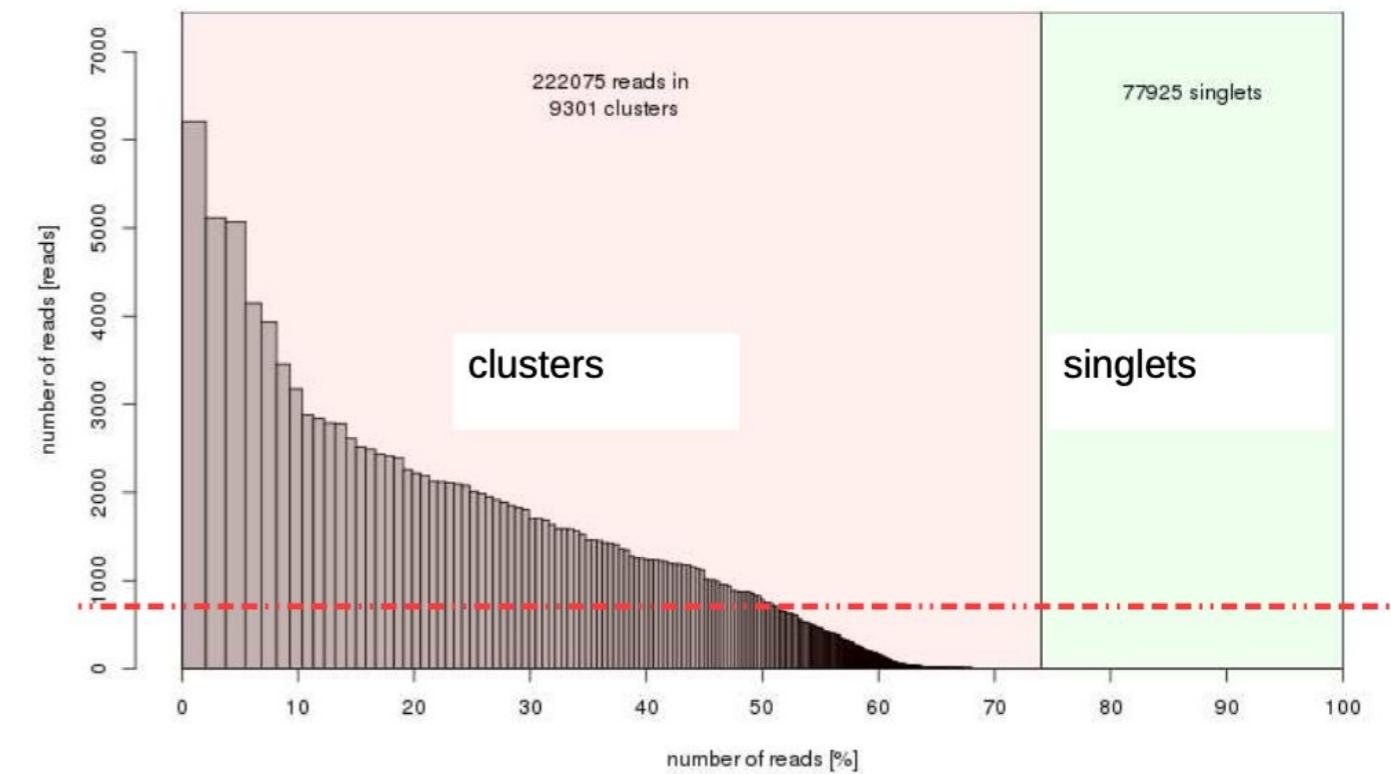
Sequence reads: 100 bp PE

100,000 to 10 M reads

55 bp overlap and 90% similarities

Characterization of cluster

RepeatMasker search (Repbase, custom db)



Typical cluster size distribution

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

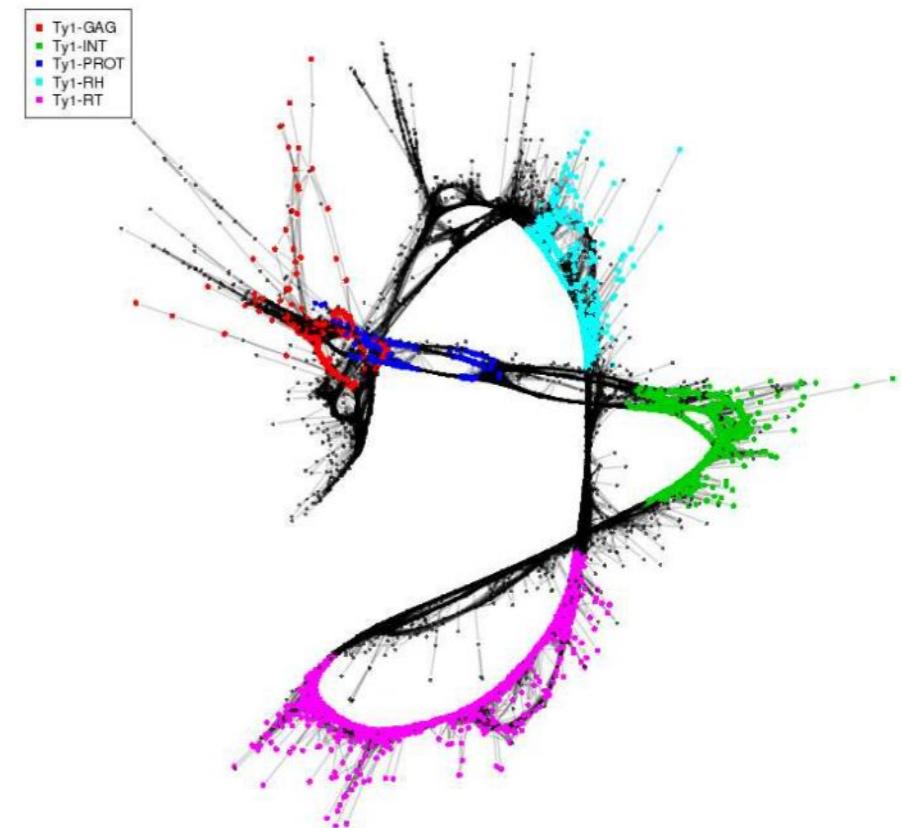
100,000 to 10 M reads

55 bp overlap and 90% similarities

Characterization of cluster

RepeatMasker search (Repbase, custom db)

Protein domain database + Graph layouts



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

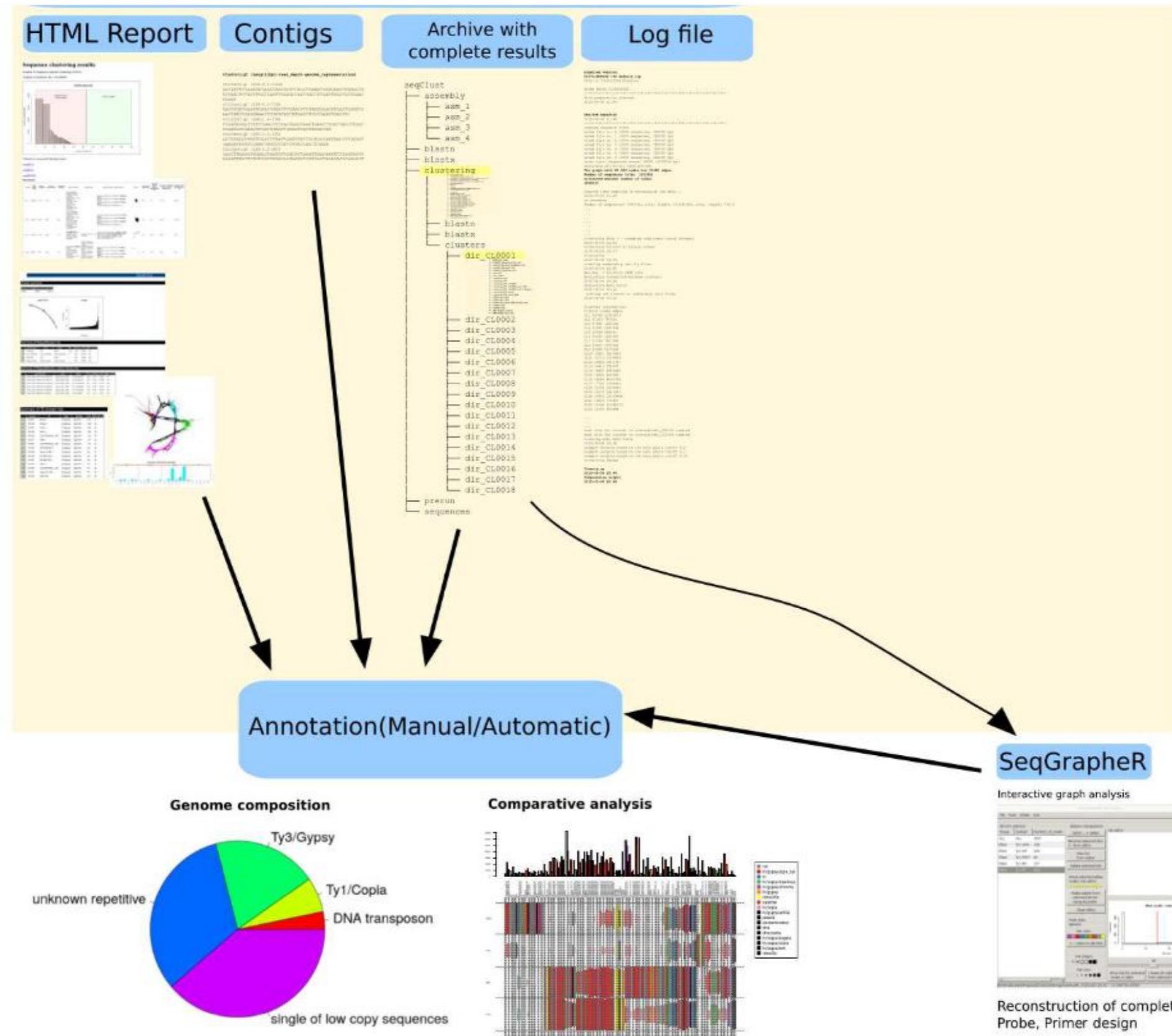
RepeatMasker search (Repbase, custom db)

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 (.....		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.EuSpn (1hits, 0.00119%) 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, 4.10%) SetD1(CL4Contig1816 (481hits, 3.73%) SetD1(CL4Contig3001 (201hits, 1.42%) SetF(CL8Contig4062 (144.....		35.9	13.05	0.074	0.378
4 CL4	869130	9657	2.040	19.2	LTR.Gypsy (1216hits, 9.51%) Low complexity (21hits, 0.6727%) 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 (350hits, 5.03%) gypsy_Ogre_TatA2(CL7Contig1310 (480hits, 4.46%) gypsy_Ogre_TatA3(CL19Contig1418 (401hits, 3.82%) gypsy_Ogre_TatA2(CL7Contig2590 (385h.....		29.1	15.15	0.015	0.072

Introduction



Introduction

Short read data

Galaxy

Workflow Visualize Données partagées Aide Utilisateur Using 19%

Tools search tools Upload Data

Get Data Send Data RepeatExplorer DANTE DANTE_LTR TideCluster Assembly annotation tools RepeatExplorer utilities NGS: QC and manipulation FASTA/FASTQ manipulation Genomic file utilities BED Text Manipulation Convert Formats Filter and Sort Join, Subtract and Group Operate on Genomic Intervals Statistics Collection Operations Jbrowse Visualisation Lift-Over Fetch Alignments/Sequences Graph/Display Data Text manipulation Phenotype Association Built-in Converters WORKFLOWS

RepeatExplorer

Discover repeats in your next generation sequencing data

Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre CAS, Ceske Budejovice, Czech Republic

This RepeatExplorer Galaxy portal is a part of services provided by ELIXIR (European Research Infrastructure for Biological Information). Please acknowledge this fact in your publications by adding a statement: "Computational resources for RepeatExplorer analysis were provided by the ELIXIR-CZ project (LM2023055), part of the international ELIXIR infrastructure."



12th Repeat Explorer Workshop, May 27–29, 2025

Join us at the 12th Repeat Explorer Workshop, May 27–29, 2025, in České Budějovice, Czech Republic. The workshop will focus on "Repetitive DNA Annotation in Genome Assemblies" and will include theoretical sessions, hands-on training, and a mini-conference for presenting research on repeat analysis in plant and animal genomes. For more detailed information, please visit our workshop page: [Workshop Details](#).

Contact information

If you need help, need to increase a data quota, or want to report a problem, please contact server administrator. If you encounter an error while running a tool, please report a bug using the bug icon in the dataset history.

Resources

- Step-by-step protocols include four protocols on how to use main RepeatExplorer tools
- RepeatExplorer channel on YouTube
- Documentation and training information
- The impact of genome coverage and sequence read sampling on reproducibility of repeat identification
- Official Galaxy Project website with information on how to use the Galaxy platform
- For the command line version of the RepeatExplorer tools, see the source code repository

REXdb

RepeatExplorer2, DANTE and DANTE_LTR tools use the REXdb database. REXdb is a comprehensive database of conserved protein domain sequences extracted from all types transposable elements found in plants. More information about REXdb can be found in related publication or in REXdb repository.

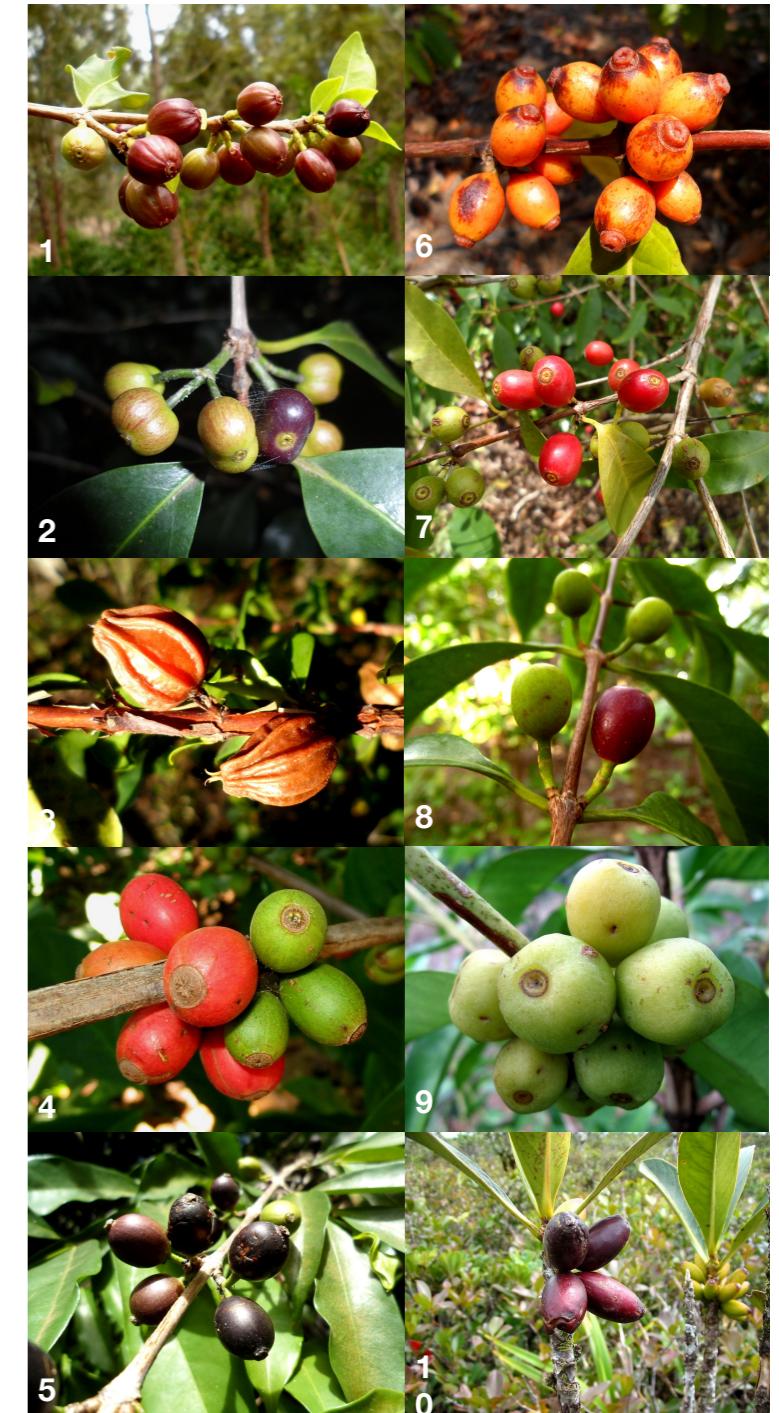
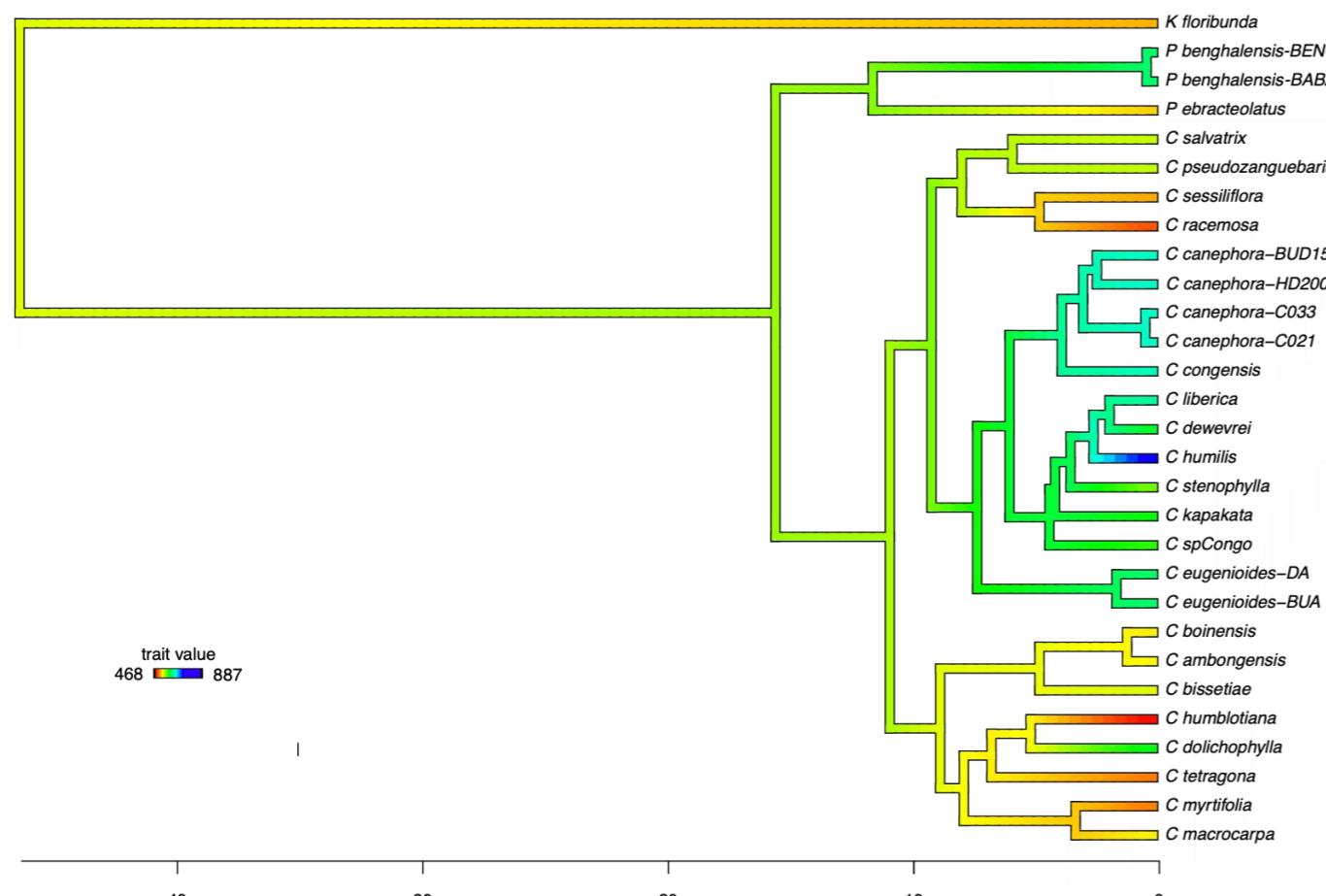
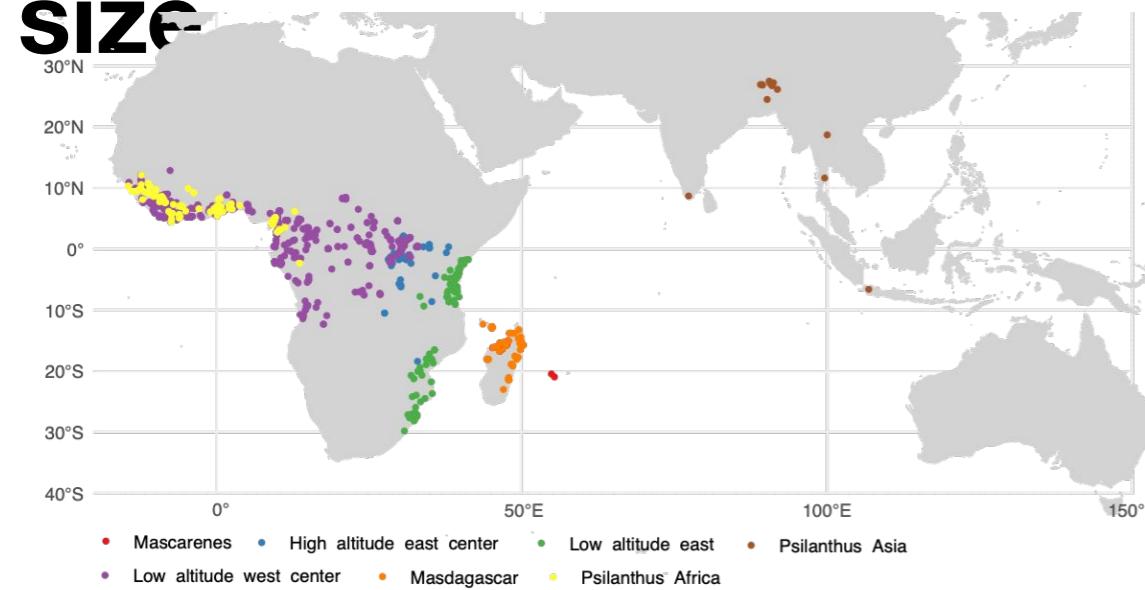
REXdb versions included in the server:

History Rechercher des données COFFEA GS 38.8 GB 52 75

127 : Comparative analysis s summary
126 : genome_sizes.txt
125 : Comparative analysis s summary
124 : COMPARATIVE_ANALYS IS_COUNTS.csv from 120
123 : SUPERCLUSTER_TABLE E.csv from 120
122 : CLUSTER_TABLE.csv fro m 120
121 : RepeatExplorer2 - HTM L report from data 118
120 : RepeatExplorer2 - Arch ived with HTML report from d ata 118
119 : RepeatExplorer2 - log fi le
118 : ALL2.fasta
117 : Comparative analysis su mmary
116 : genome_sizes.txt
115 : Krona on data 114: HTM L
114 : RepeatExplorer cluster annotation formatted for Kro na visualization from data 11

Results

Short read data: Application to *Coffea* genome size



<https://www.wildcoffee.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

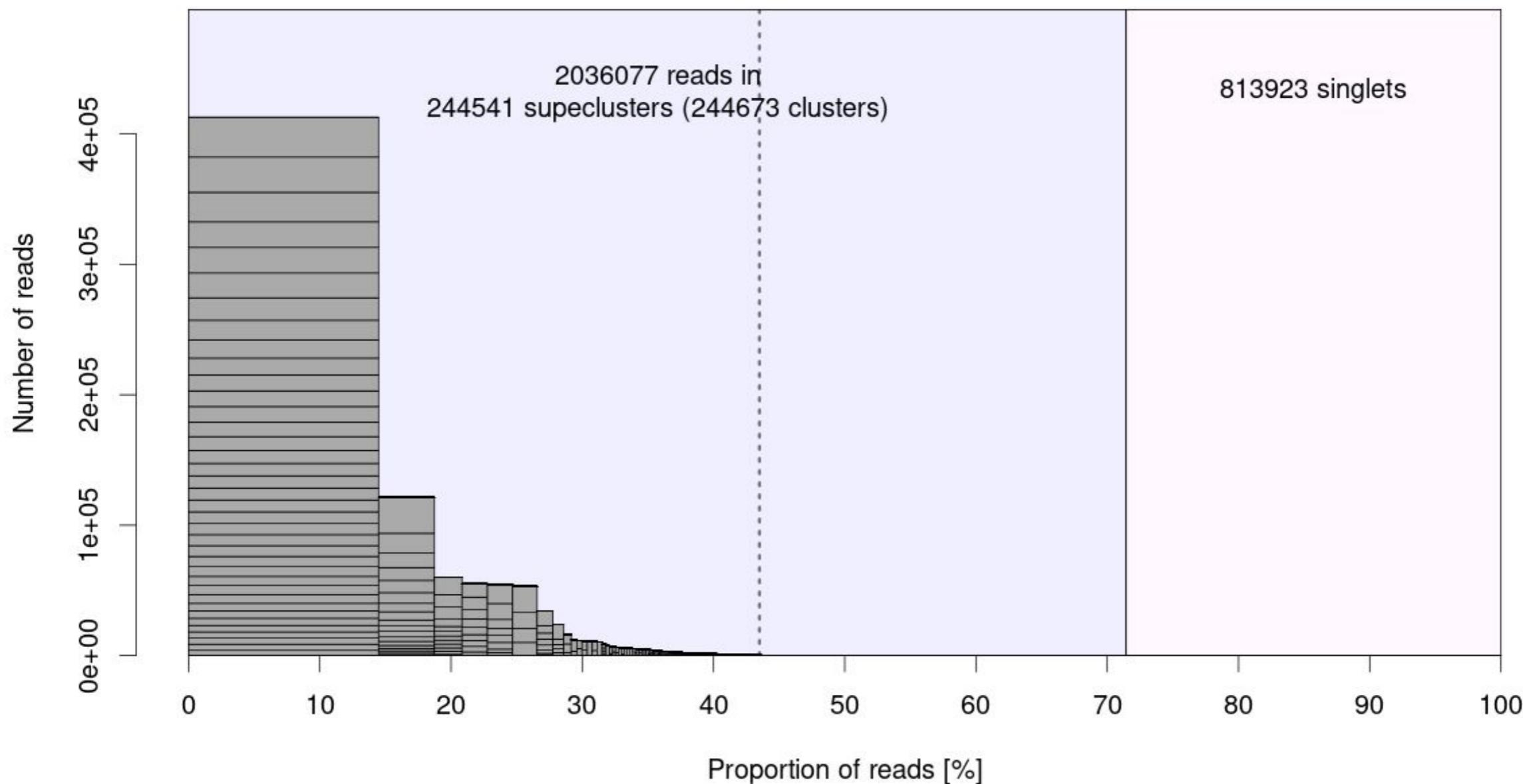
_step3 Analyse Results (RepeatExplorer2 html report)

_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

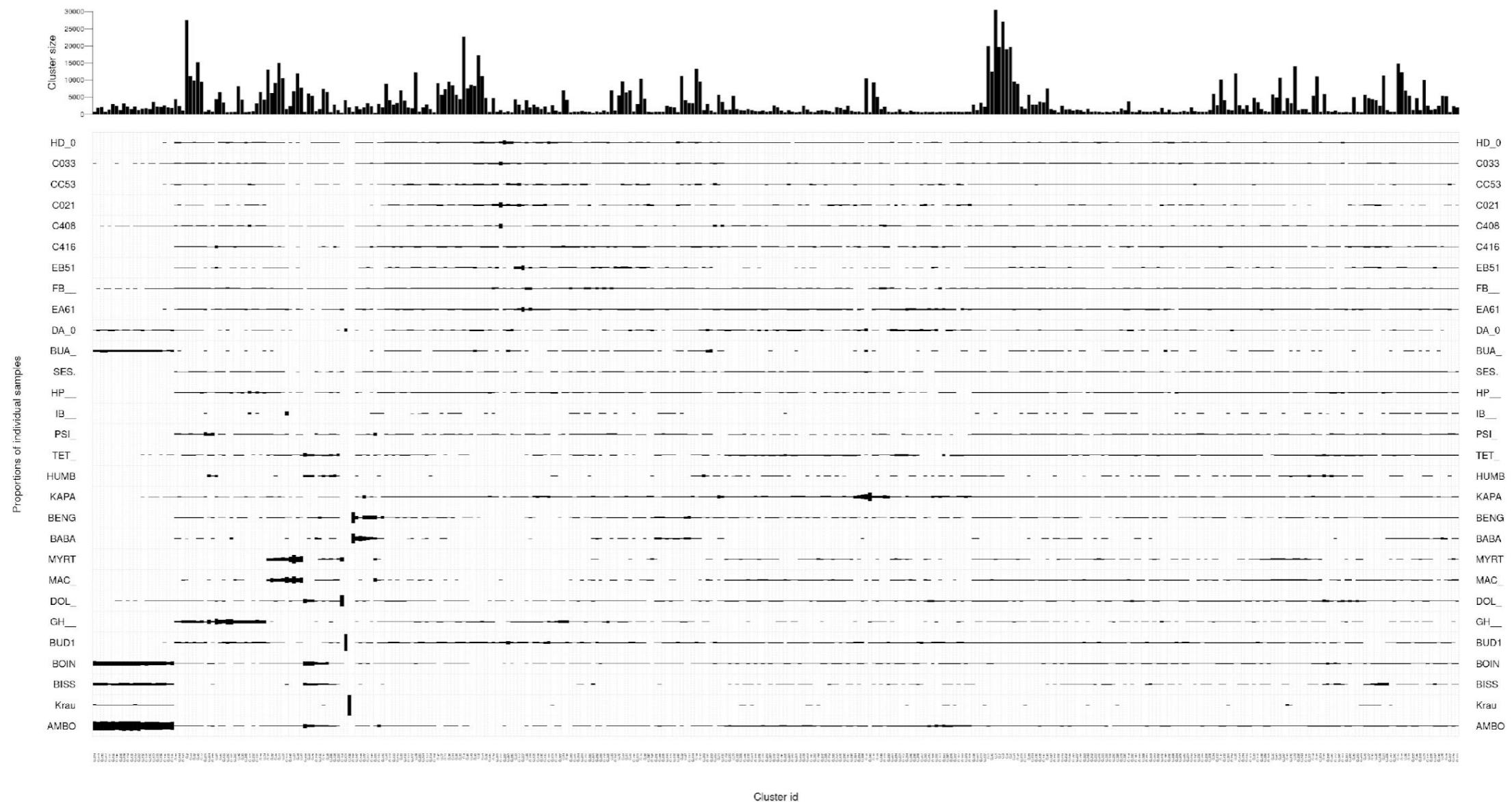
Short read data: Application to *Coffea* genome size

2850000 reads total



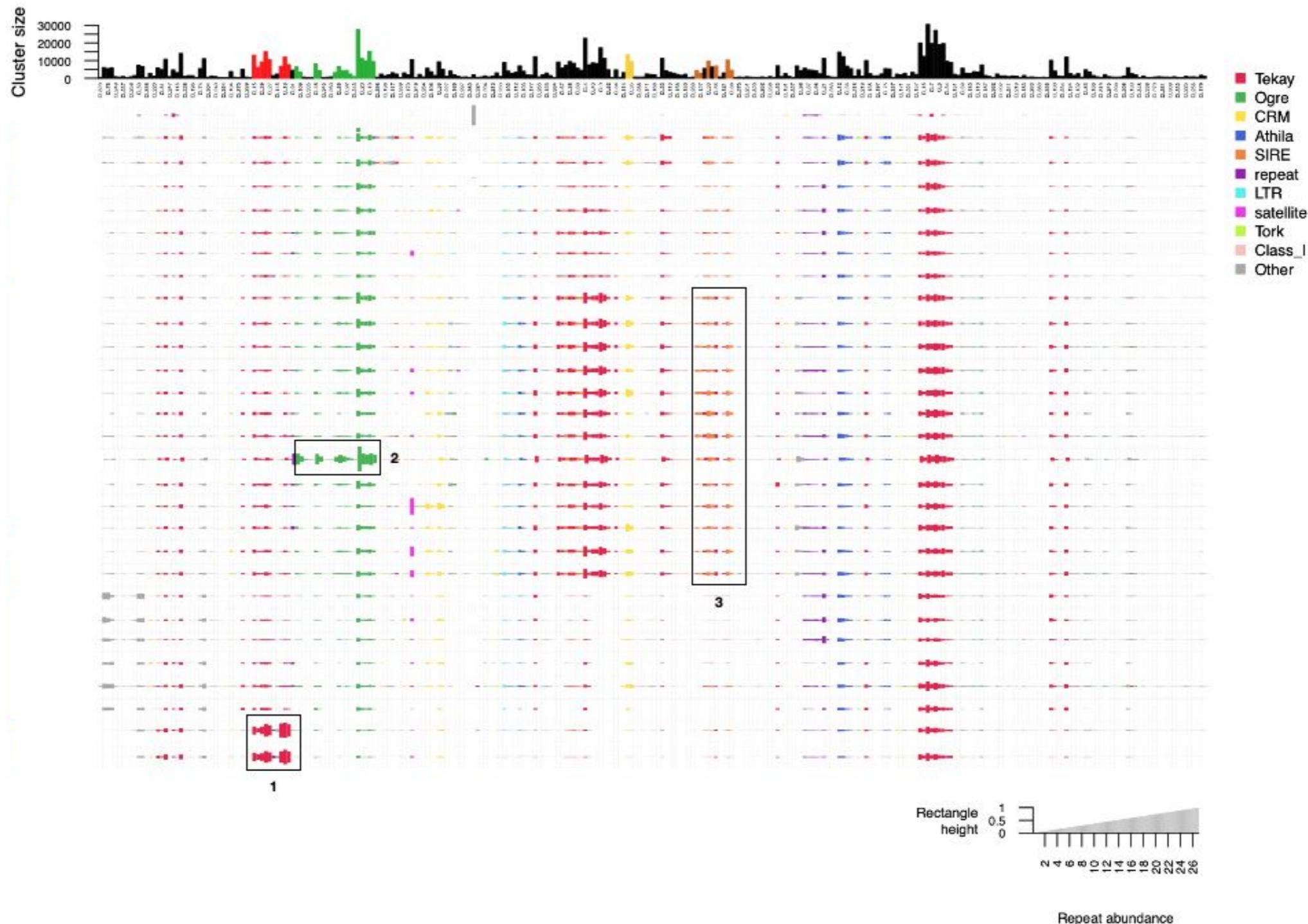
Results

Short read data: Application to *Coffea* genome size



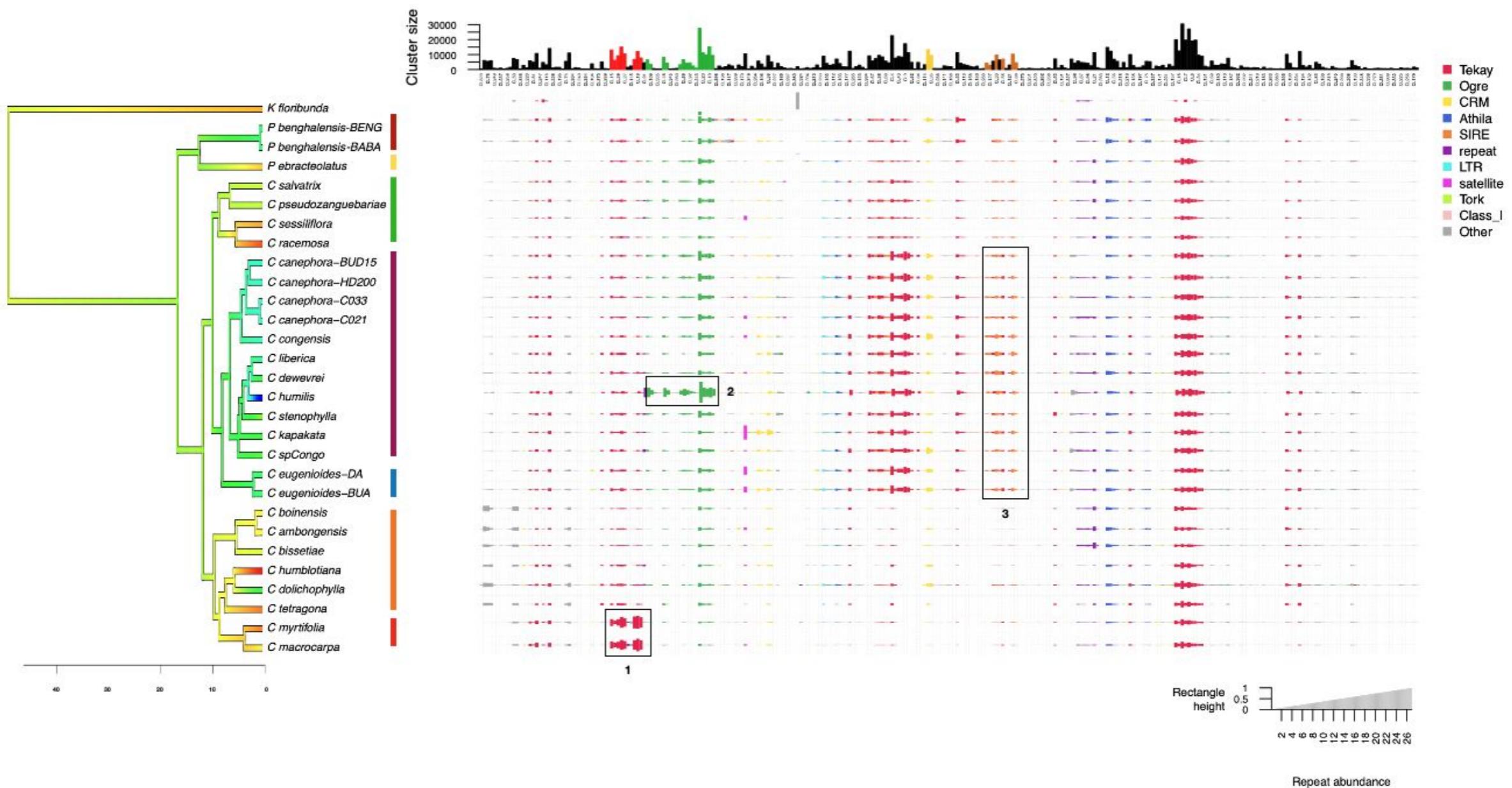
Results

Short read data: Application to *Coffea* genome size

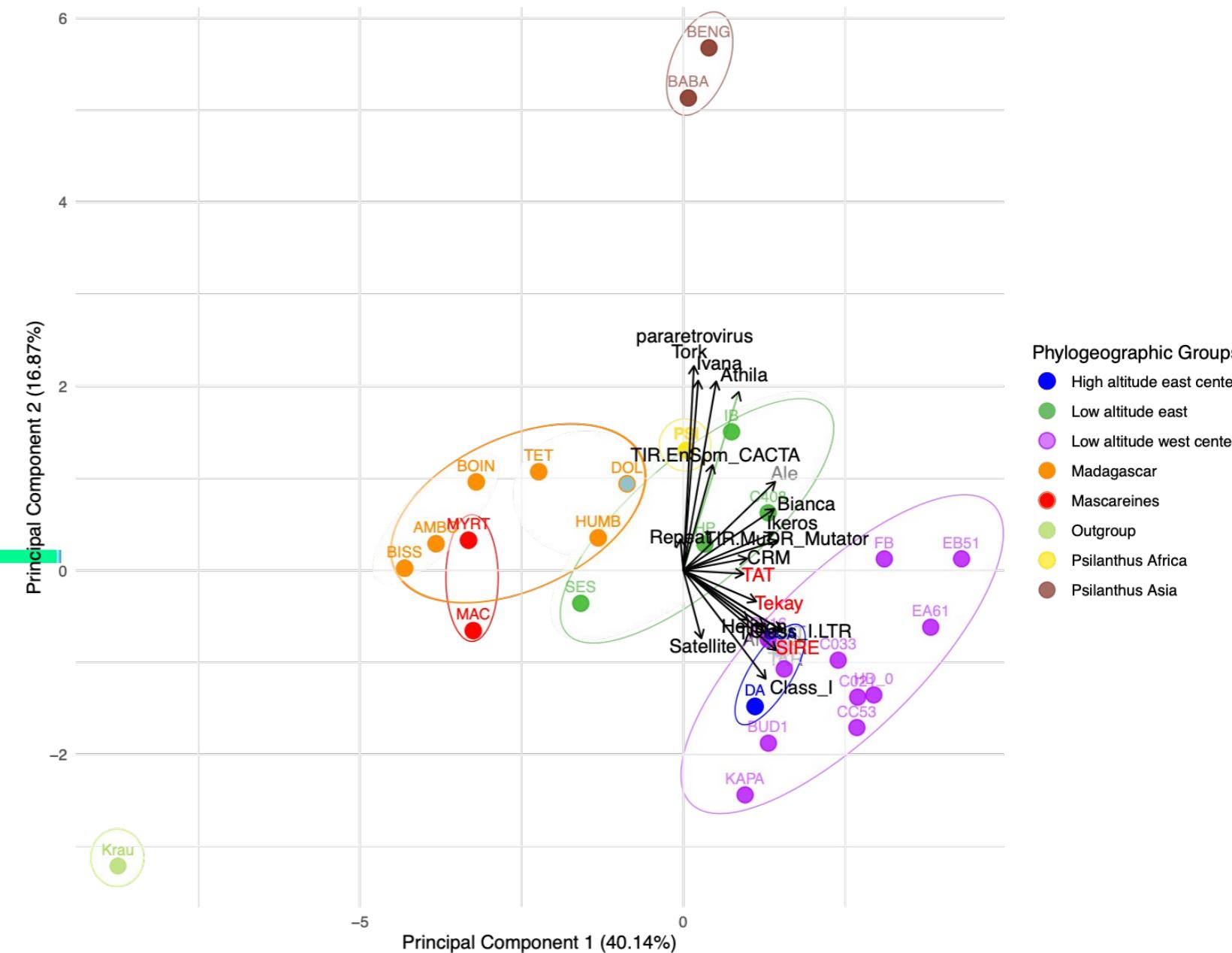
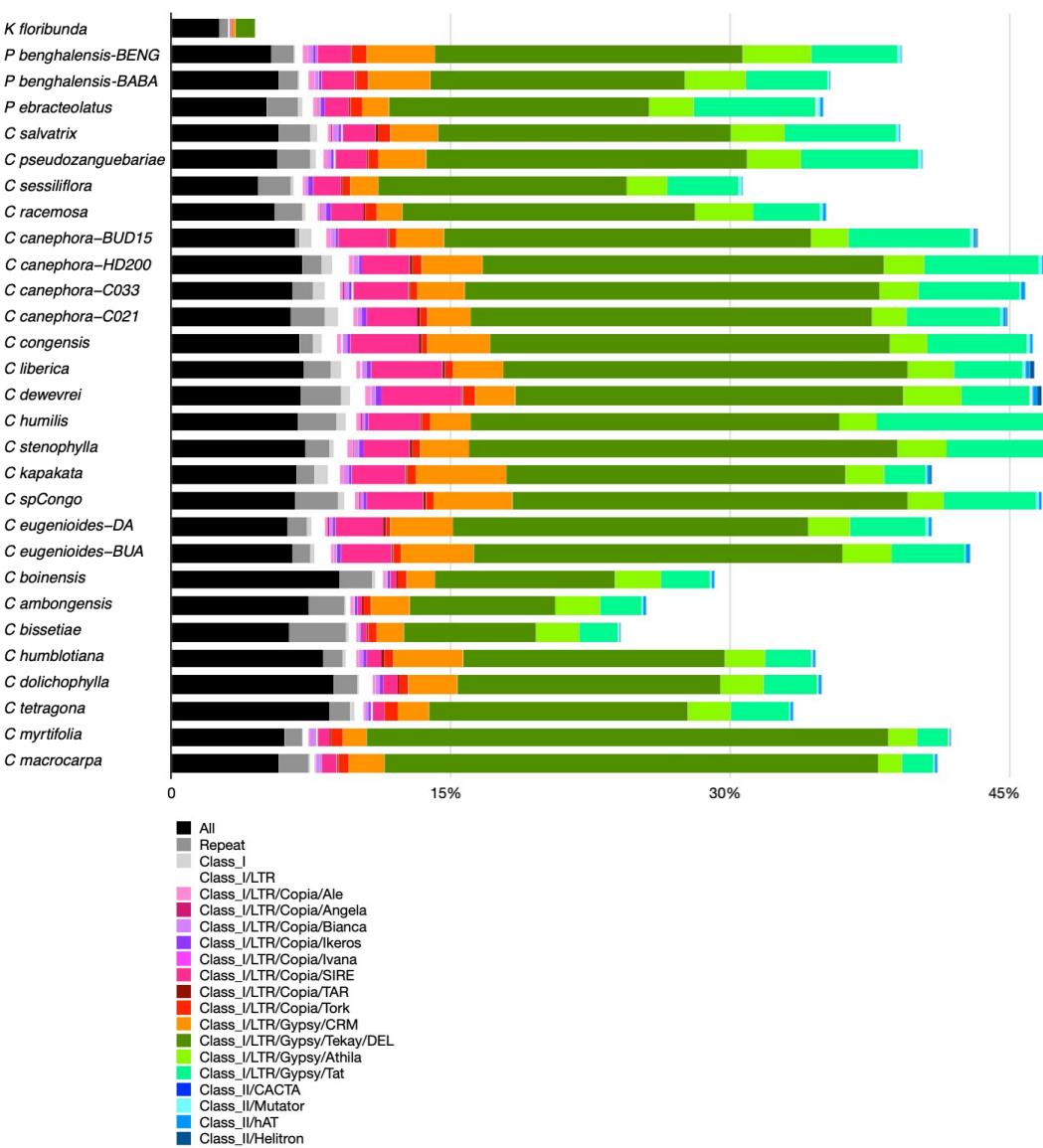


Results

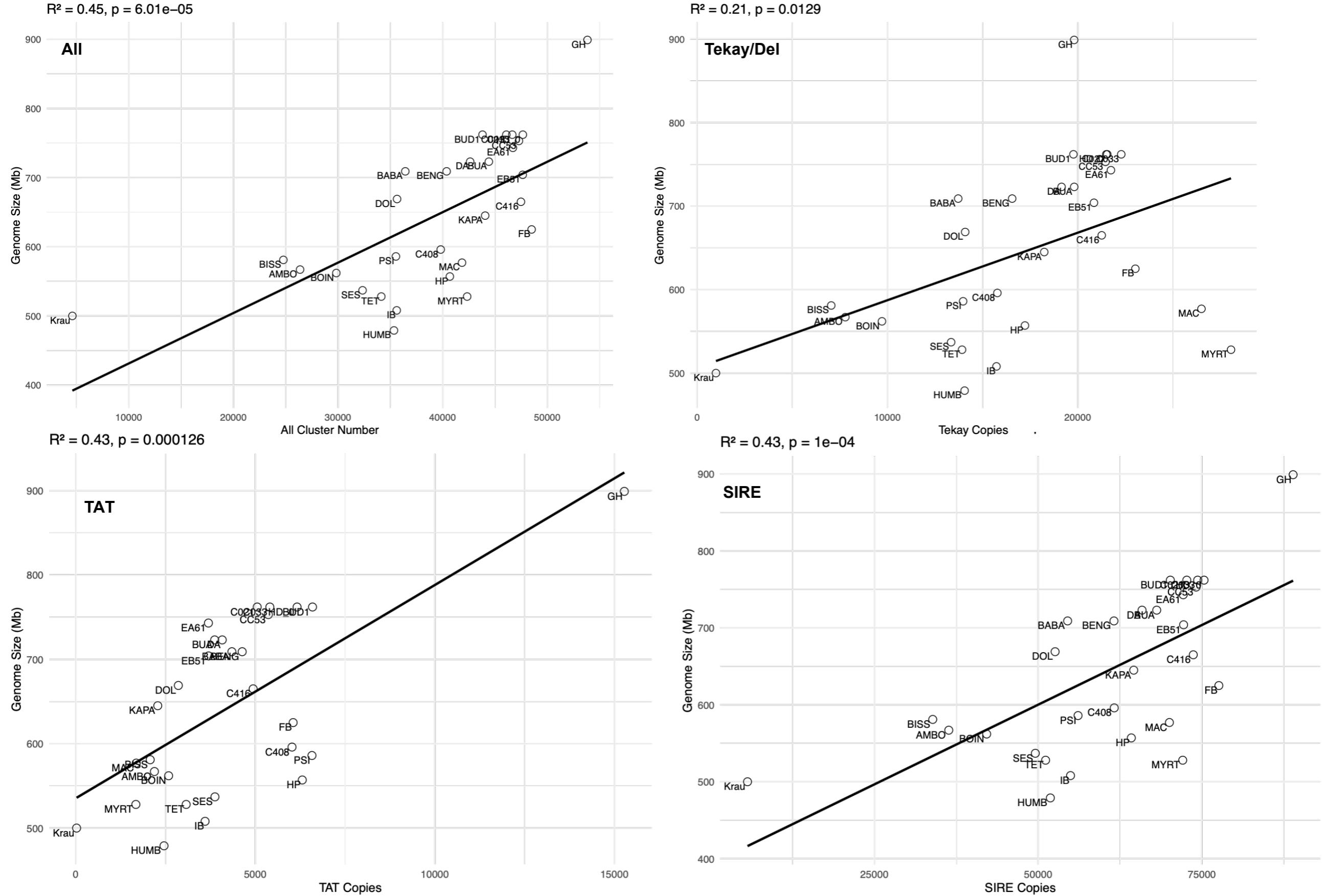
Short read data: Application to *Coffea* genome size



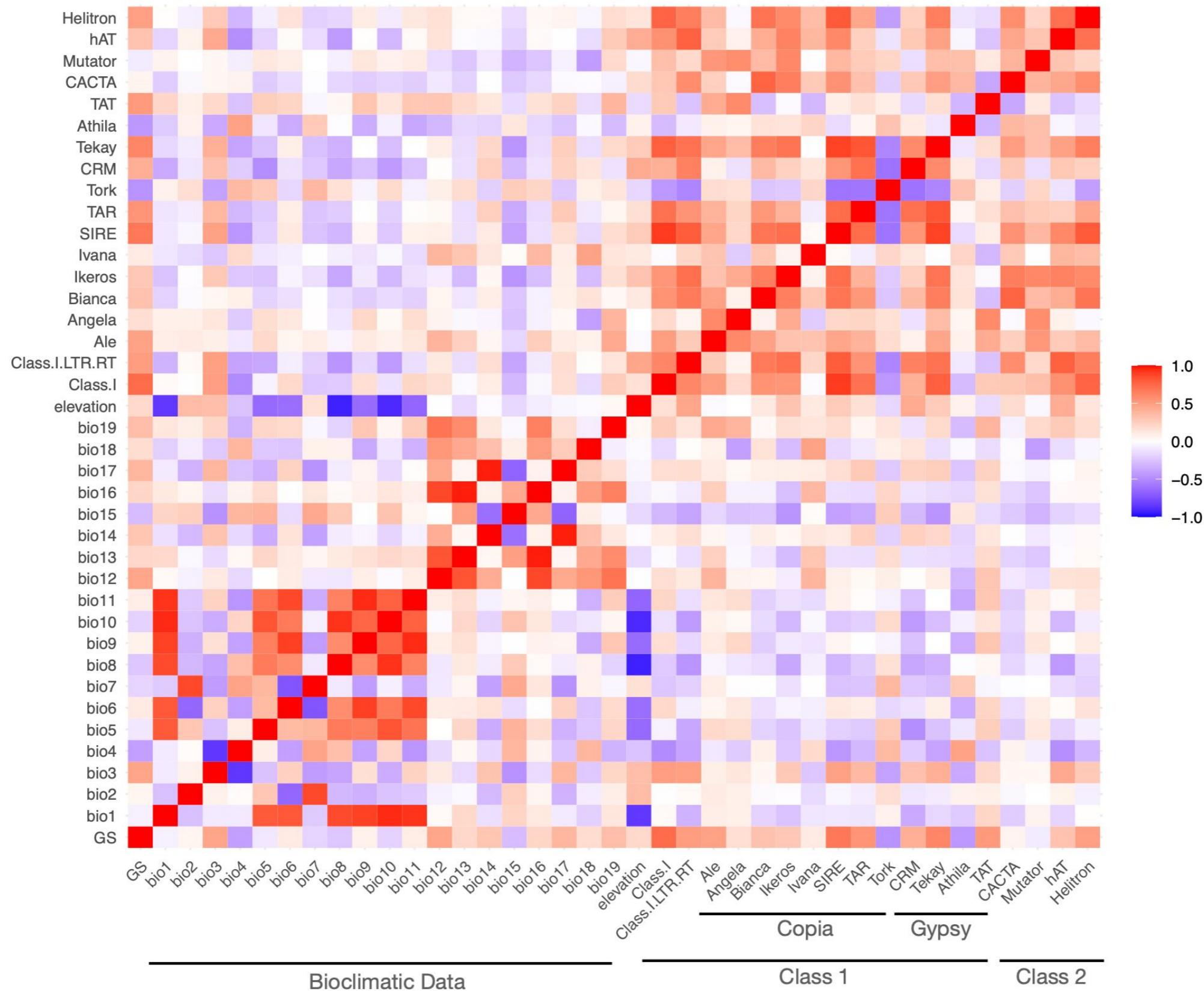
Short read data: Application to *Coffea* genome size



EMBO
Results



Results



TE detection and genotyping

After library building

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