ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1699694
ToLID	**ihLyrPleb1**
Species	Lyristes plebejus
Class	Insecta
Order	Hemiptera

Genome Traits	Expected	Observed
Haploid size (bp)	4,208,280,295	4,458,562,265
Haploid Number	9 (source: ancestor)	10
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q65

The following metrics were automatically flagged as below EBP recommended standards or different from expected:

. Observed Haploid Number is different from Expected

. Kmer completeness value is less than 90 for collapsed

Curator notes

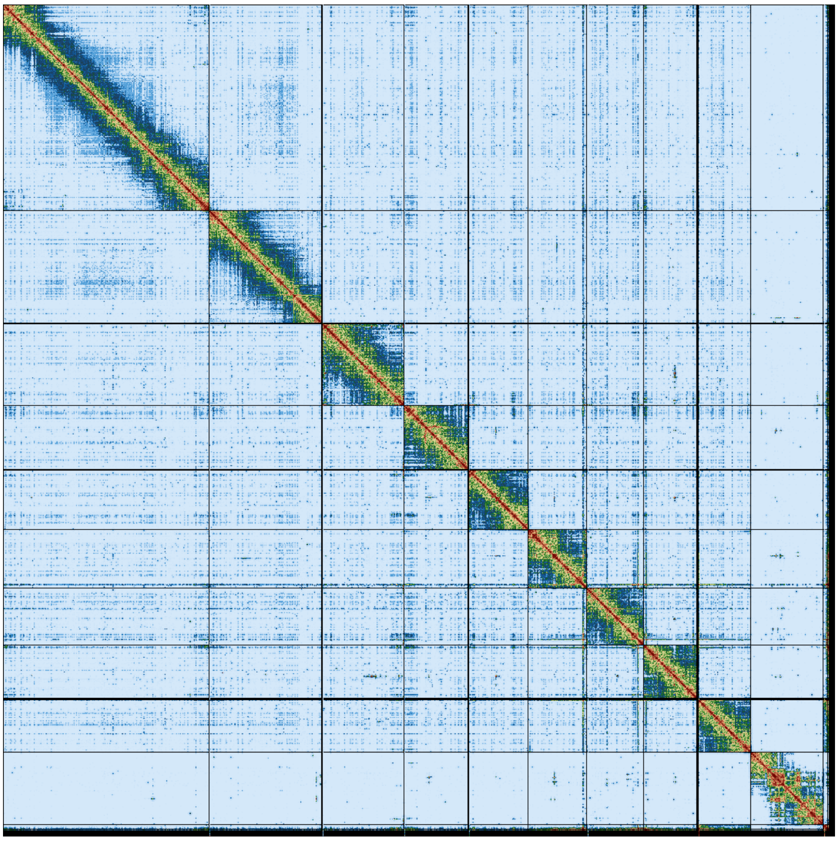
. Interventions/Gb: 7
. Contamination notes: ""
. Other observations: "The assembly of Lyristes plebejus (ihLyrPleb1) is based on 32X PacBio data and 307X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe project (BGE, https://biodiversitygenomics.eu/).The assembly process included the following steps: initial PacBio assembly generation with Hifiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 23 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 15 Mb (with the largest being 13.7 Mb). Additionally, 1273 regions totaling 288 Mb (with the largest being 7.5 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using OATK (linear sequence). Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 1 haplotypic region was removed, totaling 7.4 Mb. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	4,466,092,279	4,458,562,265
GC %	33.89	33.89
Gaps/Gbp	43.89	48.22
Total gap bp	19,600	23,700
Scaffolds	165	146
Scaffold N50	349,842,518	388,221,567
Scaffold L50	4	4
Scaffold L90	10	9
Contigs	361	361
Contig N50	35,864,000	35,864,000
Contig L50	33	33
Contig L90	120	120
QV	65.3812	65.3739
Kmer compl.	84.3792	84.3086
BUSCO sing.	96.2%	96.5%
BUSCO dupl.	1.8%	1.5%
BUSCO frag.	0.8%	0.7%
BUSCO miss.	1.2%	1.3%

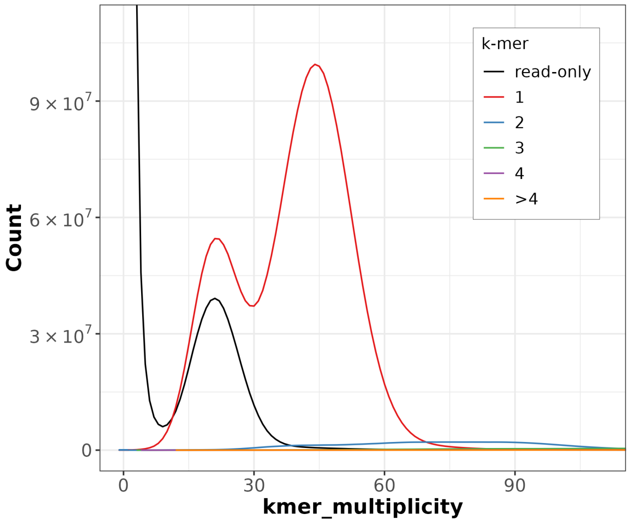
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: hemiptera_odb12 (genomes:32, BUSCOs:3396)

HiC contact map of curated assembly

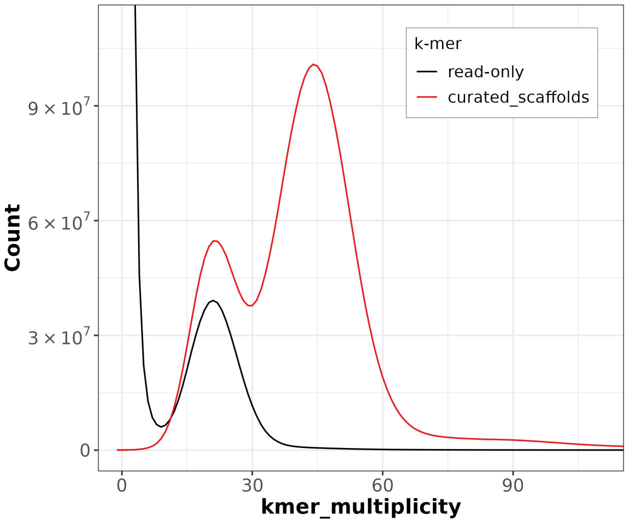


collapsed [LINK]

K-mer spectra of curated assembly

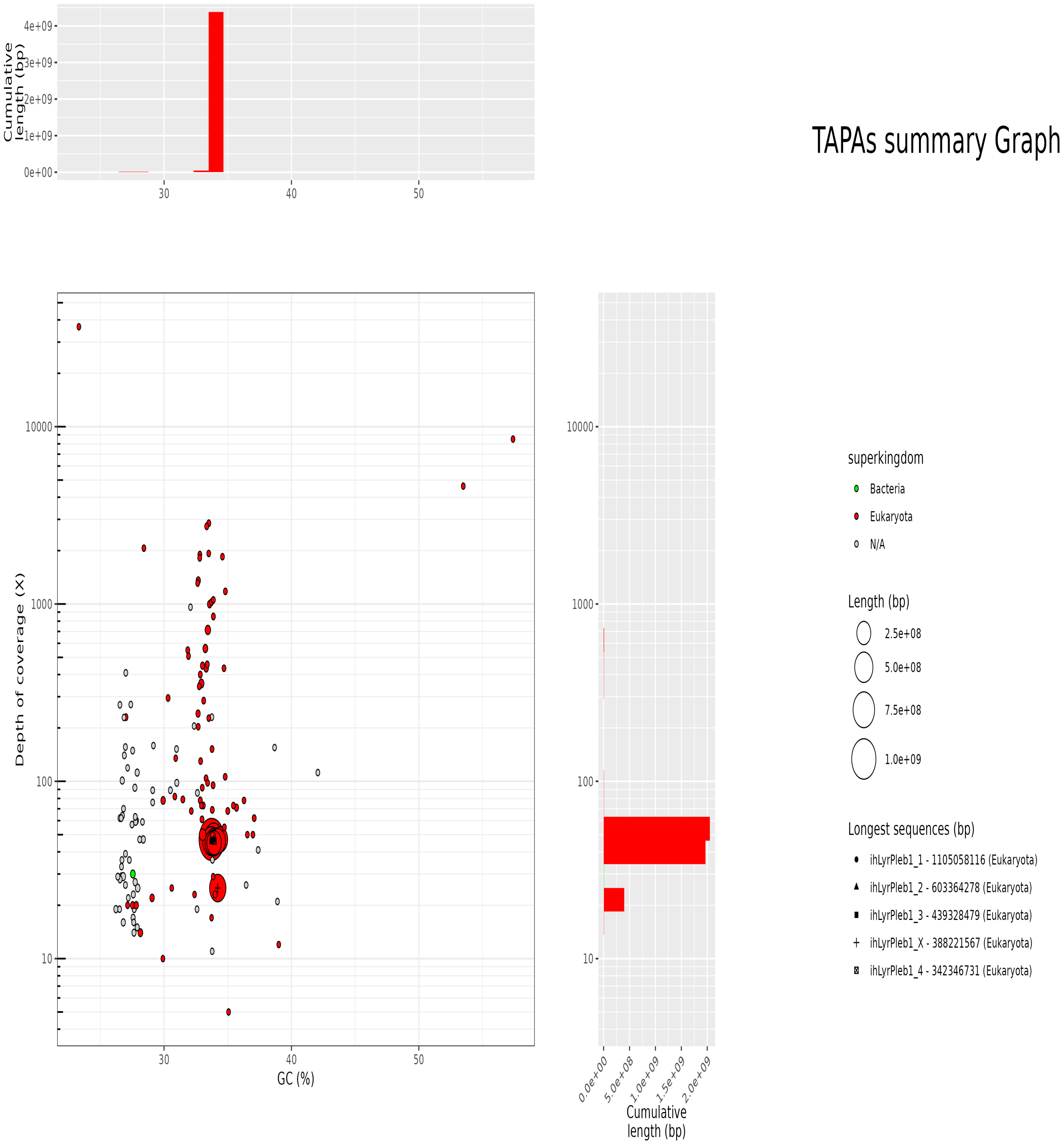


Distribution of k-mer counts per copy numbers found in asm



Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



collapsed. Bubble plot circles are scaled by sequence length, positioned by coverage and GC proportion, and coloured by taxonomy. Histograms show total assembly length distribution on each axis.

Data profile

Data	Long reads	Arima
Coverage	47	307

Assembly pipeline

```
- Hifiasm
    |_ ver: 0.19.5-r593
    |_ key param: NA
- purge_dups
    |_ ver: 1.2.5
    |_ key param: NA
- YaHS
    |_ ver: 1.2
    |_ key param: NA
```

Curation pipeline

```
- PretextMap
    |_ ver: 0.1.9
    |_ key param: NA
- PretextView
    |_ ver: 0.2.5
    |_ key param: NA
```

Submitter: Caroline Menguy
Affiliation: Genoscope

Date and time: 2026-01-12 01:07:06 CET