ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1606511
ToLID	**ucChlMari1**
Species	Chloropicon mariensis
Class	Chloropicophyceae
Order	Chloropicales

Genome Traits	Expected	Observed
Haploid size (bp)	34,624,677	19,767,358
Haploid Number	7 (source: ancestor)	22
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.5.Q32

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

. Observed Haploid size (bp) has >20% difference with Expected
. Observed Haploid Number is different from Expected

. QV value is less than 40 for collapsed
. Kmer completeness value is less than 90 for collapsed
. BUSCO single copy value is less than 90% for collapsed
. Assembly length loss > 3% for collapsed
. More than 1000 gaps/Gbp for collapsed

Curator notes

. Interventions/Gb: 1970
. Contamination notes: ""
. Other observations: "The assembly of Chloropicon mariensis RCC997 (ucChlMari1) is based on 110X ONT data and Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr).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, 40 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 22.4 Mb (with the largest being 3.9Mb). Additionally, 18 regions totaling 479 Kb (with the largest being 68 Kb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using ptgaul. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 3 haplotypic regions were

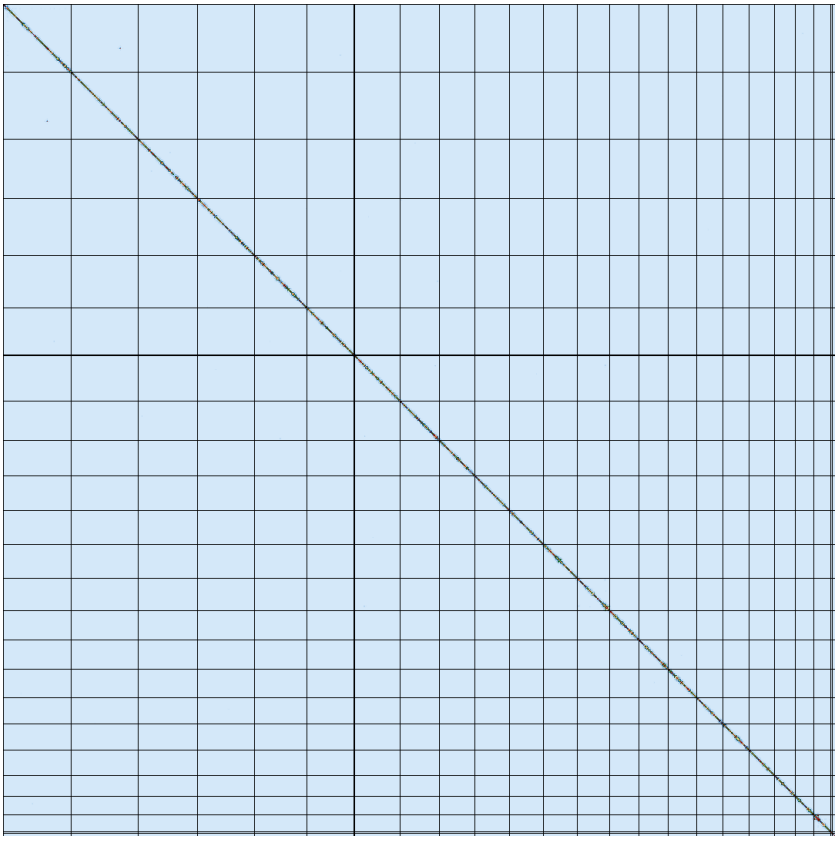
removed,totaling 37 Kb (with the largest being 18 Kb) and 76 contigs were tagged as contaminants (Eukaryotes), totaling 21.2Mb (with the largest being 1.7 Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	41,114,077	19,767,358
GC %	58.11	49.19
Gaps/Gbp	0	1,871.77
Total gap bp	0	7,400
Scaffolds	139	25
Scaffold N50	590,000	930,869
Scaffold L50	26	8
Scaffold L90	80	19
Contigs	139	62
Contig N50	590,000	623,381
Contig L50	26	12
Contig L90	80	36
QV	16.9219	32.9859
Kmer compl.	40.0376	39.9365
BUSCO sing.	28.8%	88.0%
BUSCO dupl.	70.5%	1.3%
BUSCO frag.	0.5%	4.1%
BUSCO miss.	0.3%	6.6%

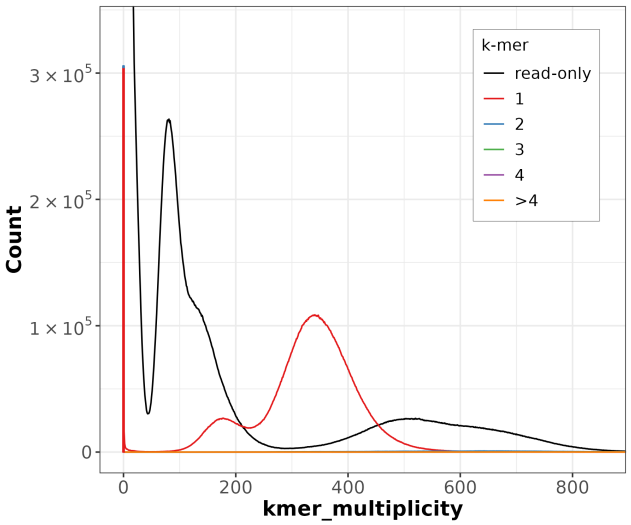
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)

HiC contact map of curated assembly

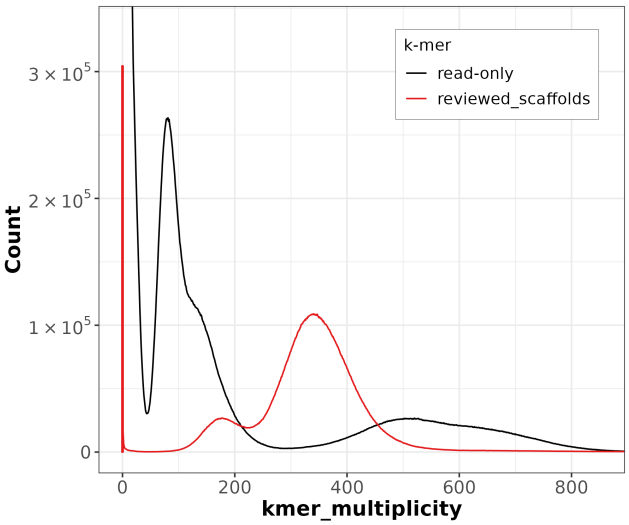


collapsed [LINK]

K-mer spectra of curated assembly

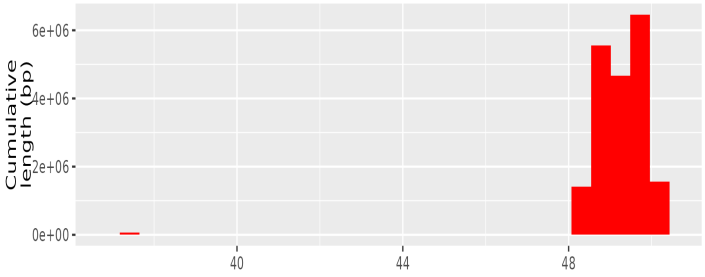


Distribution of k-mer counts per copy numbers found in asm

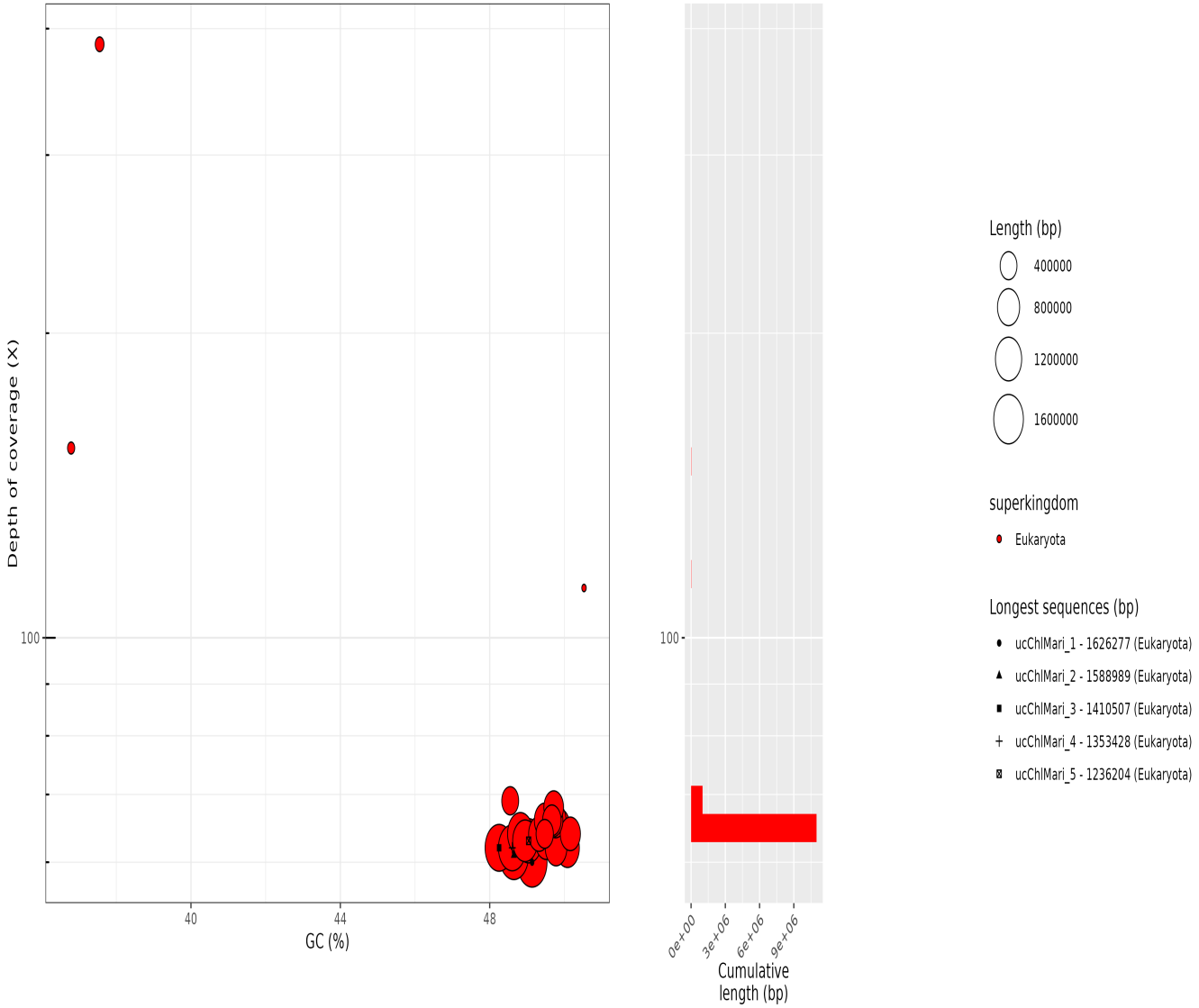


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



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	PACBIO Hifi	Arima
Coverage	110	269

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: 2025-07-30 03:44:23 CEST