ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	148945
ToLID	**ucCauCupr1**
Species	Caulerpa cupressoides
Class	Ulvophyceae
Order	Bryopsidales

Genome Traits	Expected	Observed
Haploid size (bp)	134,711,051	31,269,095
Haploid Number	7 (source: ancestor)	7
Ploidy	1 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q42

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 Ploidy is different from Expected

. Kmer completeness value is less than 90 for collapsed
. BUSCO single copy value is less than 90% for collapsed
. BUSCO duplicated value is more than 5% for collapsed
. Assembly length loss > 3% for collapsed

Curator notes

. Interventions/Gb: 772
. Contamination notes: ""
. Other observations: "The assembly of Caulerpa cupressoides (ucCauCupr1) is based on 112X PacBio HiFi data from an amplified DNA library and 327X 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, 348 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 6.555 Mb (with the largest being 0.095 Mb). Additionally, 632 regions totaling 11.727 Mb (with the largest being 1.288 Mb) were identified as haplotypic duplications and removed. The chloroplast genome was assembled using Oatk and shows an unusual size compared to currently available reference. The mitochondrial genome was assembled using Oatk and shows an unusual size compared to currently available reference.

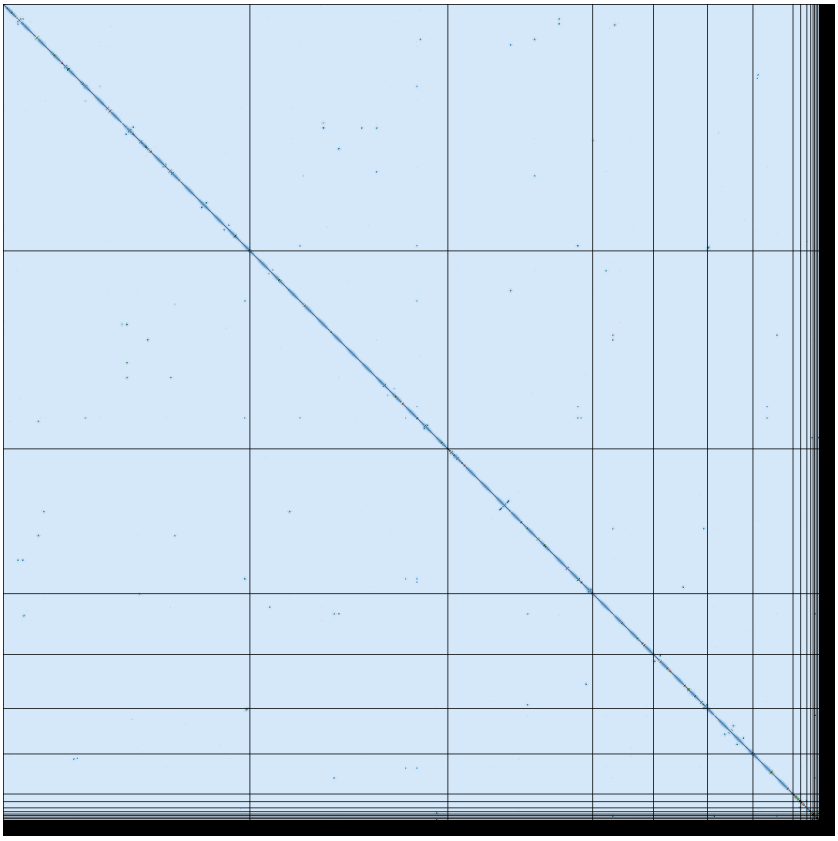
Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 19 haplotypic regions and 74 contaminant sequences were removed, totaling 2.301Mb and 1.902Mb, respectively (with the largest being 0.76Mb and 0.317Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	59,761,704	31,269,095
GC %	39.38	42.83
Gaps/Gbp	368.13	703.57
Total gap bp	2,200	3,400
Scaffolds	1,911	47
Scaffold N50	46,000	7,451,908
Scaffold L50	129	2
Scaffold L90	1,454	6
Contigs	1,933	69
Contig N50	42,477	2,113,839
Contig L50	144	4
Contig L90	1,474	14
QV	32.8667	42.8202
Kmer compl.	64.3408	61.0929
BUSCO sing.	71.4%	74.7%
BUSCO dupl.	14.3%	10.8%
BUSCO frag.	1.8%	1.8%
BUSCO miss.	12.5%	12.7%

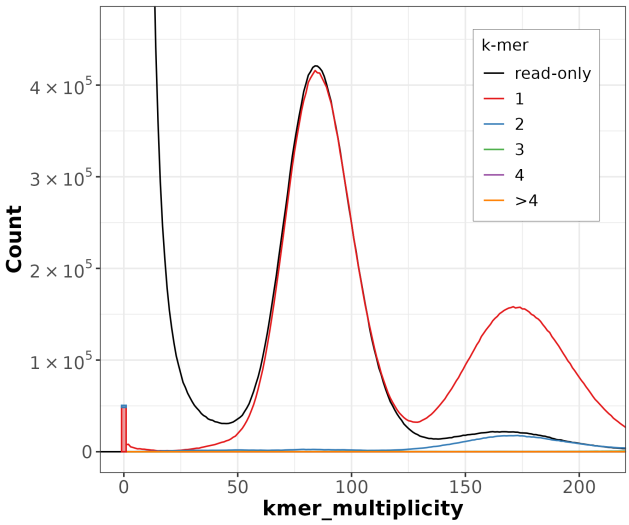
BUSCO: 6.1.0 (euk_genome_min, miniprot) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)

HiC contact map of curated assembly

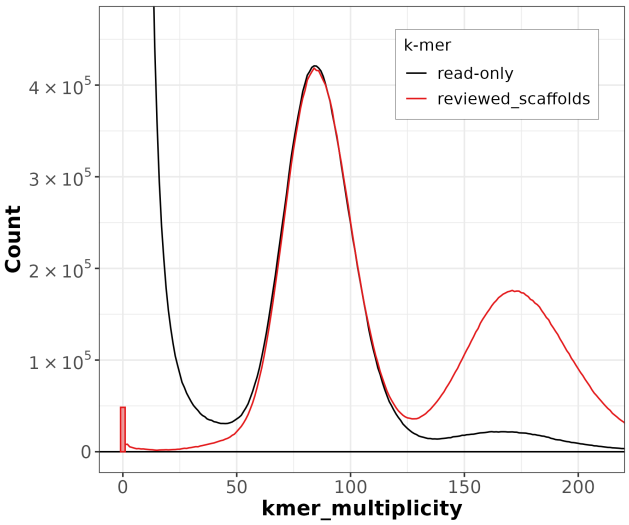


collapsed [LINK]

K-mer spectra of curated assembly

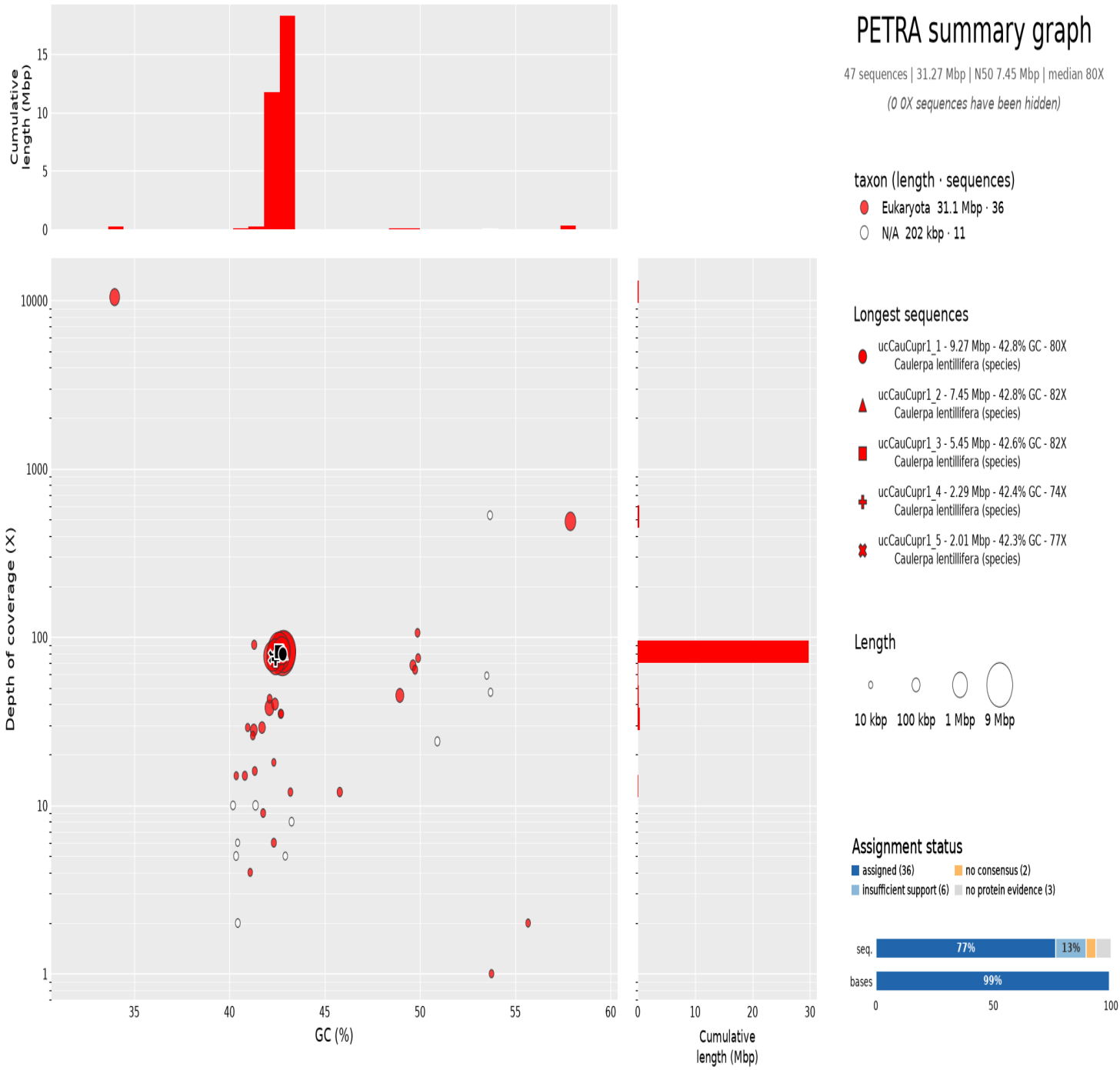


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	214	625

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:* 0.2.4
  |_ *key param:* NA
- **PretextView**
  |_ *ver:* 1.0.7
  |_ *key param:* NA

Submitter: Jean-Marc Aury
Affiliation: Genoscope

Date and time: 2026-11-07 19:44:48 CET