ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	76314
ToLID	**ucCauProl1**
Species	Caulerpa prolifera
Class	Ulvophyceae
Order	Bryopsidales

Genome Traits	Expected	Observed
Haploid size (bp)	62,589,150	29,600,868
Haploid Number	7 (source: ancestor)	0
Ploidy	1 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.5.Q41

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
. 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
. Not 90% of assembly in chromosomes for collapsed

Curator notes

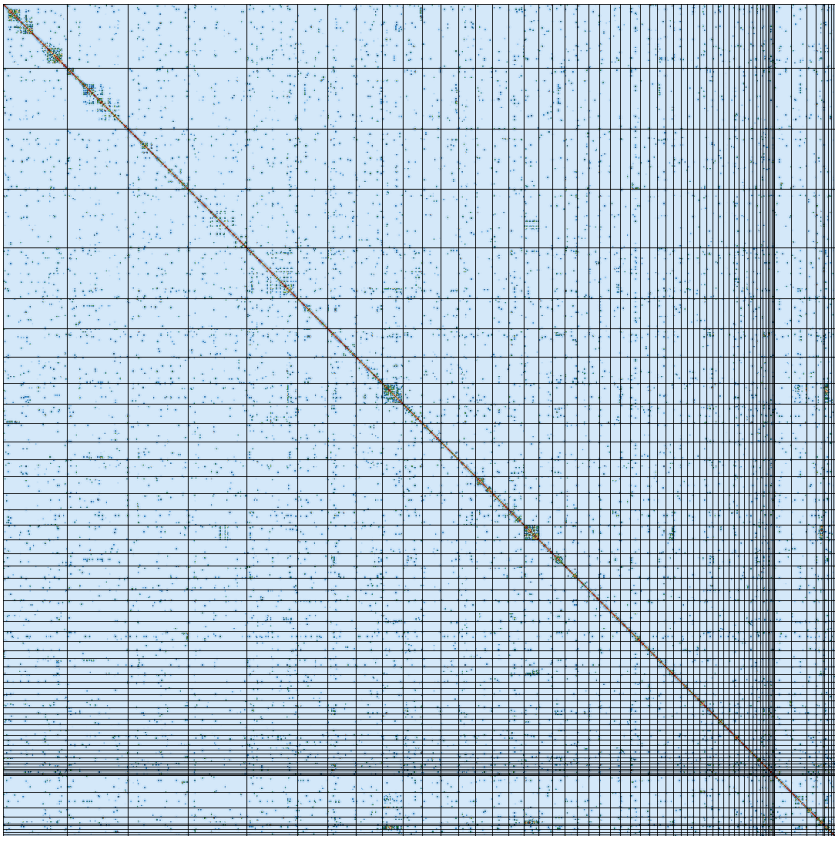
. Interventions/Gb: 0
. Contamination notes: ""
. Other observations: "The assembly of Caulerpa prolifera (ucCauProl1) is based on 157X PacBio data and 844X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe project (BGE, https://biodiversitygenomics.eu/). The assembly process included the following steps: initial PacBio assembly generation with Nextdenovo, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 9 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 8.836 Mb (with the largest being 6.026 Mb). Additionally, 40 regions totaling 2.493 Mb (with the largest being 0.124 Mb) were identified as haplotypic duplications and removed. Mitochondrial genome was assembled using OATK. Genome submitted to the contig scale "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	29,609,268	29,600,868
GC %	43.59	43.59
Gaps/Gbp	2,836.95	0
Total gap bp	8,400	0
Scaffolds	236	62
Scaffold N50	360,413	683,620
Scaffold L50	21	10
Scaffold L90	116	39
Contigs	320	62
Contig N50	166,000	683,620
Contig L50	39	10
Contig L90	188	39
QV	41.5187	41.5188
Kmer compl.	58.5376	58.5526
BUSCO sing.	77.4%	77.4%
BUSCO dupl.	9.2%	9.5%
BUSCO frag.	2.4%	2.0%
BUSCO miss.	11.0%	11.1%

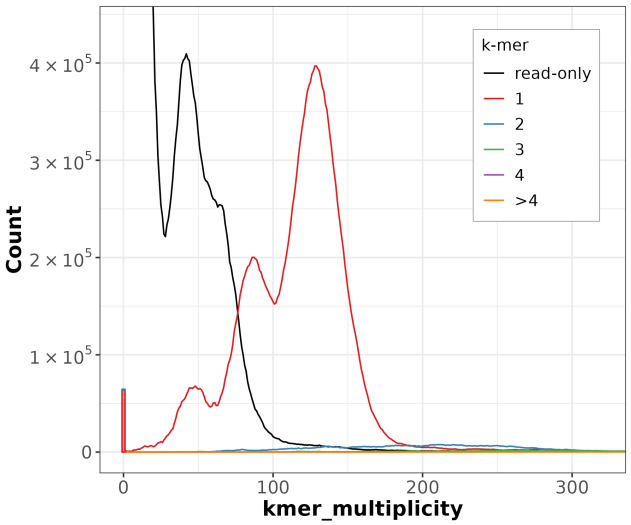
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)

HiC contact map of curated assembly

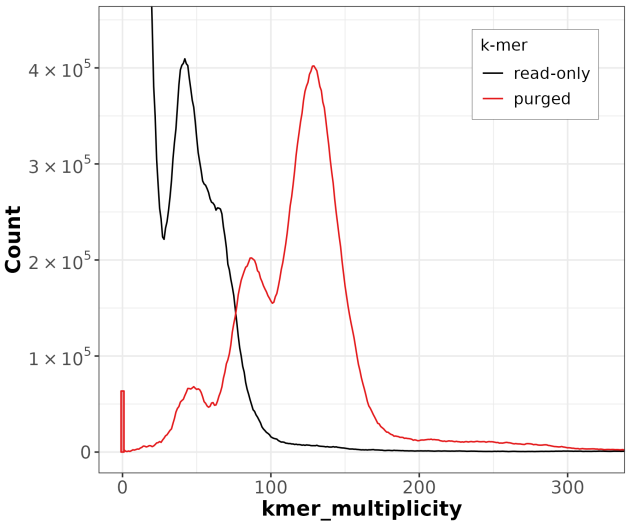


collapsed [LINK]

K-mer spectra of curated assembly

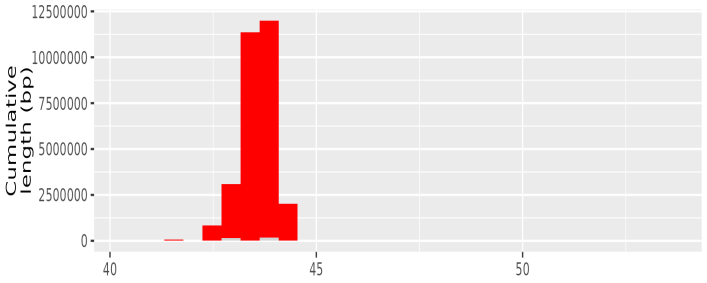


Distribution of k-mer counts per copy numbers found in asm

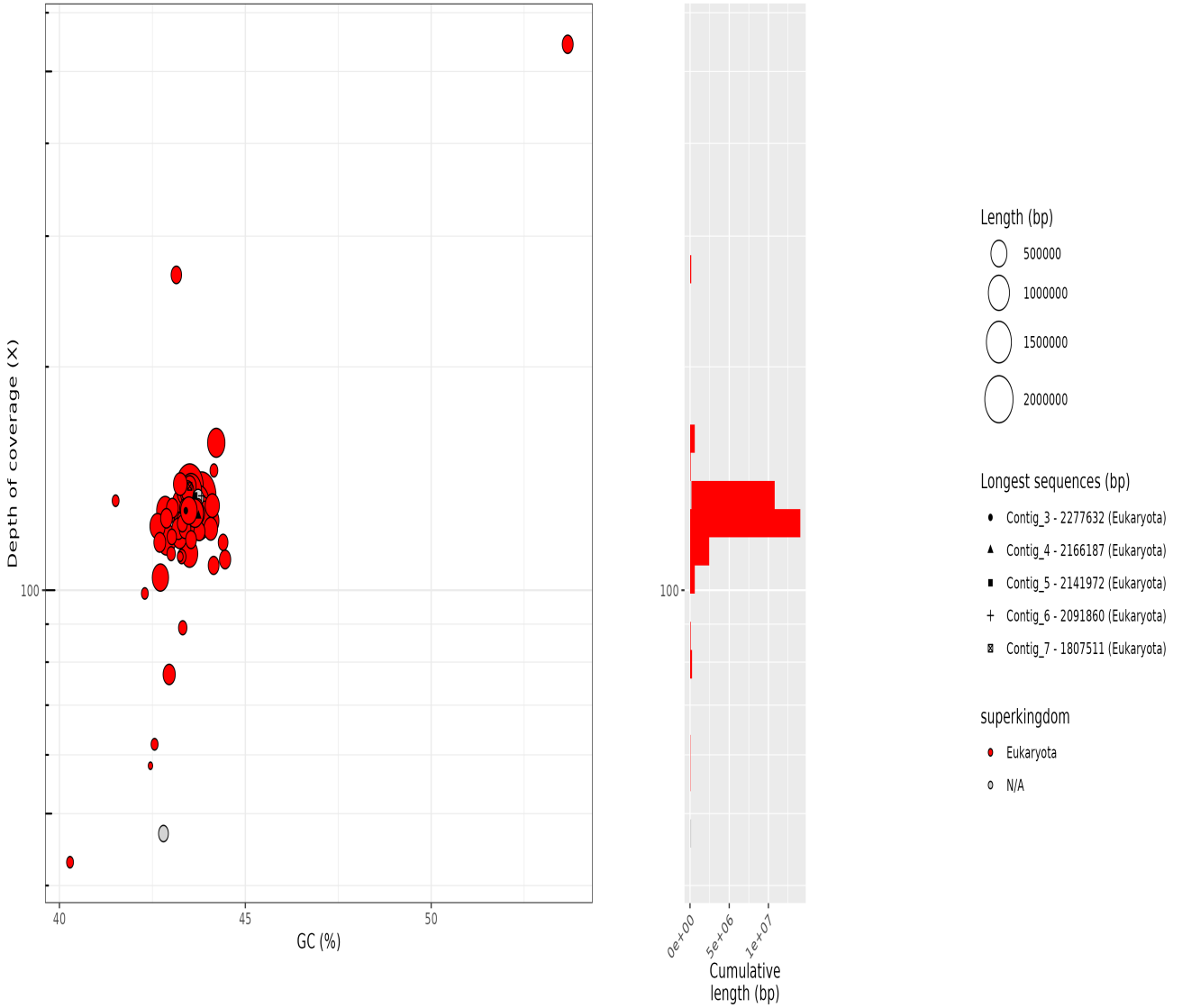


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



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	Long reads	Arima
Coverage	157	844

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: Lola Demirdjian
Affiliation: Genoscope

Date and time: 2026-02-20 20:02:07 CET