ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	41886
ToLID	**ucPseMara1**
Species	Pseudoscourfieldia marina
Class	Pseudoscourfieldiophyceae
Order	Pseudoscourfieldiales

Genome Traits	Expected	Observed
Haploid size (bp)	26,422,947	22,931,100
Haploid Number	7 (source: ancestor)	45
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.5.Q48

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

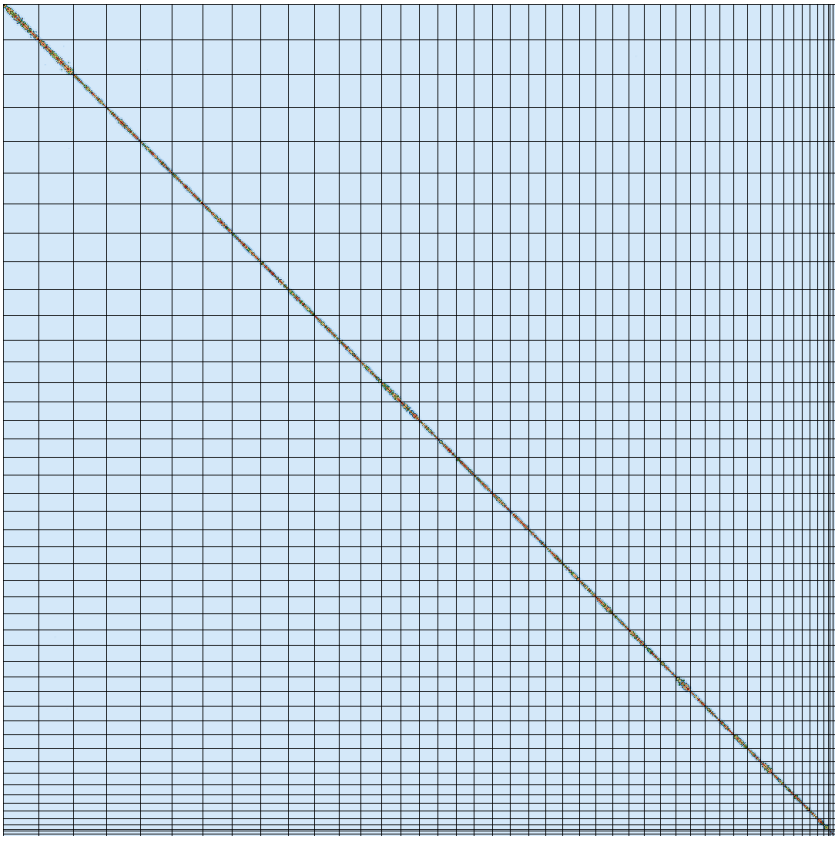
. Interventions/Gb: 0
. Contamination notes: ""
. Other observations: "The assembly of Pseudoscourfieldia marina str. RCC4076 (ucPseMara1) is based on 103X PacBio data and 403X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr).The assembly process included the following 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, 74 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 6.4 Mb (with the largest being 3.4Mb). Additionally, 3 regions totaling 0.11 Mb (with the largest being 0.08 Mb) were identified as haplotypic duplications and removed. The mitochondrial and chloroplastic genomes were assembled using oatk. The obtained mitochondrial genome is not circular. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, no supplementary haplotypic regions were removed. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. Each chromosome is constituted by a single contig without gaps. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	22,882,928	22,931,100
GC %	57.4	57.36
Gaps/Gbp	0	0
Total gap bp	0	0
Scaffolds	46	47
Scaffold N50	510,611	510,611
Scaffold L50	15	15
Scaffold L90	36	36
Contigs	46	47
Contig N50	510,611	510,611
Contig L50	15	15
Contig L90	36	36
QV	48.6655	48.6746
Kmer compl.	86.2662	86.2662
BUSCO sing.	90.7%	90.7%
BUSCO dupl.	1.2%	1.2%
BUSCO frag.	0.9%	0.9%
BUSCO miss.	7.2%	7.2%

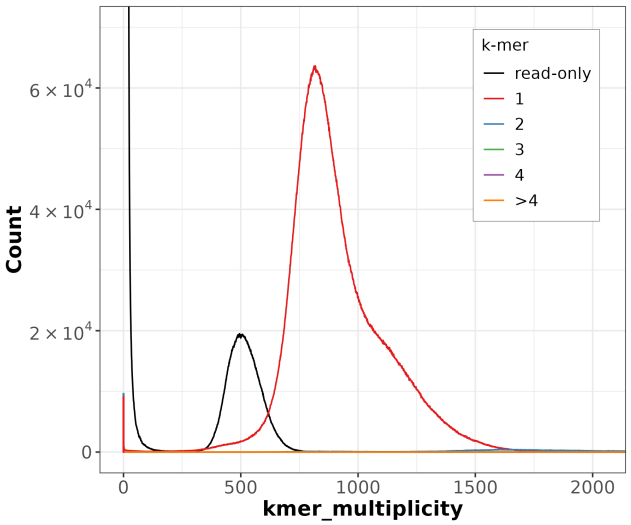
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)

HiC contact map of curated assembly

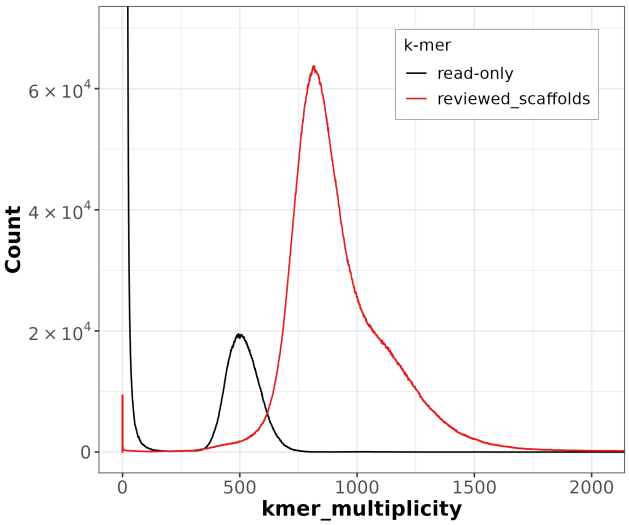


collapsed [LINK]

K-mer spectra of curated assembly

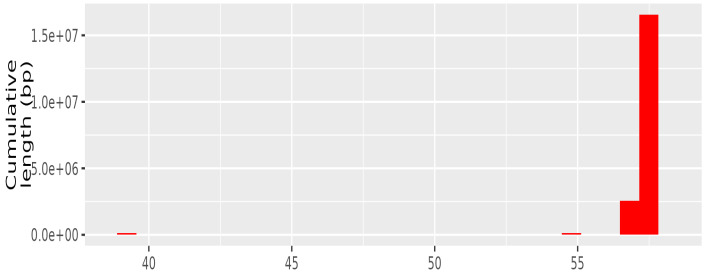


Distribution of k-mer counts per copy numbers found in asm

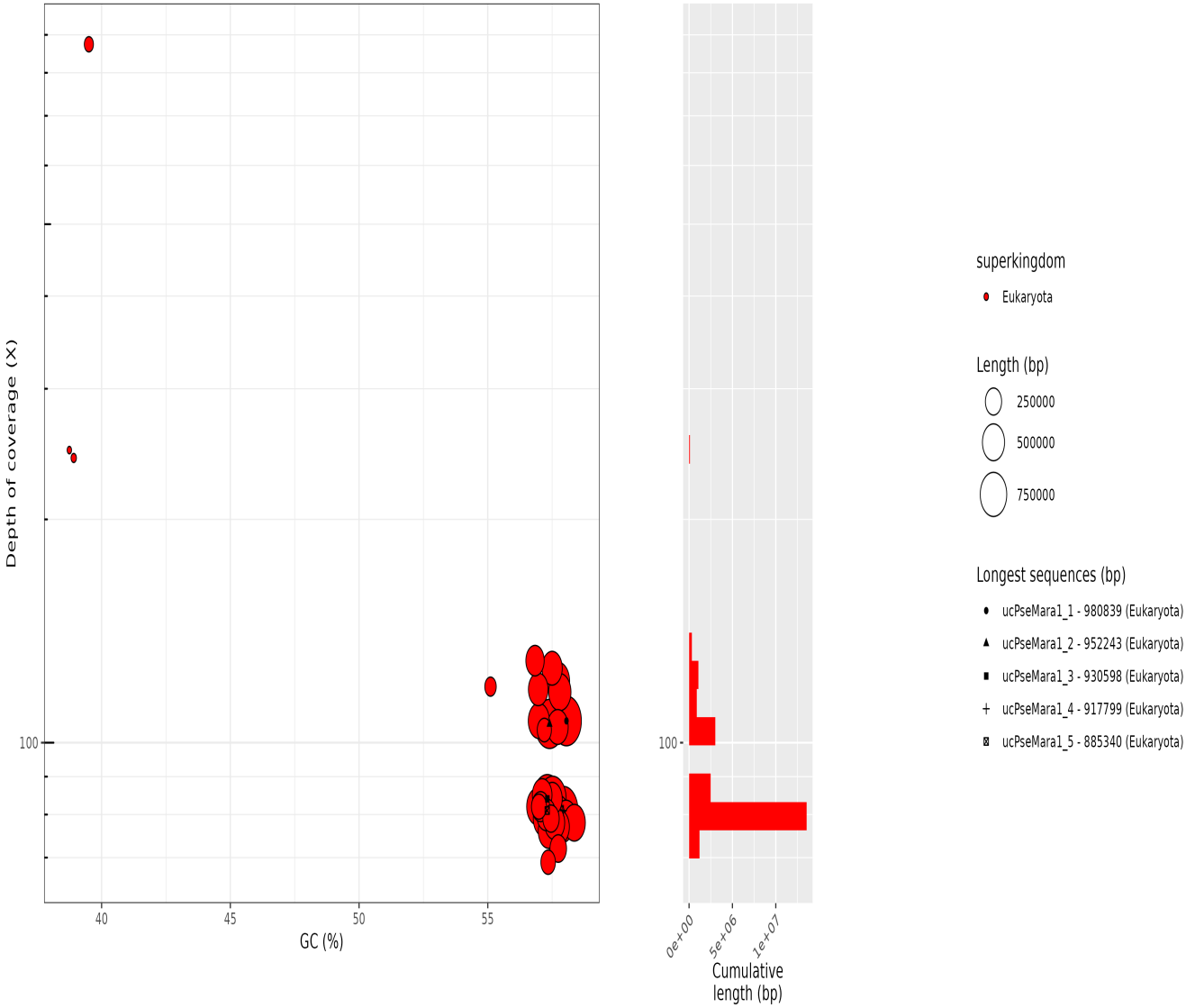


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	103	403

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: 2025-11-26 03:10:46 CET