ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	41886
ToLID	**ucPseMara2**
Species	Pseudoscourfieldia marina
Class	Pseudoscourfieldiophyceae
Order	Pseudoscourfieldiales

Genome Traits	Expected	Observed
Haploid size (bp)	33,089,377	30,651,986
Haploid Number	7 (source: ancestor)	43
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
. Assembly length loss > 3% for collapsed

Curator notes

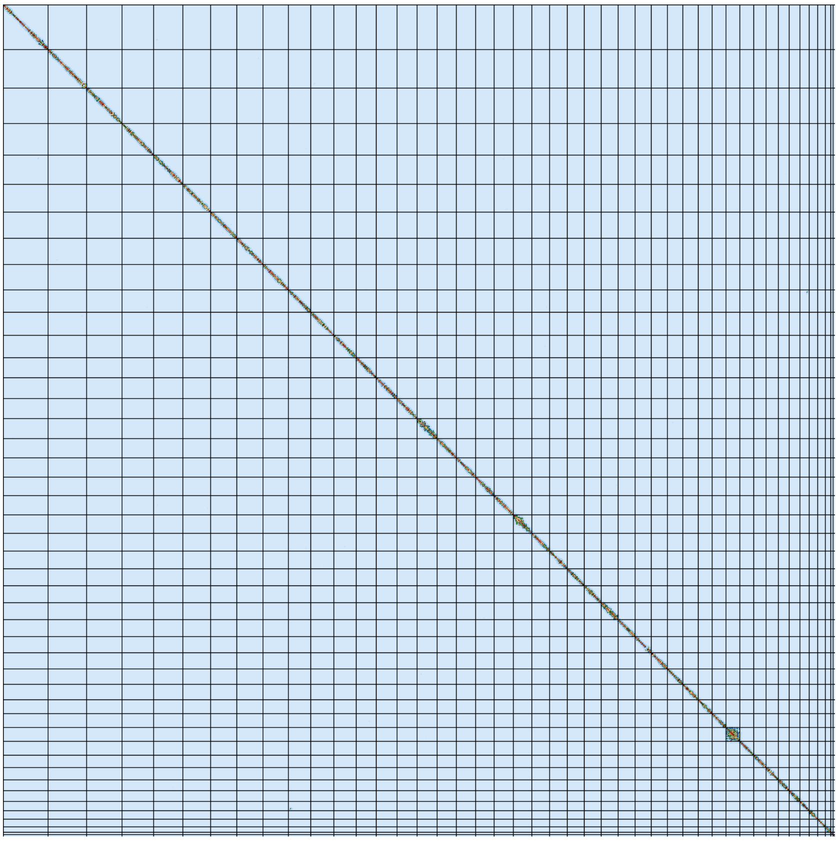
. Interventions/Gb: 0
. Contamination notes: ""
. Other observations: "The assembly of \'Pseudoscourfieldia marina sp.RCC7688\' (ucPseMara2) is based on 76X PacBio data and 368X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following 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, 16 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 8.721 Mb (with the largest being 3.423 Mb). Additionally, 3 regions totaling 0.192 Mb (with the largest being 0.087 Mb) were identified as haplotypic duplications and removed. Mitochondrial and chloroplastic genomes were assembled using OATK. During manual curation, 3 haplotypic regions were removed, totaling 1.615Mb (with the largest being 0.671Mb). 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	32,202,882	30,651,986
GC %	57.61	57.57
Gaps/Gbp	0	0
Total gap bp	0	0
Scaffolds	47	45
Scaffold N50	713,050	733,193
Scaffold L50	17	16
Scaffold L90	38	35
Contigs	47	45
Contig N50	713,050	733,193
Contig L50	17	16
Contig L90	38	35
QV	48.6801	48.6925
Kmer compl.	67.6506	67.1209
BUSCO sing.	85.2%	90.8%
BUSCO dupl.	6.6%	1.6%
BUSCO frag.	1.8%	0.9%
BUSCO miss.	6.4%	6.7%

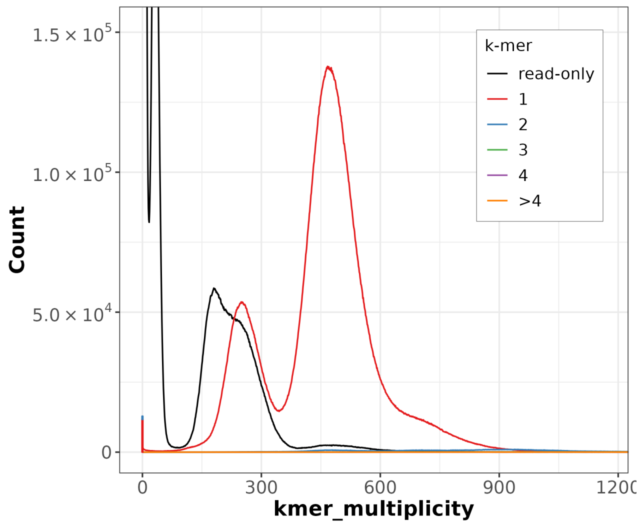
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)

HiC contact map of curated assembly

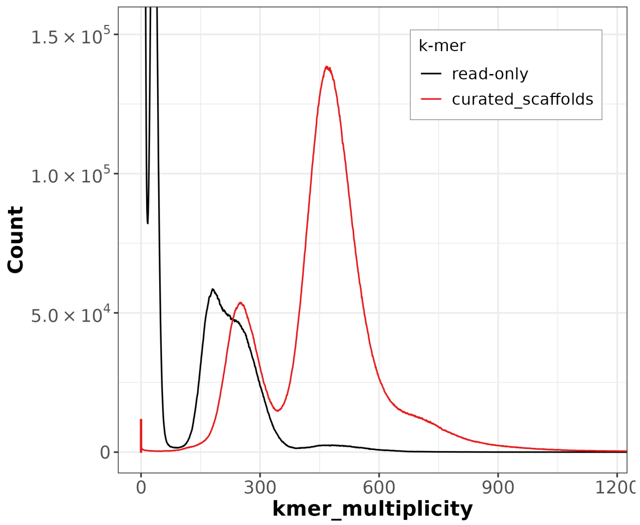


collapsed [LINK]

K-mer spectra of curated assembly

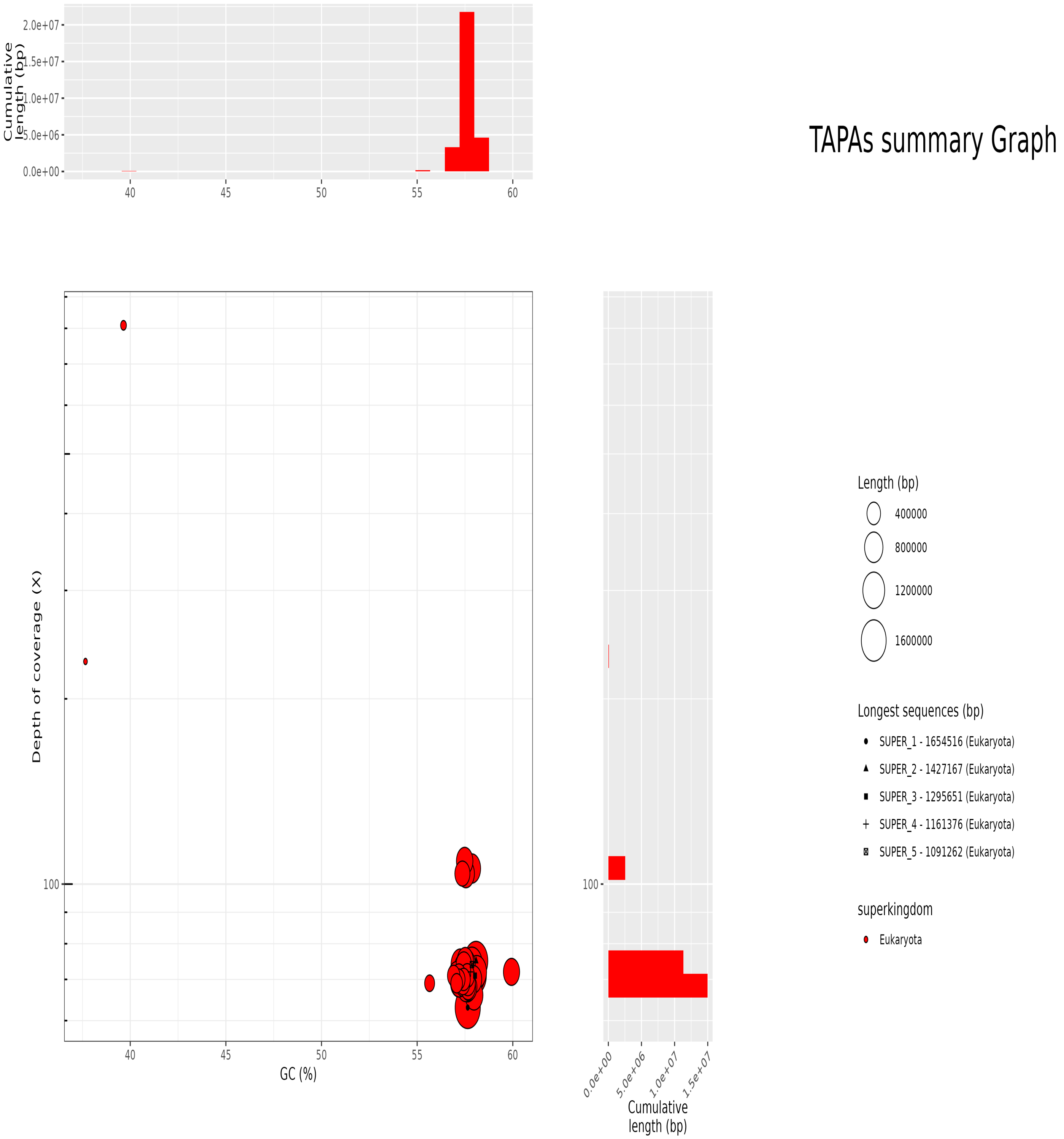


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	76	368

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: Simone Duprat
Affiliation: Genoscope

Date and time: 2025-12-13 10:47:32 CET