ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3451506
ToLID	**ucMicSpeb1**
Species	Micromonas sp. RCC1109
Class	Mamiellophyceae
Order	Mamiellales

Genome Traits	Expected	Observed
Haploid size (bp)	21,346,823	20,054,854
Haploid Number	7 (source: ancestor)	19
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q39

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

Curator notes

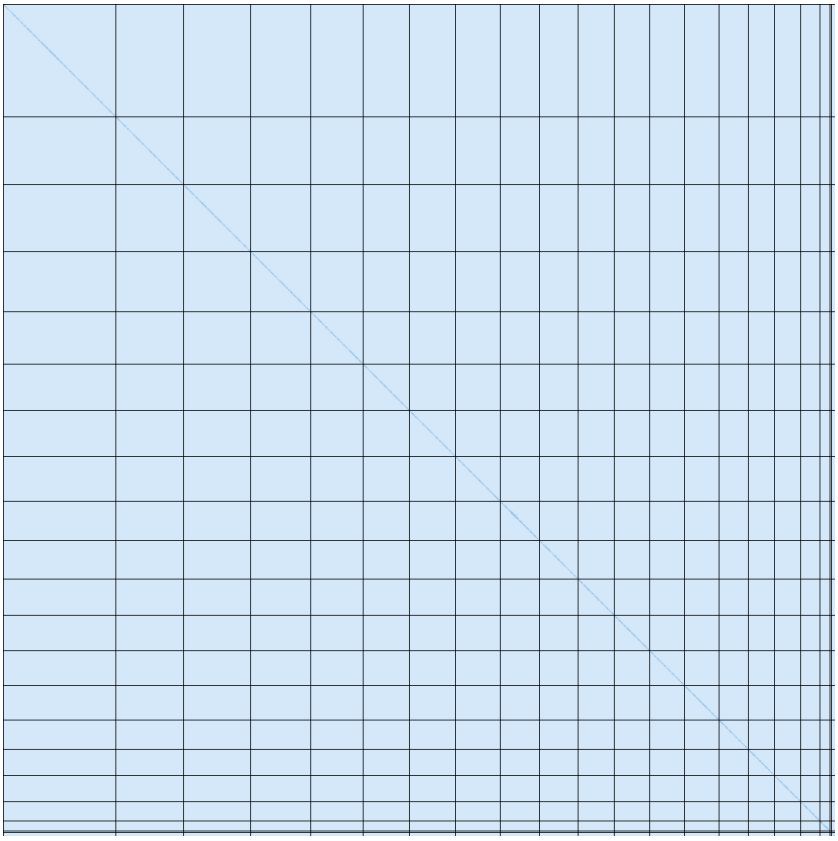
. Interventions/Gb: 200
. Contamination notes: ""
. Other observations: "The assembly of Micromonas sp RCC1109 (ucMicSpeb1) is based on 388X ONT data and 321X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial ONT assembly generation with Flye, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 8 contigs were identified as contaminants (bacterial) totaling 4.3 Mb (with the largest being 3 Mb). Additionally, 4 regions totaling 79 kb (with the largest being 26 kb) were identified as haplotypic duplications and removed. The mitochondrial genome and chloroplast genome were assembled using ptGAUL. Finally, the primary assembly was analyzed and manually improved using Pretext.During manual curation, 8 contaminant sequences were removed, totaling 0.46 Mb (with the largest being 0.12 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	20,397,286	20,054,854
GC %	65.89	66.43
Gaps/Gbp	441.24	698.09
Total gap bp	900	1,900
Scaffolds	31	21
Scaffold N50	1,107,173	1,107,173
Scaffold L50	7	7
Scaffold L90	17	16
Contigs	40	35
Contig N50	1,107,173	1,107,173
Contig L50	7	7
Contig L90	22	21
QV	30.7844	39.0396
Kmer compl.	66.4014	66.6603
BUSCO sing.	90.5%	90.5%
BUSCO dupl.	0.5%	0.5%
BUSCO frag.	3.0%	3.0%
BUSCO miss.	6.0%	6.0%

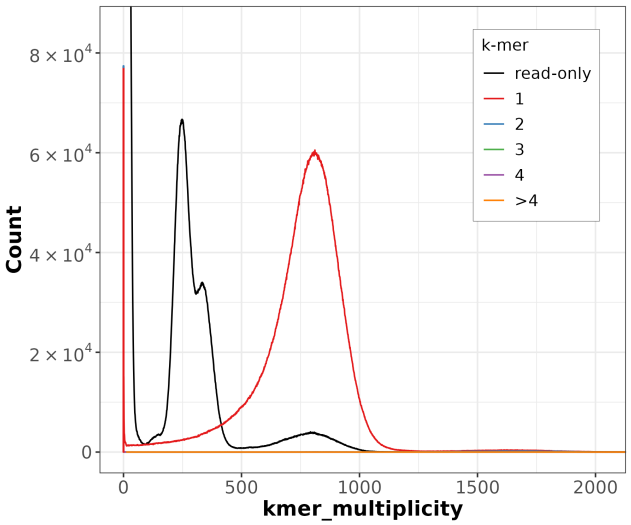
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)

HiC contact map of curated assembly

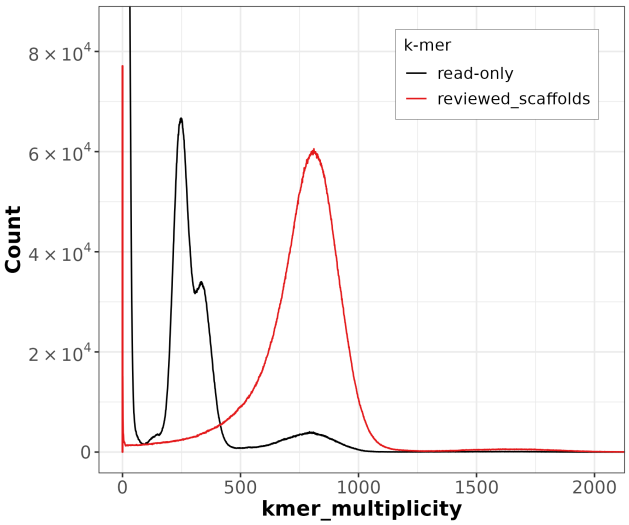


collapsed [LINK]

K-mer spectra of curated assembly

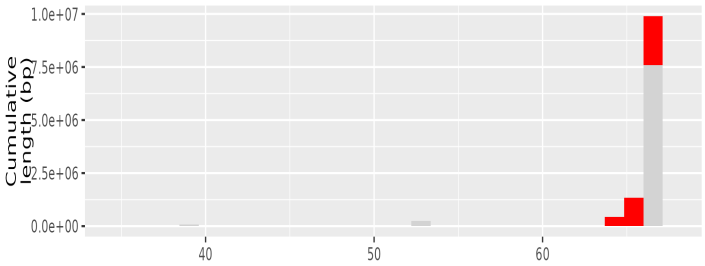


Distribution of k-mer counts per copy numbers found in asm

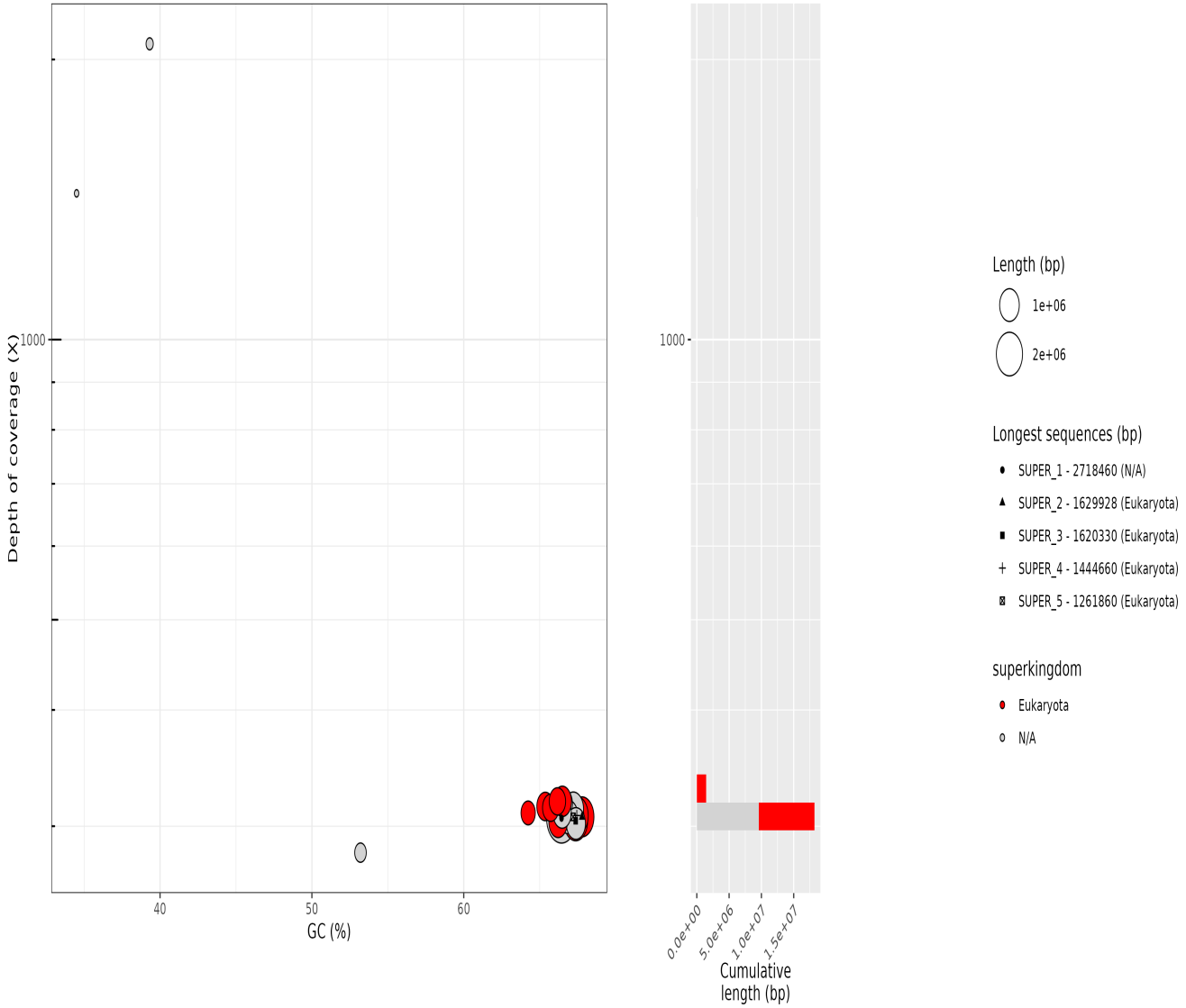


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	Long reads	Arima
Coverage	388	321

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: Sophie Layac
Affiliation: Genoscope

Date and time: 2025-09-21 01:21:01 CEST