ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3451514
ToLID	**uoAstSpea1**
Species	Asterionellopsis sp. RCC1712
Class	Fragilariophyceae
Order	Fragilariales

Genome Traits	Expected	Observed
Haploid size (bp)	692,893,854	78,201,717
Haploid Number	19 (source: ancestor)	16
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q48

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

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

Curator notes

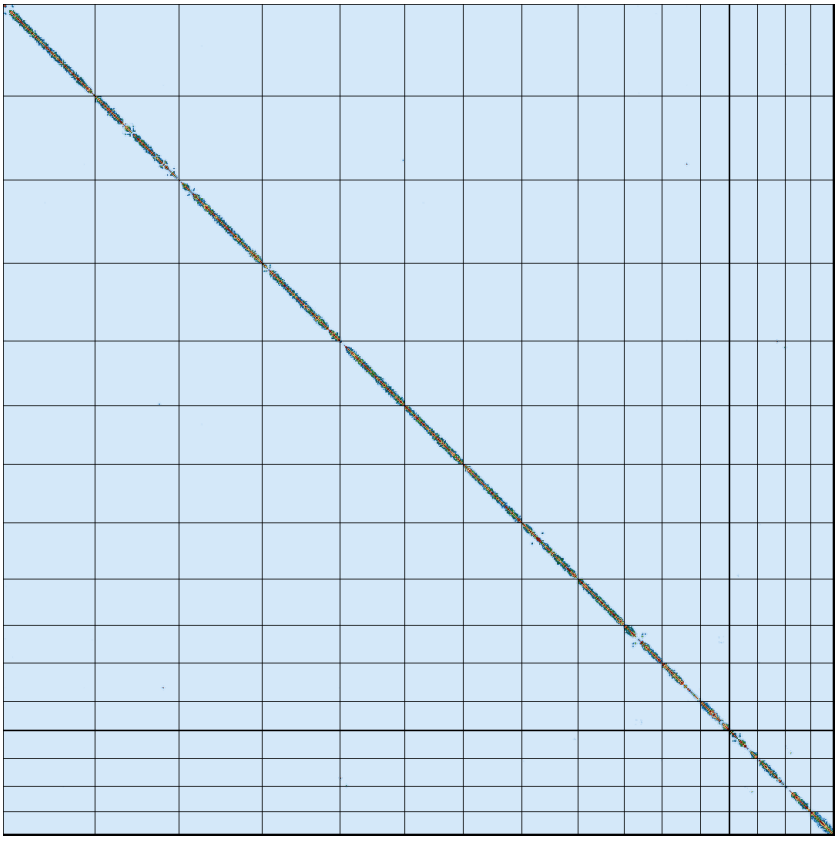
. Interventions/Gb: 235
. Contamination notes: ""
. Other observations: "The assembly of Asterionellopsis sp._RCC1712 (uoAstSpea1) is based on 56X PacBio data and 10X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr).The assembly process included the following steps: 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, 4 646 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 176 Mb (with the largest being 4.8Mb). Additionally, 185 regions totaling 8.8 Mb (with the largest being 0.8Mb) were identified as haplotypic duplications and removed. The mitochondrial and chloroplastic genomes were assembled using oatk. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 4 haplotypic regions were removed, totaling 2.6 Gb (with the largest being 1.79 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	80,905,597	78,201,717
GC %	42.62	42.63
Gaps/Gbp	197.76	332.47
Total gap bp	1,600	3,700
Scaffolds	30	20
Scaffold N50	5,523,572	5,523,572
Scaffold L50	6	6
Scaffold L90	15	13
Contigs	46	46
Contig N50	2,972,000	2,972,000
Contig L50	9	8
Contig L90	24	25
QV	48.1709	48.62
Kmer compl.	71.0503	70.5946
BUSCO sing.	85.9%	88.3%
BUSCO dupl.	6.3%	4.0%
BUSCO frag.	2.0%	2.0%
BUSCO miss.	5.8%	5.8%

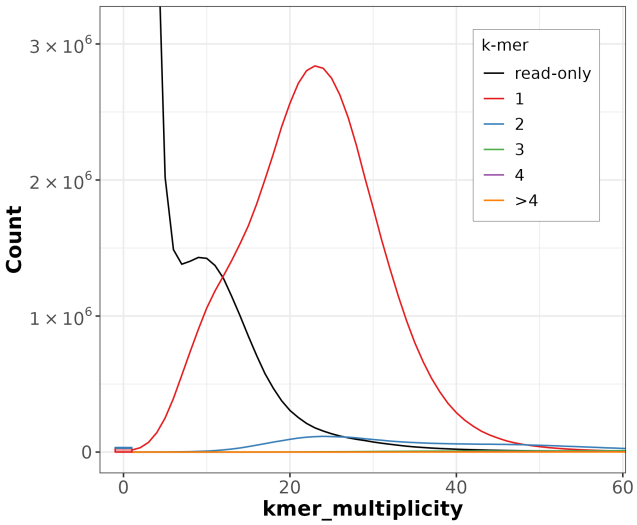
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: bacillariophyta_odb12 (genomes:8, BUSCOs:2944)

HiC contact map of curated assembly

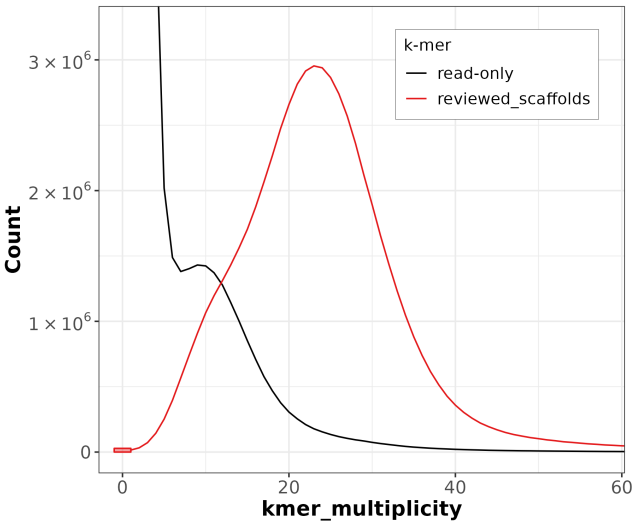


collapsed [LINK]

K-mer spectra of curated assembly

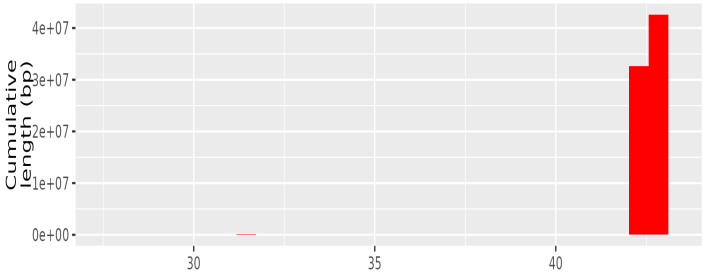


Distribution of k-mer counts per copy numbers found in asm

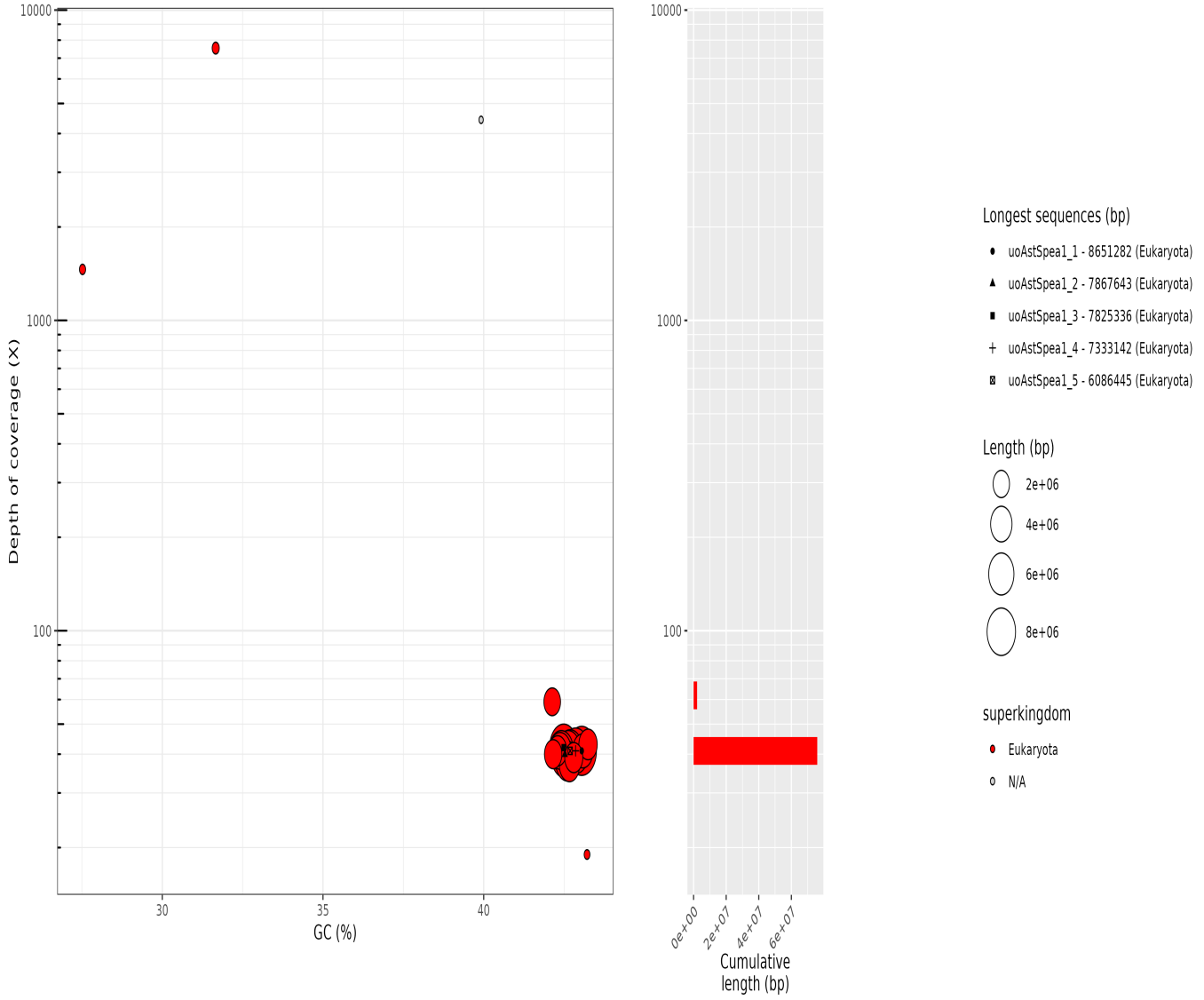


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	56	10

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: 2025-12-17 22:04:21 CET