ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1606638
ToLID	**uoPelSpea1**
Species	Pelagomonas sp. RCC986
Class	Pelagophyceae
Order	Pelagomonadales

Genome Traits	Expected	Observed
Haploid size (bp)	62,316,895	33,111,529
Haploid Number	4 (source: ancestor)	7
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q43

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

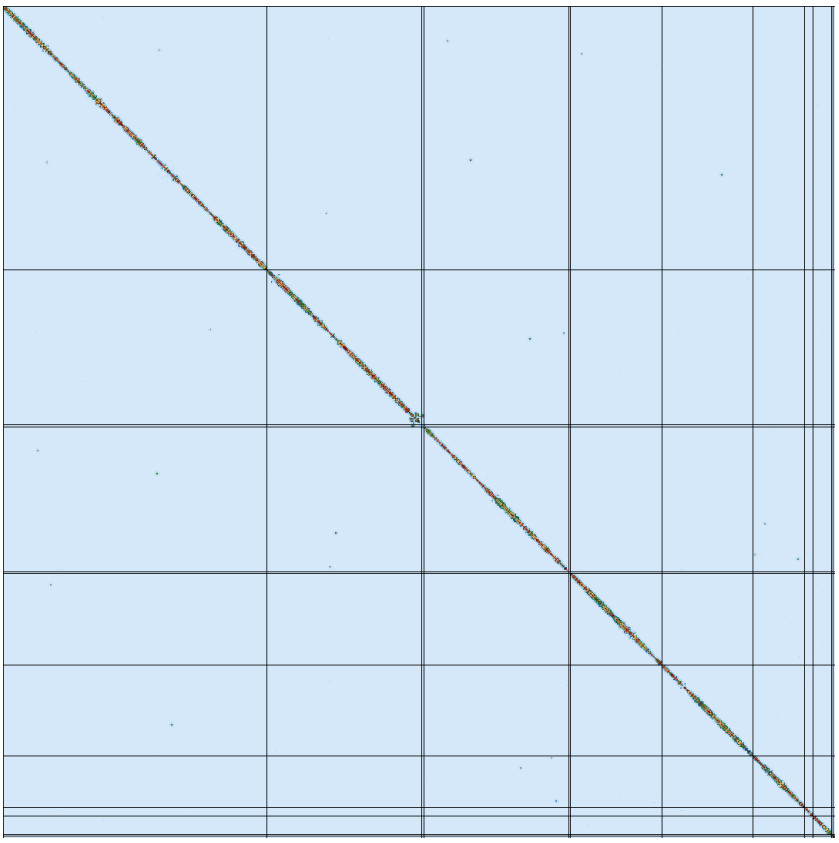
. Interventions/Gb: 140
. Contamination notes: ""
. Other observations: "The assembly of Pelagomonas sp. RCC986 (uoPelSpea1) is based on 31X PacBio data and 360X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr).The assembly process included the following steps: initial 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, 11 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 378 kb (with the largest being 66 kb). Additionally, 15 regions totaling 6 Mb (with the largest being 3 Mb) 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 supplementary haplotypic regions were removed, totaling 2.7 Mb (with the largest being 1.7 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	35,808,998	33,111,529
GC %	63.24	63.23
Gaps/Gbp	83.78	120.8
Total gap bp	300	500
Scaffolds	13	12
Scaffold N50	5,763,302	6,154,543
Scaffold L50	3	2
Scaffold L90	7	6
Contigs	16	16
Contig N50	3,917,259	3,917,259
Contig L50	4	4
Contig L90	9	8
QV	43.9008	43.8474
Kmer compl.	71.821	69.5867
BUSCO sing.	82.1%	88.4%
BUSCO dupl.	10.6%	4.3%
BUSCO frag.	1.0%	1.0%
BUSCO miss.	6.3%	6.3%

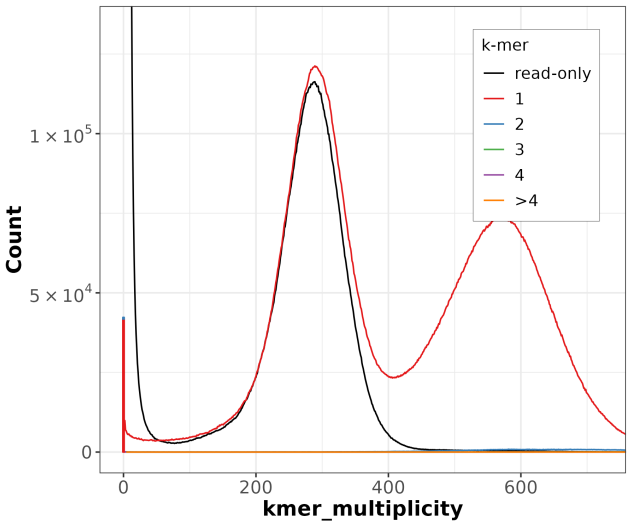
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: stramenopiles_odb12 (genomes:55, BUSCOs:697)

HiC contact map of curated assembly

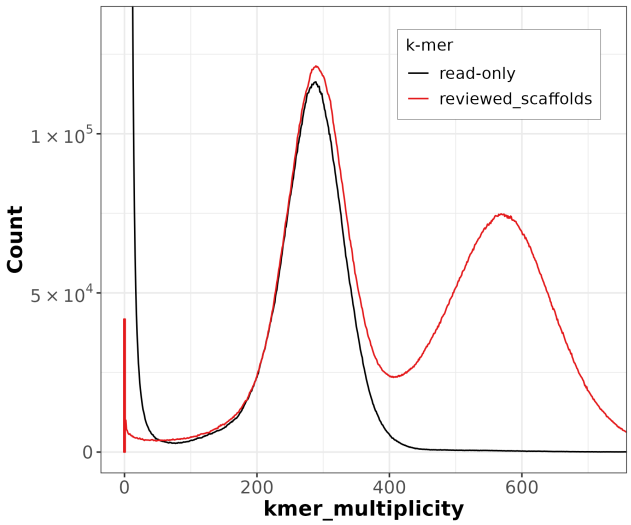


collapsed [LINK]

K-mer spectra of curated assembly

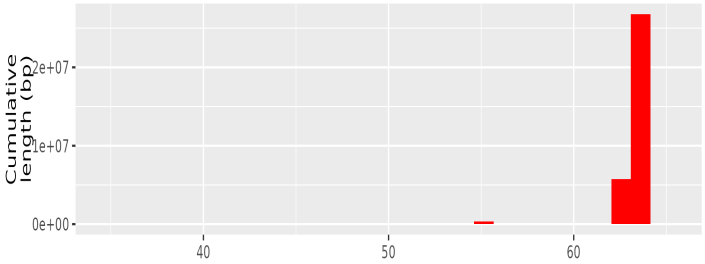


Distribution of k-mer counts per copy numbers found in asm

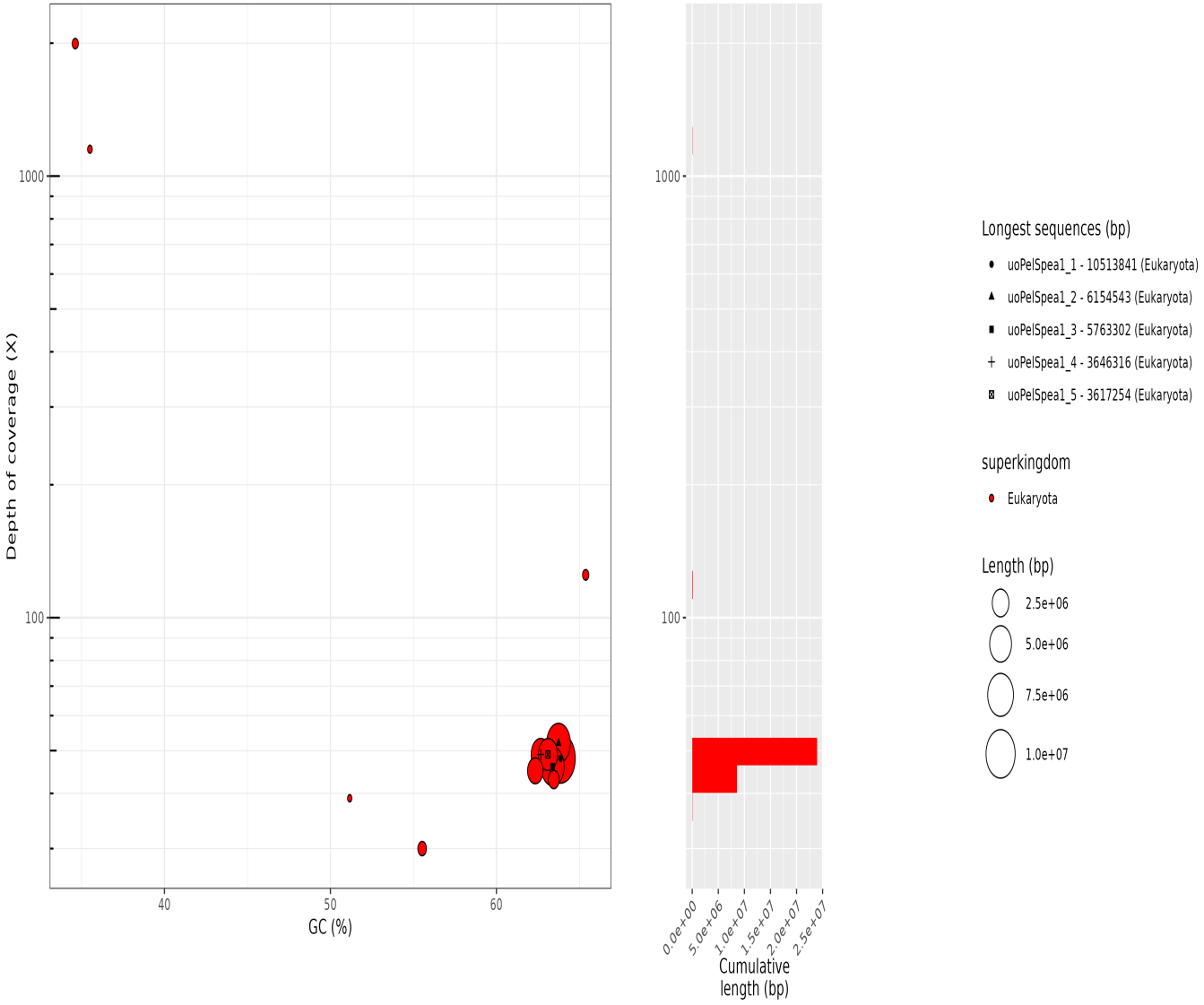


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	31	360

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: Mohamed Meftah Saoues
Affiliation: Genoscope

Date and time: 2026-01-16 06:46:42 CET