ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3451525
ToLID	**uoPelSpee1**
Species	Pelagomonas sp. RCC108
Class	Pelagophyceae
Order	Pelagomonadales

Genome Traits	Expected	Observed
Haploid size (bp)	212,016,307	32,863,366
Haploid Number	4 (source: ancestor)	6
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q41

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

Curator notes

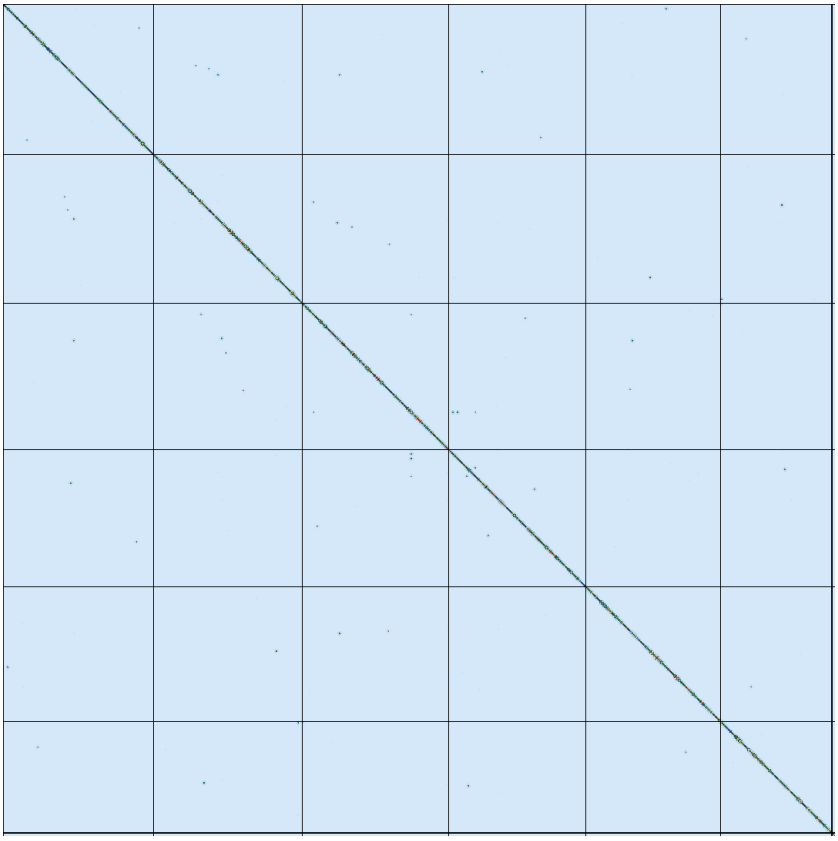
. Interventions/Gb: 0
. Contamination notes: ""
. Other observations: "The assembly of Pelagomonas calceolata str. RCC108 (uoPelSpee1) is based on 295X PacBio data and 589X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following 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, 34 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 19.959 Mb (with the largest being 5.168 Mb). Additionally, 7 regions totaling 0.35 Mb (with the largest being 0.091 Mb) were identified as haplotypic duplications and removed. Mitochondrial and chloroplastic genomes were assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. 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	32,745,634	32,863,366
GC %	63.53	63.43
Gaps/Gbp	61.08	60.86
Total gap bp	200	200
Scaffolds	7	8
Scaffold N50	5,779,576	5,779,576
Scaffold L50	3	3
Scaffold L90	6	6
Contigs	9	10
Contig N50	5,405,630	5,405,630
Contig L50	3	3
Contig L90	7	7
QV	41.6888	41.7044
Kmer compl.	30.3173	30.4304
BUSCO sing.	94.4%	94.1%
BUSCO dupl.	1.4%	1.9%
BUSCO frag.	1.1%	1.1%
BUSCO miss.	3.0%	2.9%

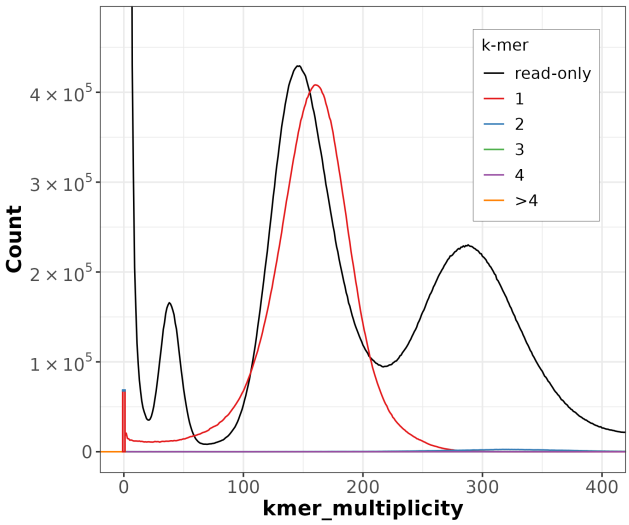
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: stramenopiles_odb12 (genomes:55, BUSCOs:697)
BUSCO: 5.8.2 (euk_genome_min, miniprot) / Lineage: stramenopiles_odb12 (genomes:55, BUSCOs:697)

HiC contact map of curated assembly

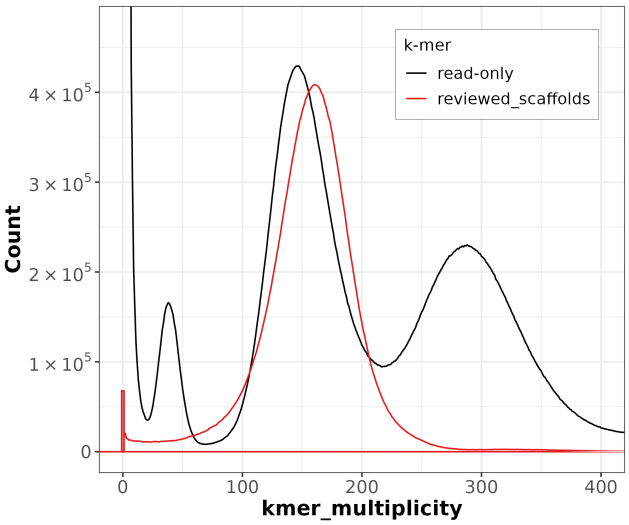


collapsed [LINK]

K-mer spectra of curated assembly

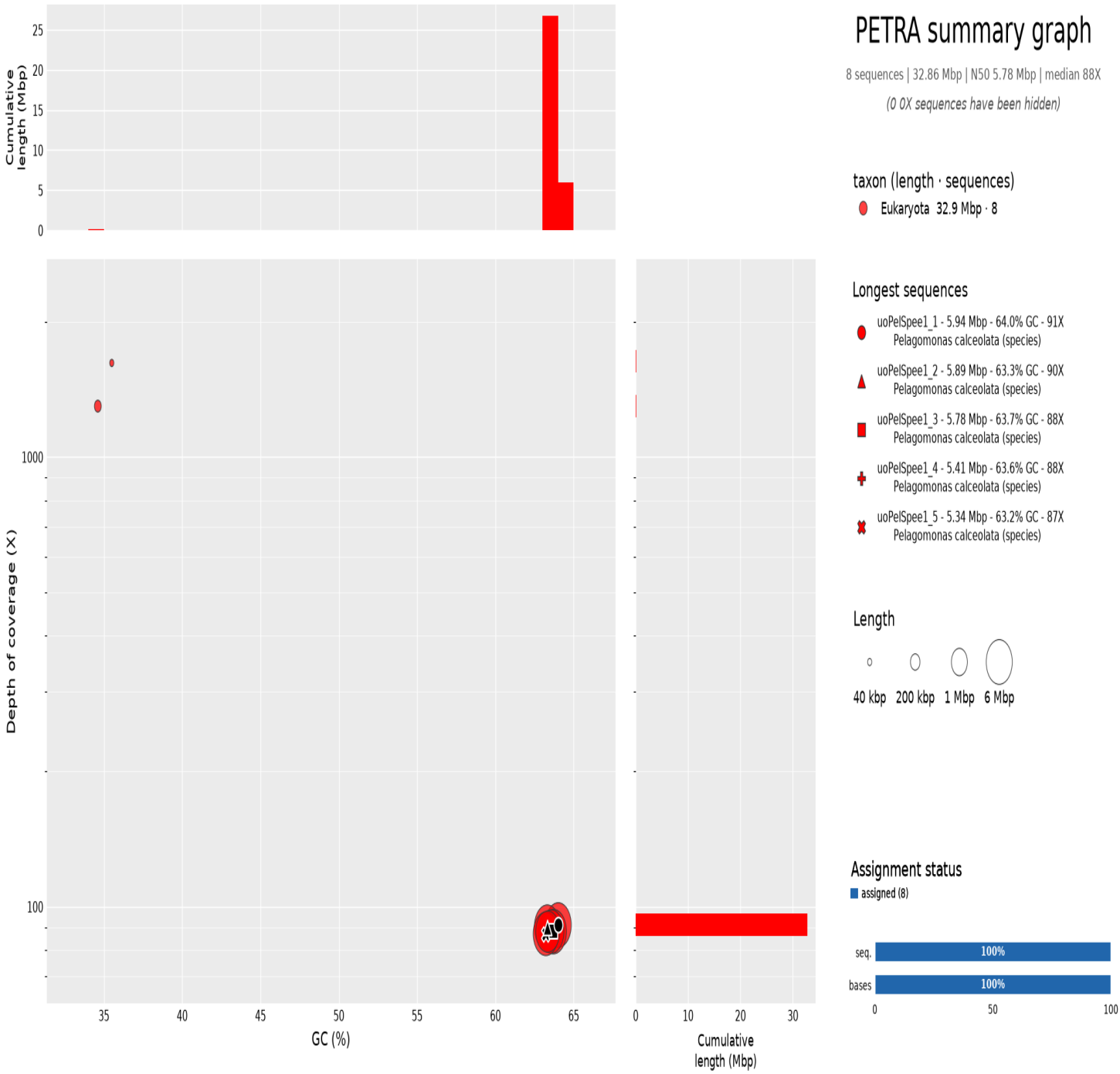


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	294	586

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:* 0.2.4
  |_ *key param:* NA
- **PretextView**
  |_ *ver:* 1.0.7
  |_ *key param:* NA

Submitter: Mohamed Meftah Saoues
Affiliation: Genoscope

Date and time: 2026-10-06 09:58:21 CEST