ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3451521
ToLID	**uoFraSpea1**
Species	Fragilariaceae sp. RCC3059
Class	Fragilariophyceae
Order	Fragilariales

Genome Traits	Expected	Observed
Haploid size (bp)	64,339,657	42,819,093
Haploid Number	4 (source: ancestor)	23
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: 87
. Contamination notes: ""
. Other observations: "The assembly of Fragilariaceae sp. RCC3059 (uoFraSpea1) is based on 328X PacBio data and 821X 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, 45 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 14.15 Mb (with the largest being 4.031 Mb). Additionally, 235 regions totaling 4.821 Mb (with the largest being 0.098 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. During manual curation, 13 haplotypic regions and 10 contaminant sequences were removed, totaling 2.896Mb and 0.327Mb, respectively (with the largest being 0.833Mb and 0.048Mb). The telomeric pattern

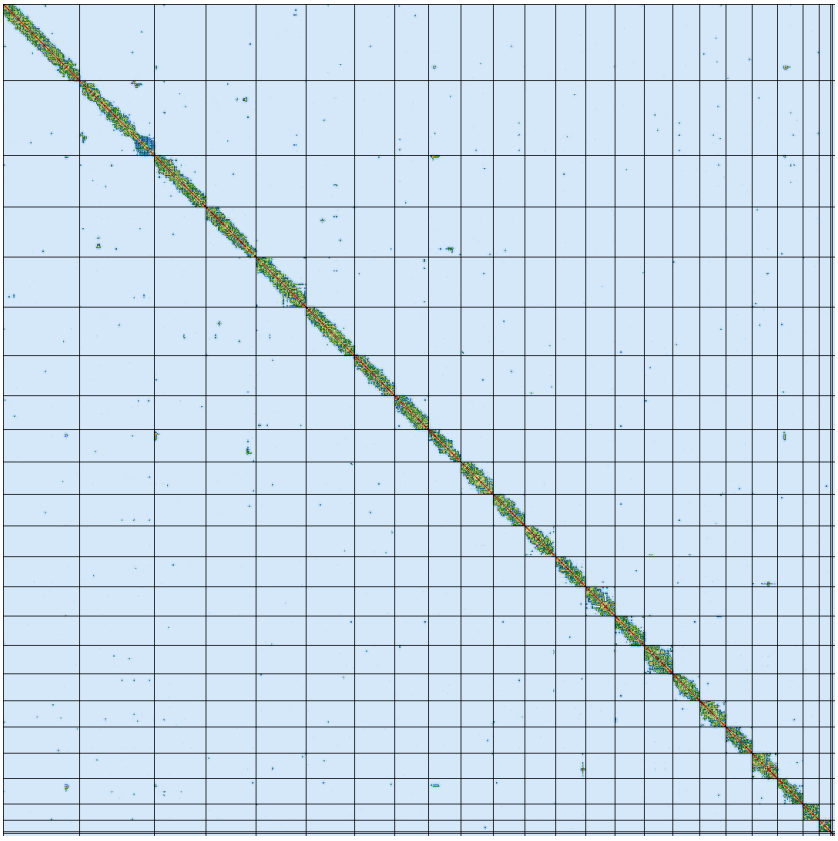
TTAGGG was identified with TelFinder and used to generate the Pretext Telomeres tracks. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. Chromosomes 2, 4, and 9 appear to be linked, but the structure of each chromosome is normal, notably including the presence of telomeres. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	45,880,140	42,819,093
GC %	49.17	49.23
Gaps/Gbp	87.18	46.71
Total gap bp	400	200
Scaffolds	46	25
Scaffold N50	1,664,037	1,766,758
Scaffold L50	9	8
Scaffold L90	21	19
Contigs	50	27
Contig N50	1,664,037	1,766,758
Contig L50	9	8
Contig L90	22	20
QV	47.2034	48.7402
Kmer compl.	16.8076	61.6597
BUSCO sing.	88.3%	89.9%
BUSCO dupl.	5.2%	2.3%
BUSCO frag.	2.9%	2.9%
BUSCO miss.	3.5%	4.9%

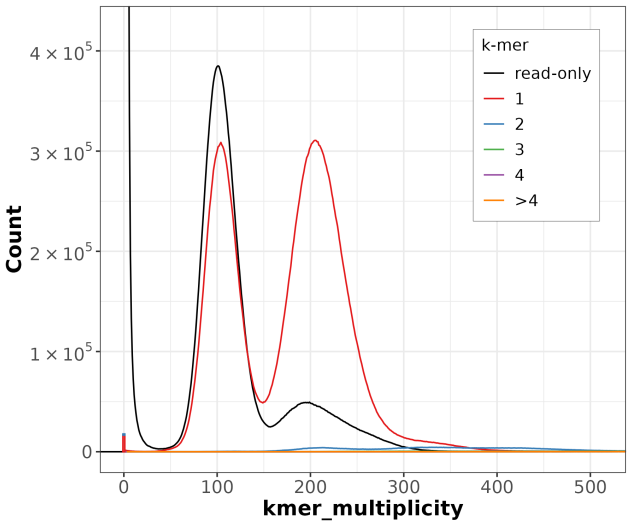
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: bacillariophyta_odb12 (genomes:8, BUSCOs:2944)
BUSCO: 5.8.2 (euk_genome_min, miniprot) / Lineage: bacillariophyta_odb12 (genomes:8, BUSCOs:2944)

HiC contact map of curated assembly

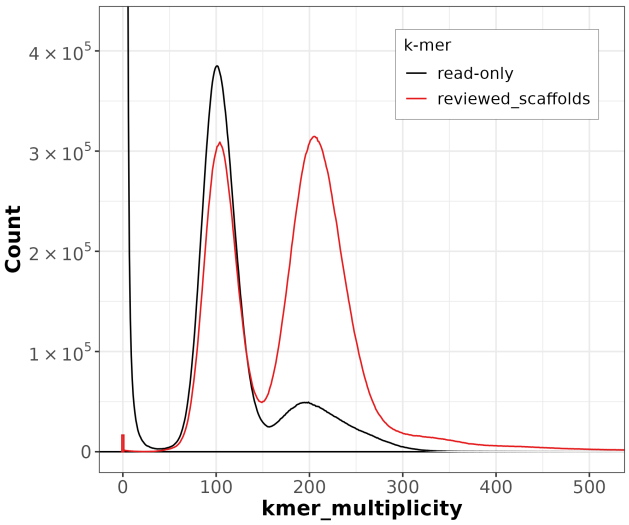


collapsed [LINK]

K-mer spectra of curated assembly

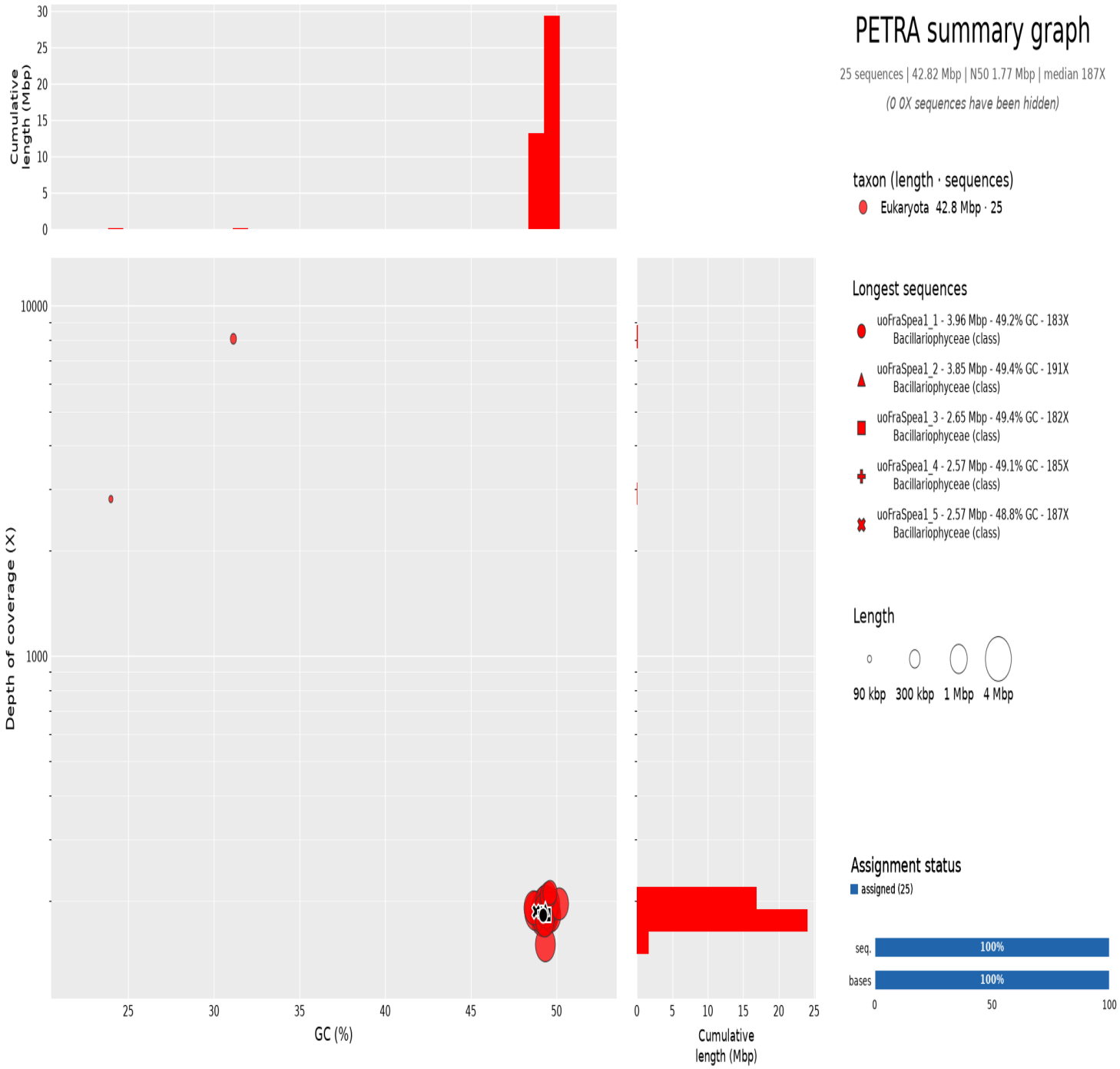


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	351	880

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: Sophie Layac
Affiliation: Genoscope

Date and time: 2026-10-06 09:59:47 CEST