ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	2528228
ToLID	**uoAndSpea1**
Species	Ankylochrysis sp. RCC6043
Class	Pelagophyceae
Order	Pelagomonadales

Genome Traits	Expected	Observed
Haploid size (bp)	75,637,279	68,528,649
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: 6.7.Q47

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

. Observed Haploid Number is different from Expected

. Kmer completeness value is less than 90 for collapsed

Curator notes

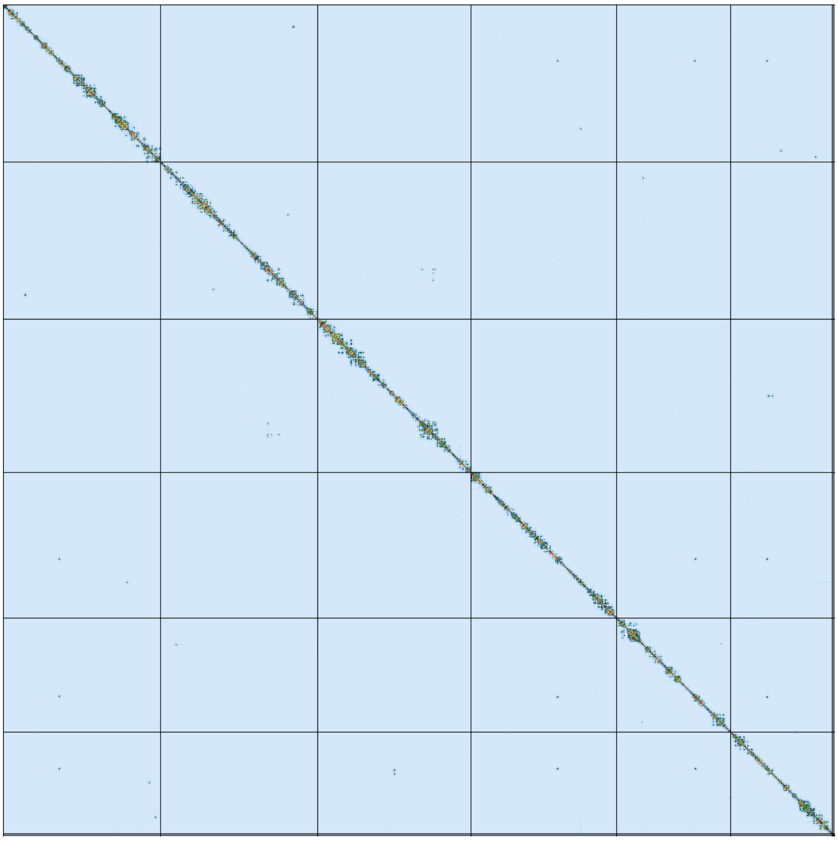
. Interventions/Gb: 204
. Contamination notes: ""
. Other observations: "The assembly of \'Veerella sp. RCC6043\' (uoAndSpea1) is based on 68X PacBio data and 313X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial 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, 122 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 20.715 Mb (with the largest being 4.248 Mb). Additionally, 69 regions totaling 6.782 Mb (with the largest being 1.364 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, 6 contaminant sequences were removed, totaling 0.18Mb (with the largest being 0.068Mb). 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	68,485,336	68,528,649
GC %	59.72	59.7
Gaps/Gbp	248.23	306.44
Total gap bp	1,700	2,800
Scaffolds	12	8
Scaffold N50	12,841,458	12,627,815
Scaffold L50	3	3
Scaffold L90	5	6
Contigs	29	29
Contig N50	4,410,003	4,209,093
Contig L50	6	6
Contig L90	15	16
QV	46.4908	47.2428
Kmer compl.	88.4756	88.6029
BUSCO sing.	85.2%	91.1%
BUSCO dupl.	0.7%	0.9%
BUSCO frag.	7.9%	2.2%
BUSCO miss.	6.2%	5.9%

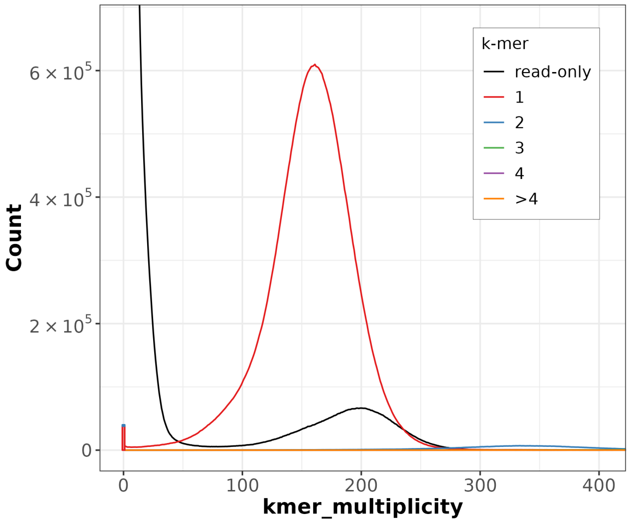
```
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: stramenopiles_odb12 (genomes:55, BUSCOs:697)
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: stramenopiles_odb12 (genomes:55, BUSCOs:697)
```

HiC contact map of curated assembly

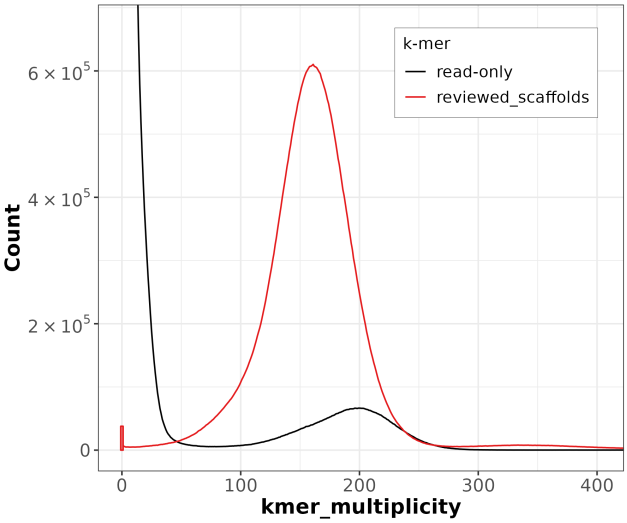


collapsed [LINK]

K-mer spectra of curated assembly

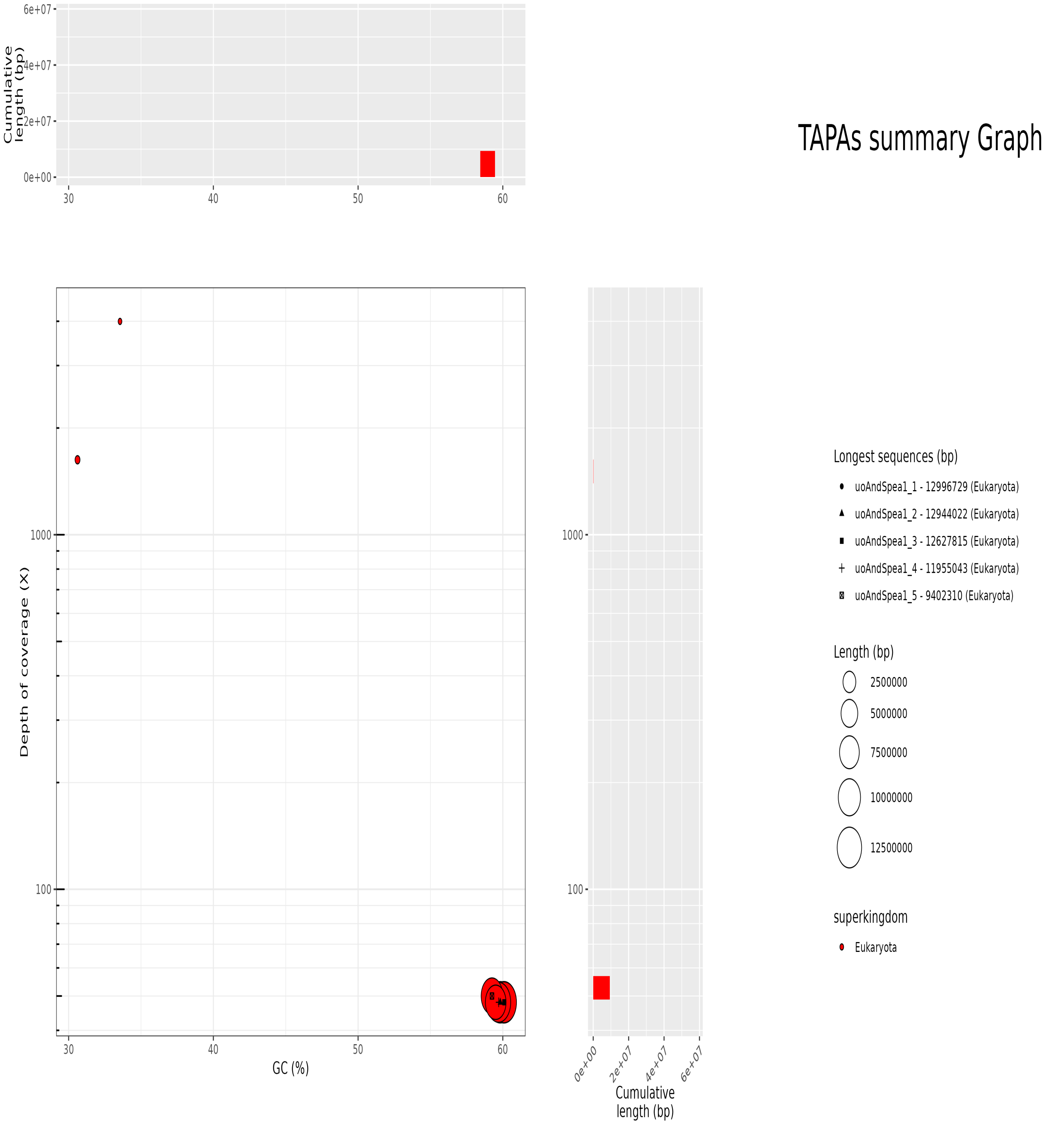


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	65	302

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: Simone Duprat
Affiliation: Genoscope

Date and time: 2025-11-26 11:01:21 CET