ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3451515
ToLID	**uoAurSpea1**
Species	Aureoumbra sp. RCC10688
Class	Pelagophyceae
Order	Sarcinochrysidales

Genome Traits	Expected	Observed
Haploid size (bp)	80,526,491	42,933,339
Haploid Number	4 (source: ancestor)	5
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

Curator notes

. Interventions/Gb: 47
. Contamination notes: ""
. Other observations: "The assembly of Aureoumbra sp. RCC10688 (uoAurSpea1) is based on 116X PacBio data and 602X 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, 108 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 25.228 Mb (with the largest being 4.852 Mb). Additionally, 51 regions totaling 2.029 Mb (with the largest being 0.069 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. The Hi-C contact map contained relatively few long-range interactions, making chromosome validation and orientation more challenging. However, the assembly was highly contiguous, with very few gaps, which facilitated the reconstruction of chromosome-scale scaffolds. Chromosome-scale

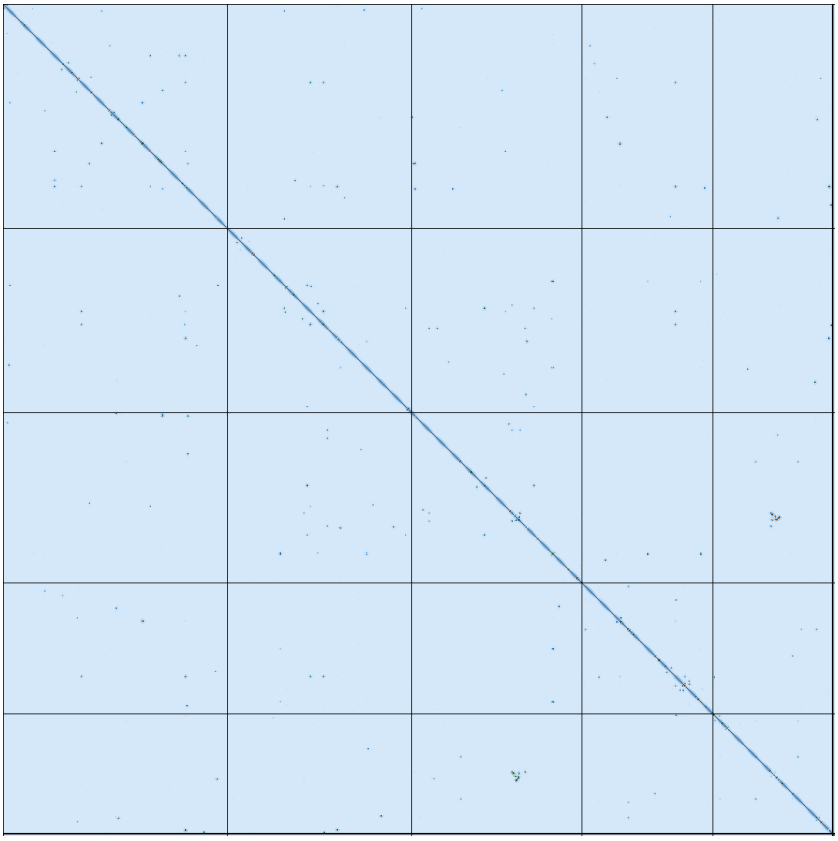
scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	43,008,665	42,933,339
GC %	42.23	42.24
Gaps/Gbp	232.51	46.58
Total gap bp	1,000	300
Scaffolds	391	7
Scaffold N50	1,307,000	8,760,684
Scaffold L50	9	3
Scaffold L90	204	5
Contigs	401	9
Contig N50	1,307,000	8,760,684
Contig L50	9	3
Contig L90	212	5
QV	47.795	48.5649
Kmer compl.	67.4341	67.45
BUSCO sing.	88.5%	89.0%
BUSCO dupl.	2.3%	2.3%
BUSCO frag.	2.4%	2.3%
BUSCO miss.	6.7%	6.5%

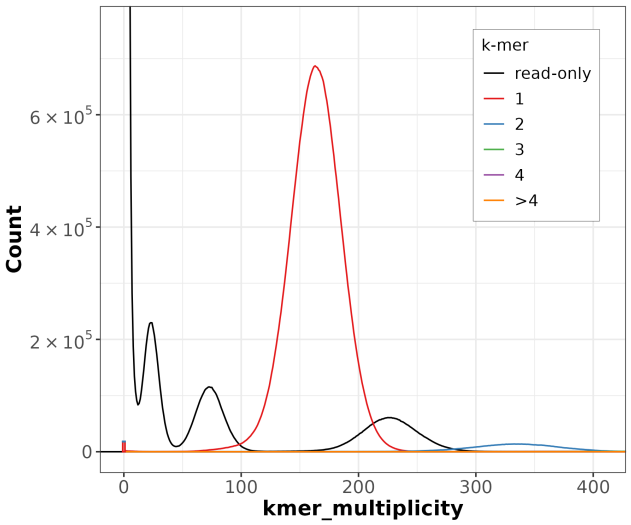
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: stramenopiles_odb12 (genomes:55, BUSCOs:697)
BUSCO: 5.8.2 (euk_genome_min, miniprot) / Lineage: stramenopiles_odb12 (genomes:55, BUSCOs:697)

HiC contact map of curated assembly

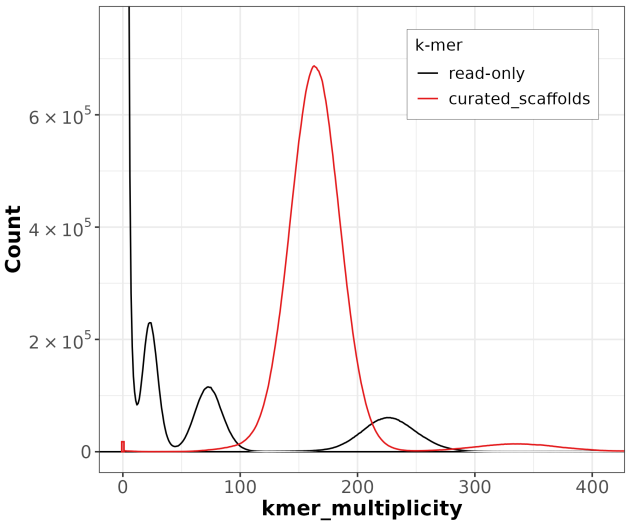


collapsed [LINK]

K-mer spectra of curated assembly

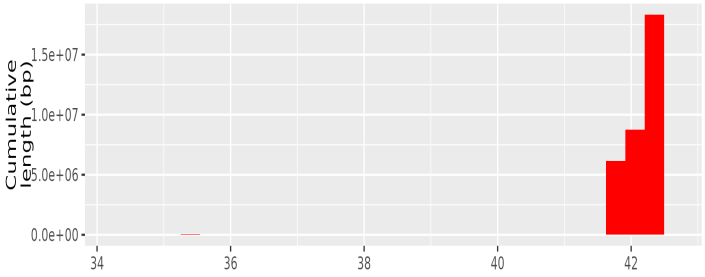


Distribution of k-mer counts per copy numbers found in asm

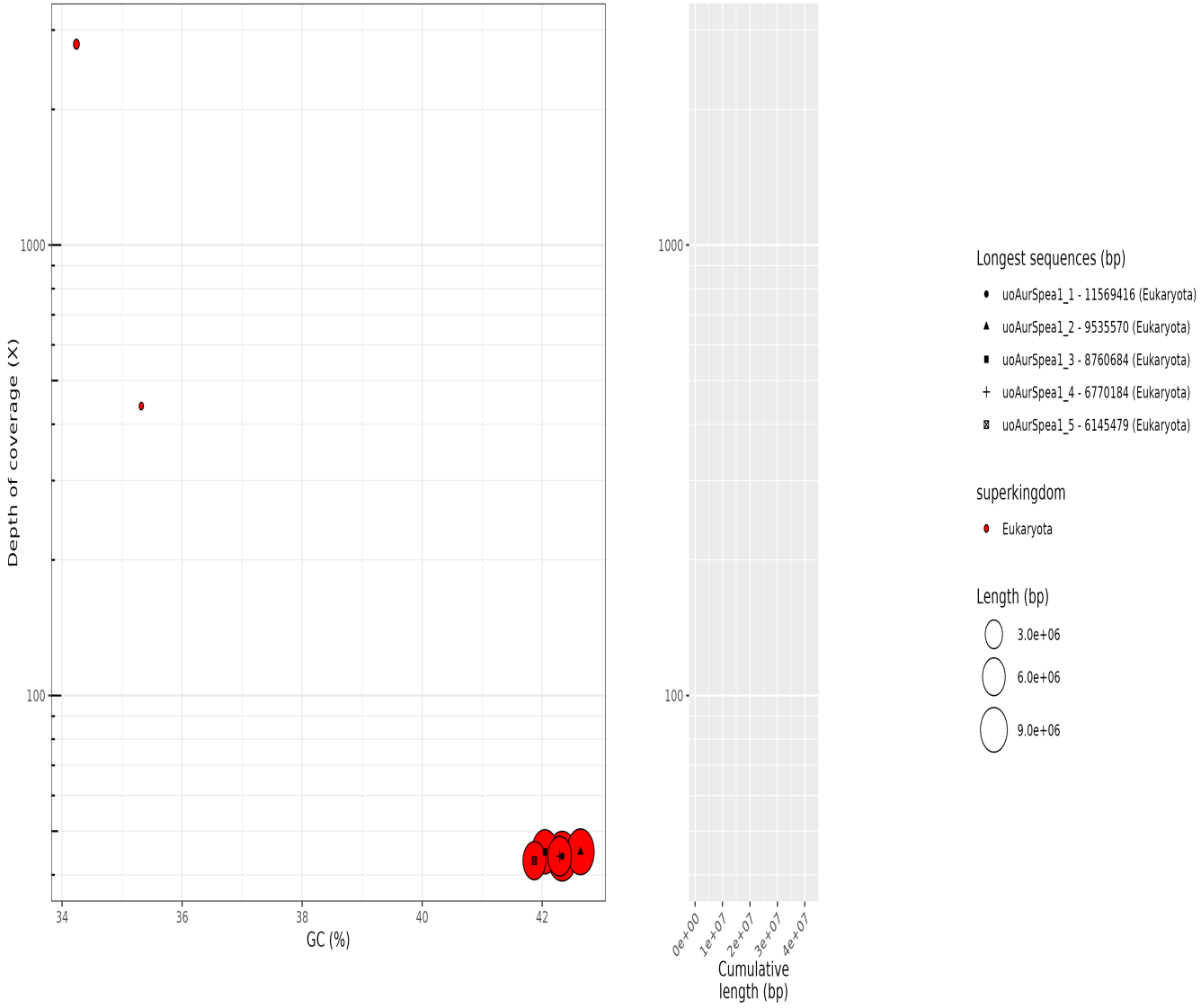


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



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	116	603

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: Jean-Marc Aury
Affiliation: Genoscope

Date and time: 2026-09-10 14:08:20 CEST