

ERGA Assembly Report

v24.10.15

Tags: ATLASEa[INVALID TAG]

TxID	2590921
ToLID	uoMedComi1
Species	Mediolabrus comicus
Class	Thalassiosirophyceae
Order	Thalassiosirales

Genome Traits	Expected	Observed
Haploid size (bp)	38,811,502	42,633,054
Haploid Number	4 (source: ancestor)	24
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 Number is different from Expected
- . Kmer completeness value is less than 90 for collapsed
- . BUSCO single copy value is less than 90% for collapsed
- . BUSCO duplicated value is more than 5% for collapsed
- . Assembly length loss > 3% for collapsed

Curator notes

. Interventions/Gb: 506
. Contamination notes: ""
. Other observations: "The assembly of *Mediolabrus comicus* str. RCC4660 (uoMedComi1) is based on 320.29X PacBio data and 384.09X Arima Hi-C data generated as part of the ATLASEa programme (<https://www.atlasea.fr>). The assembly process included the following steps: initial PacBio+HiC assembly generation with Hifiiasm, removal of contaminant sequences using Context, and Hi-C-based scaffolding with YaHS of each haplotype. In total, 37 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 1.272 Mb (with the largest being 0.119 Mb). Mitochondrial and chloroplastic genomes were assembled using OATK. Finally, Both haplotypes were analyzed and manually improved using Pretext. During manual curation, 66 haplotypic regions were removed, totaling 7.503Mb (with the largest being 0.664Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

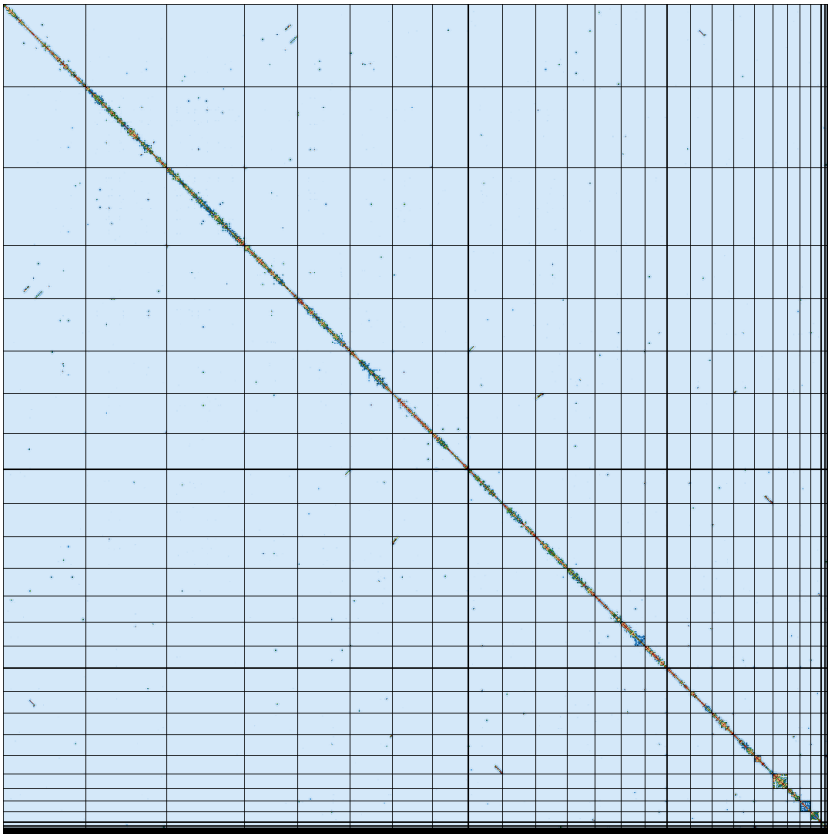
Metrics	Pre-curation collapsed	Curated collapsed
Total bp	49,053,236	42,633,054
GC %	43.27	44.17
Gaps/Gbp	81.54	258.02
Total gap bp	400	2,000
Scaffolds	98	49
Scaffold N50	2,732,976	2,054,507
Scaffold L50	7	7
Scaffold L90	23	19
Contigs	102	60
Contig N50	1,833,879	1,795,044
Contig L50	8	9
Contig L90	26	25
QV	43.8531	48.8913
Kmer compl.	76.2093	72.6001
BUSCO sing.	76.6%	82.9%
BUSCO dupl.	11.8%	5.2%
BUSCO frag.	2.6%	2.9%
BUSCO miss.	8.9%	9.0%

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)

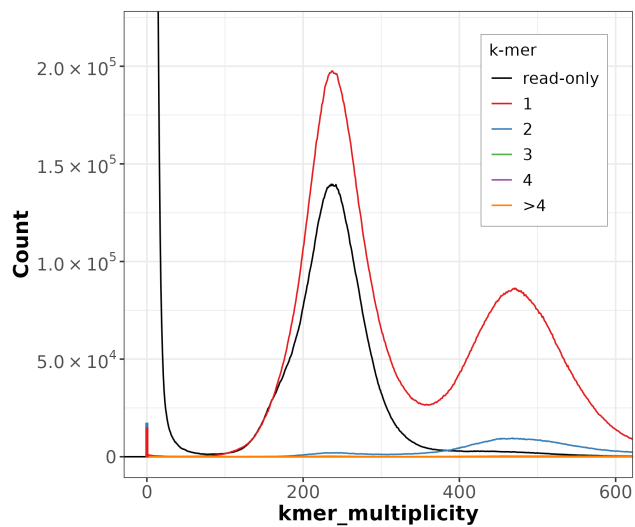
BUSCO: 5.8.2 (euk_genome_min, miniprot) / Lineage: bacillariophyta_odb12 (genomes:8, BUSCOs:2944)

HiC contact map of curated assembly

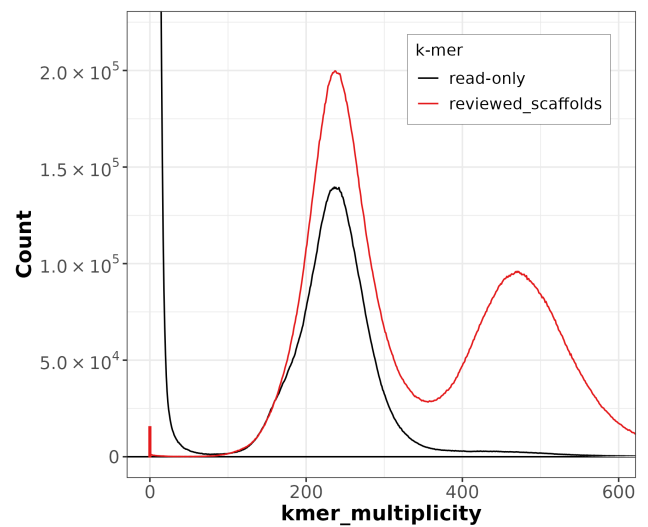


collapsed [\[LINK\]](#)

K-mer spectra of curated assembly

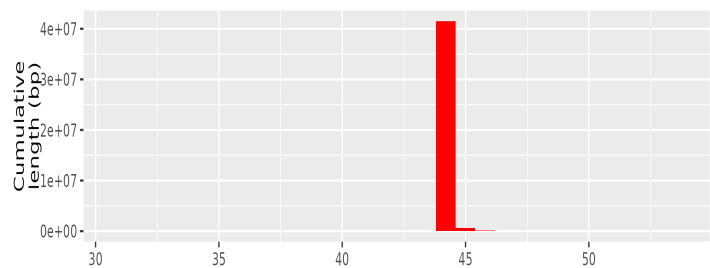


Distribution of k-mer counts per copy numbers found in asm

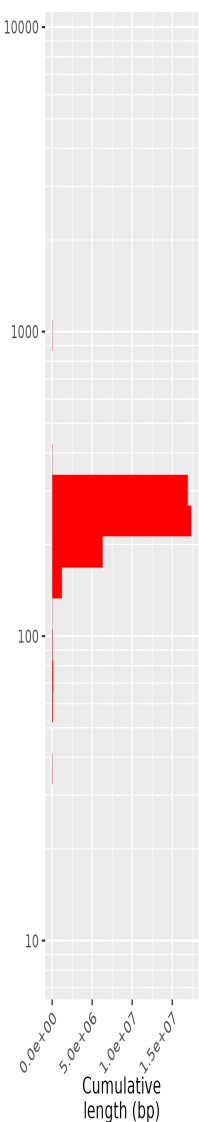
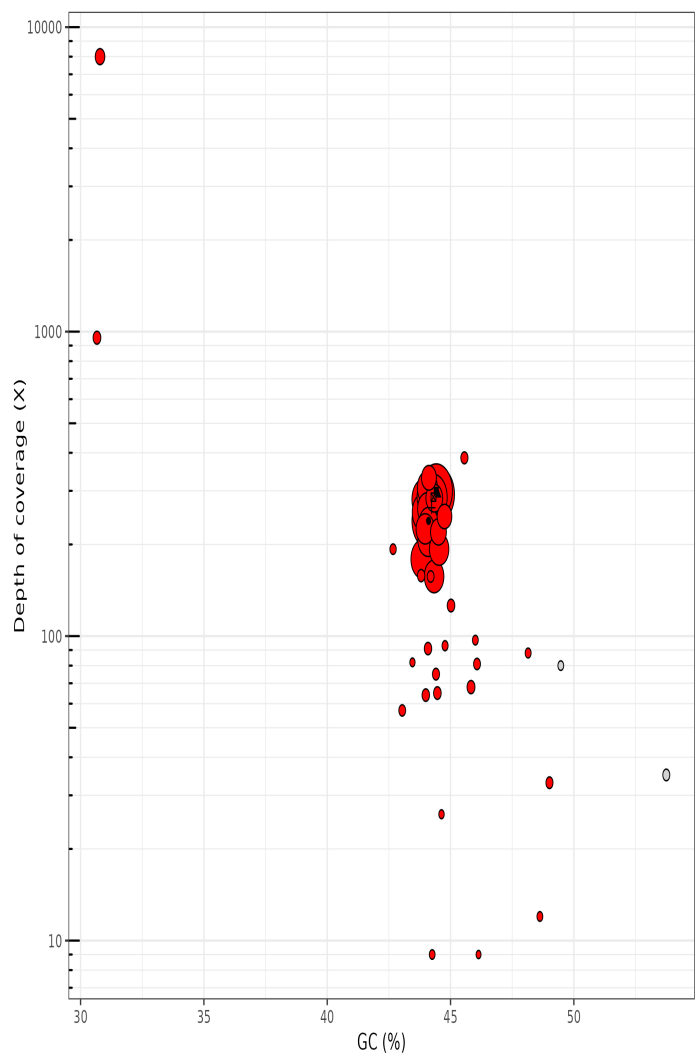


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



Longest sequences (bp)

- uoMedComi1_1 - 4239825 (Eukaryota)
- ▲ uoMedComi1_2 - 4137174 (Eukaryota)
- uoMedComi1_3 - 4010131 (Eukaryota)
- + uoMedComi1_4 - 2709276 (Eukaryota)
- ▣ uoMedComi1_5 - 2699410 (Eukaryota)

Length (bp)

- 1e+06
- 2e+06
- 3e+06
- 4e+06

superkingdom

- Eukaryota
- N/A

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	292	819

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: Emilie Teodori

Affiliation: Genoscope

Date and time: 2026-08-28 16:01:26 CEST