

ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	908974
ToLID	urHelDival
Species	<i>Helminthora divaricata</i>
Class	Florideophyceae
Order	Nemaliales

Genome Traits	Expected	Observed
Haploid size (bp)	252,682,488	46,245,877
Haploid Number	4 (source: ancestor)	24
Ploidy	4 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q40

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
- Observed Ploidy 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: 216
- Contamination notes: ""
- Other observations: "The assembly of *Helminthora divaricata* (urHelDival) is based on 67X PacBio data and 174X 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, 4205 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 197.83 Mb (with the largest being 4.163 Mb). Additionally, 575 regions totaling 6.695 Mb (with the largest being 0.019 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, 3741 supplementary contigs were removed totaling 166 Mb (corresponding to Metazoa) were

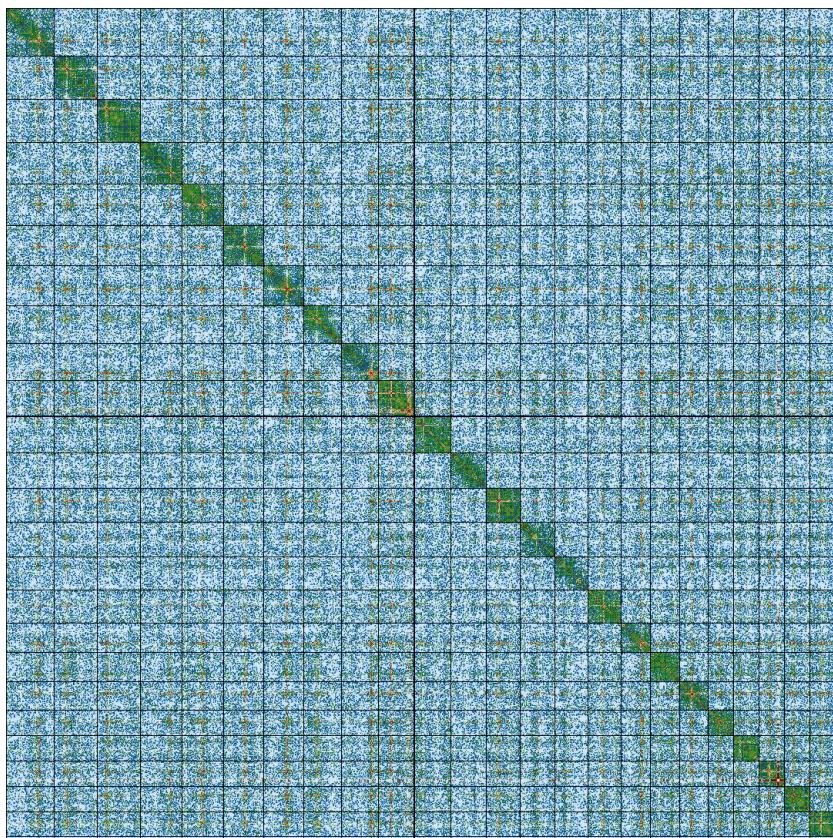
removed. The telomeric pattern CAATAGGG was identified by TelFinder. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	232,502,318	46,245,877
GC %	45.54	55.99
Gaps/Gbp	2,322.56	237.86
Total gap bp	54,000	1,700
Scaffolds	4,790	27
Scaffold N50	90,284	1,989,230
Scaffold L50	335	11
Scaffold L90	2,870	21
Contigs	5,330	38
Contig N50	63,224	1,988,032
Contig L50	485	11
Contig L90	3,410	22
QV	26.3142	40.7865
Kmer compl.	69.8311	31.5951
BUSCO sing.	75.8%	86.7%
BUSCO dupl.	13.6%	1.2%
BUSCO frag.	3.5%	3.9%
BUSCO miss.	7.1%	8.2%

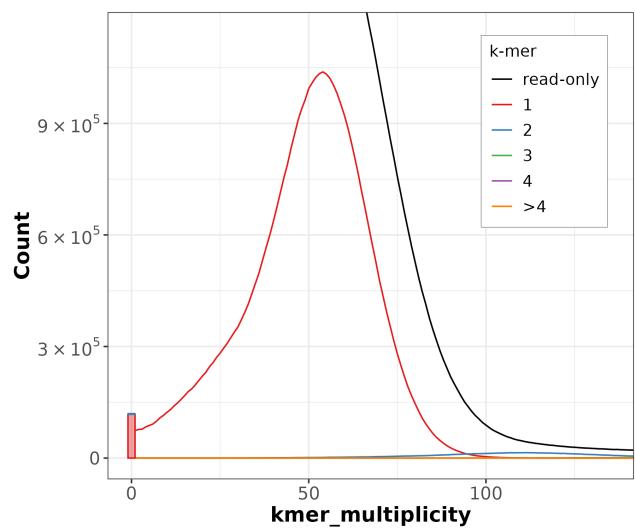
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: rhodophyta_odb12 (genomes:8, BUSCOs:1591)

HiC contact map of curated assembly

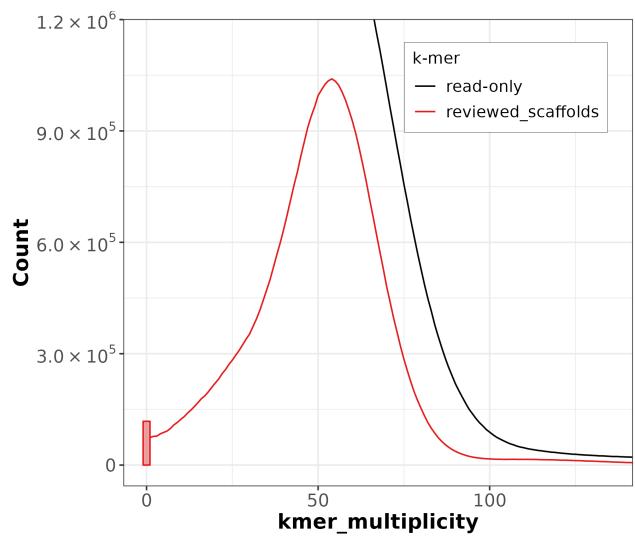


collapsed [\[LINK\]](#)

K-mer spectra of curated assembly

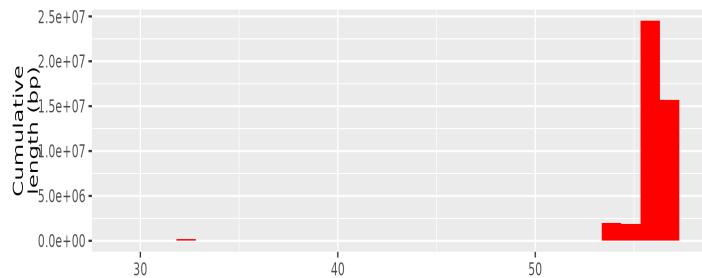


Distribution of k-mer counts per copy numbers found in asm

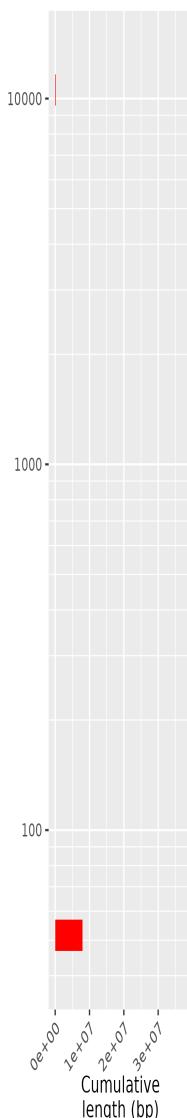
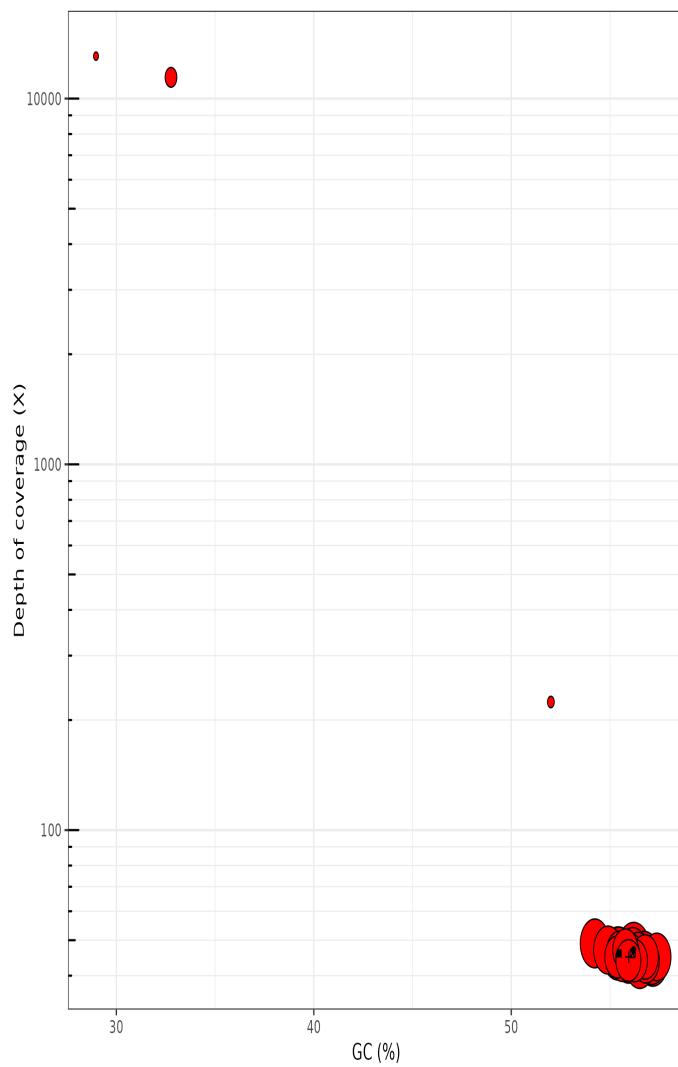


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



Longest sequences (bp)

- urHelDiva1_1 - 2657754 (Eukaryota)
- ▲ urHelDiva1_2 - 2422145 (Eukaryota)
- urHelDiva1_3 - 2372000 (Eukaryota)
- + urHelDiva1_4 - 2352000 (Eukaryota)
- ▣ urHelDiva1_5 - 2280343 (Eukaryota)

Length (bp)

- 500000
- 1000000
- 1500000
- 2000000
- 2500000

superkingdom

- Eukaryota

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	67	174

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: 2026-01-14 09:42:43 CET