

ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	378980
ToLID	cbHomMand10
Species	<i>Homalothecium mandonii</i>
Class	Bryopsida
Order	Hypnales

Genome Traits	Expected	Observed
Haploid size (bp)	565,007,130	416,463,416
Haploid Number	10 (source: ancestor)	10
Ploidy	1 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q74

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 Ploidy 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: 20
- . Contamination notes: ""
- . Other observations: "The assembly of *Homalothecium mandonii* (cbHomMand10) is based on 242X PacBio data and 87X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, <https://www.erga-biodiversity.eu/>) via the Biodiversity Genomics Europe project (BGE, <https://biodiversitygenomics.eu/>). 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, 117,724 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 5,525.781 Mb (with the largest being 12.259 Mb). Additionally, 689 regions totaling 15.958 Mb (with the largest being 0.086 Mb) were identified as haplotypic duplications and removed. Mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 2,088

contaminant sequences were removed, totaling 102,136,336 pb (with the largest being 7,115,672 pb). 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	519,220,046	416,463,416
GC %	40.28	37.92
Gaps/Gbp	1,222.99	60.03
Total gap bp	63,500	2,700
Scaffolds	2,214	20
Scaffold N50	36,023,647	38,796,100
Scaffold L50	5	4
Scaffold L90	227	9
Contigs	2,849	45
Contig N50	16,178,403	23,327,363
Contig L50	10	7
Contig L90	734	21
QV	58.7459	74.2609
Kmer compl.	12.841	12.6967
BUSCO sing.	71.6%	71.6%
BUSCO dupl.	9.4%	9.1%
BUSCO frag.	4.4%	4.3%
BUSCO miss.	14.6%	15.1%

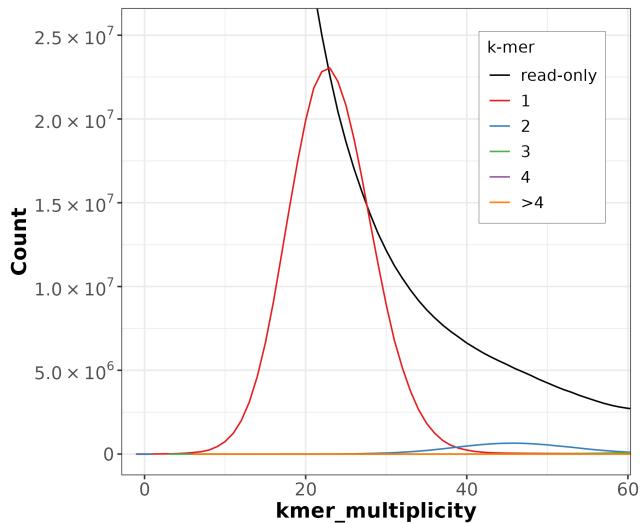
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: embryophyta_odb12 (genomes:78, BUSCOS:2026)

HiC contact map of curated assembly

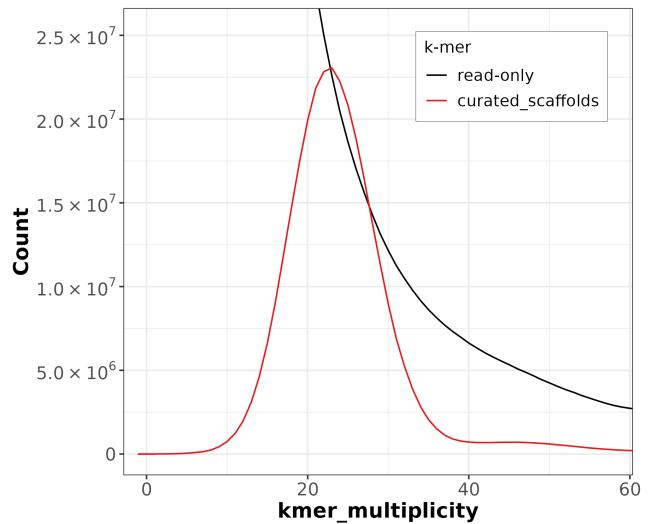


collapsed [\[LINK\]](#)

K-mer spectra of curated assembly

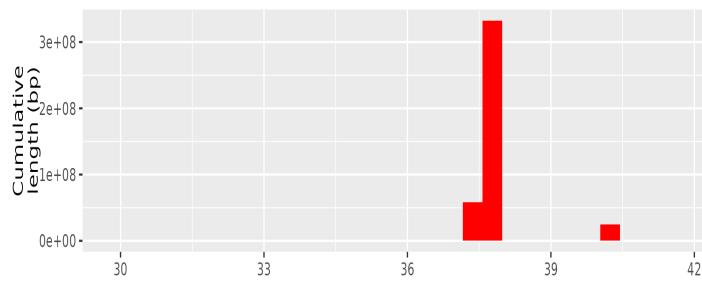


Distribution of k-mer counts per copy numbers found in asm

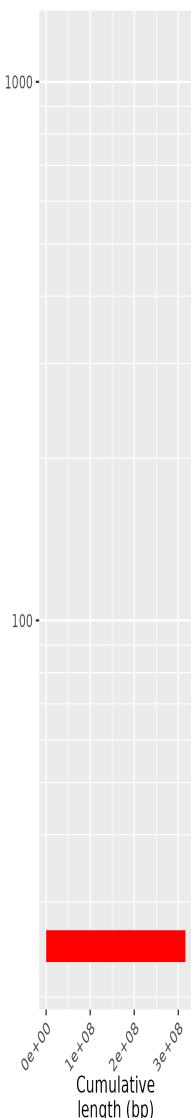
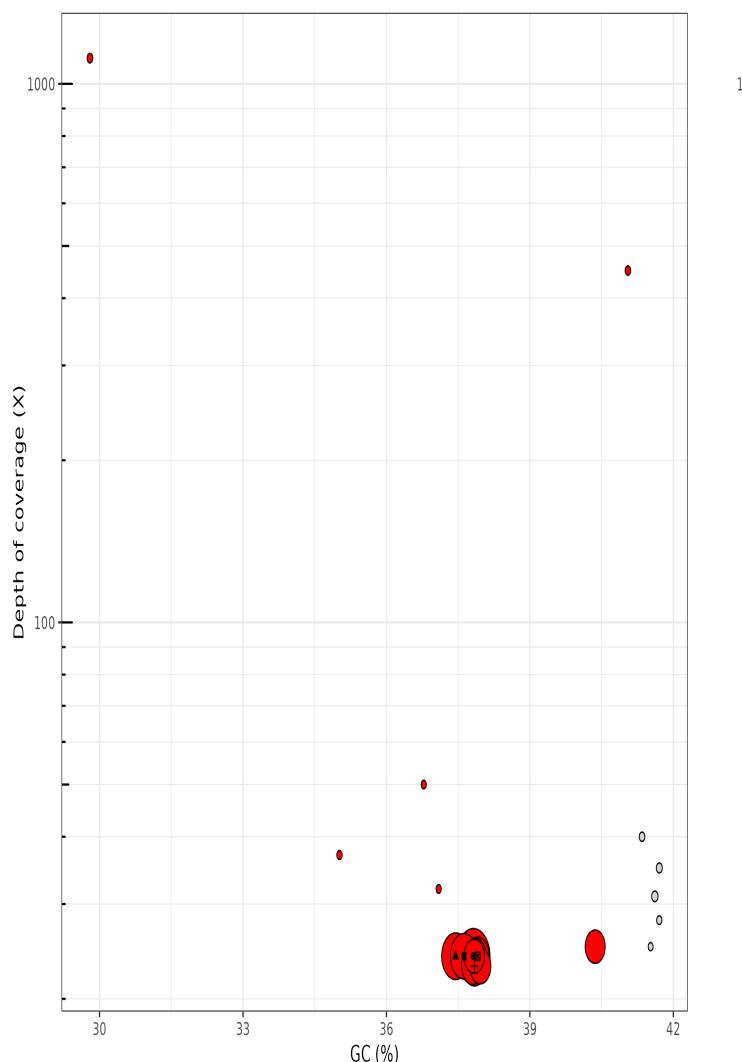


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



Length (bp)

- 2e+07
- 4e+07
- 6e+07
- 8e+07

Longest sequences (bp)

- cbHomMand10_1 - 86847434 (Eukaryota)
- ▲ cbHomMand10_2 - 57970191 (Eukaryota)
- cbHomMand10_3 - 54494190 (Eukaryota)
- + cbHomMand10_4 - 38796100 (Eukaryota)
- ▣ cbHomMand10_5 - 36023647 (Eukaryota)

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	184	87

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: Lola Demirdjian
Affiliation: Genoscope

Date and time: 2026-01-30 01:51:18 CET