ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	522324
ToLID	**cbOrtDiap4**
Species	Orthotrichum diaphanum
Class	Bryopsida
Order	Orthotrichales

Genome Traits	Expected	Observed
Haploid size (bp)	378,051,659	300,854,170
Haploid Number	10 (source: direct)	10
Ploidy	1 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q45

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
. More than 1000 gaps/Gbp for collapsed

Curator notes

. Interventions/Gb: 170
. Contamination notes: ""
. Other observations: "The assembly of ORTHOTRICHUM DIAPHANUM (cbOrtDiap4) is based on 90X PacBio data and 172X 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 Nextdenovo, removal of contaminant sequences using Context, and Hi-C-based scaffolding with YaHS. In total, 312 contigs were identified as contaminants (bacterial and virus), totaling 102,717,377 pb (with the largest being 1,315,266 pb). Additionally, 215 regions totaling 16,655,681 pb (with the largest being 6,394,324 pb) were identified as haplotypic duplications and removed. Additionally, 209 regions totaling 12,347,622 pb (with the largest being 297,312 pb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using OATK. Finally, the primary

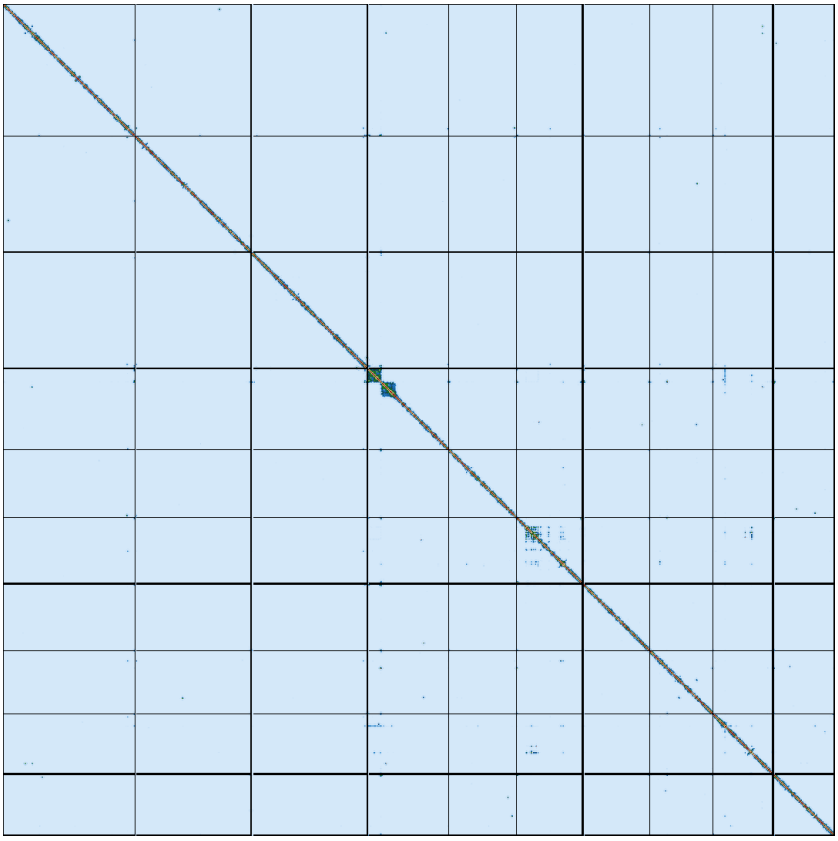
assembly was analyzed and manually improved using Pretext. During manual curation, 14 scaffolds were identified as contaminants and removed. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size and the telomere pattern found is TTAGG. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	301,715,765	300,854,170
GC %	37.25	37.27
Gaps/Gbp	1,000.94	1,070.29
Total gap bp	30,200	37,300
Scaffolds	58	32
Scaffold N50	29,060,192	29,290,605
Scaffold L50	4	4
Scaffold L90	9	9
Contigs	360	354
Contig N50	1,432,891	1,432,891
Contig L50	60	60
Contig L90	200	199
QV	45.275	45.2836
Kmer compl.	87.2371	87.1535
BUSCO sing.	72.0%	72.0%
BUSCO dupl.	7.3%	7.3%
BUSCO frag.	4.1%	4.0%
BUSCO miss.	16.6%	16.7%

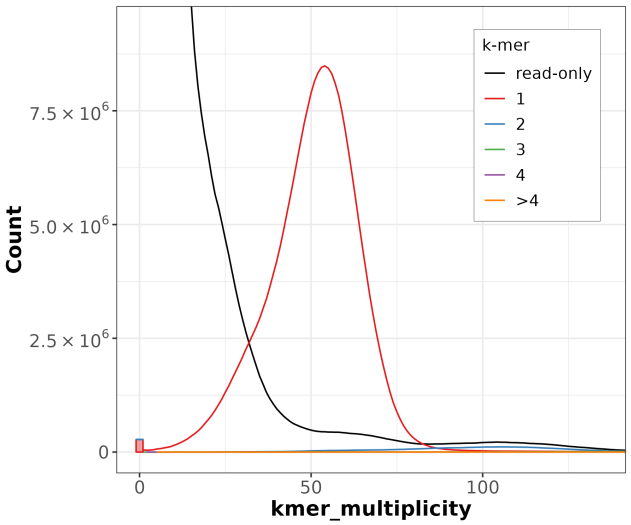
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: embryophyta_odb12 (genomes:78, BUSCOs:2026)

HiC contact map of curated assembly

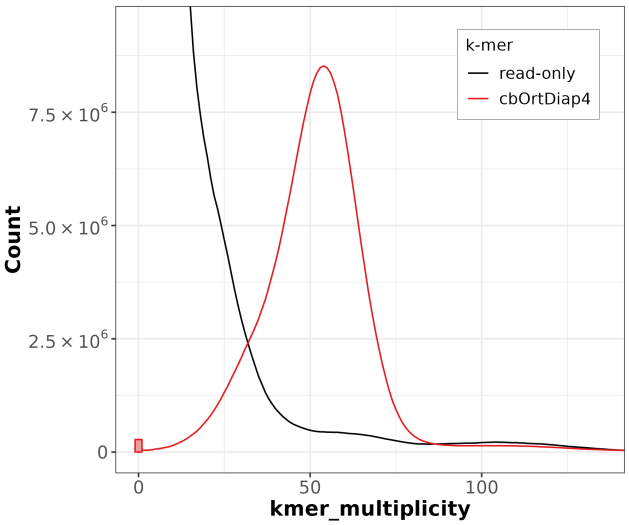


collapsed [LINK]

K-mer spectra of curated assembly

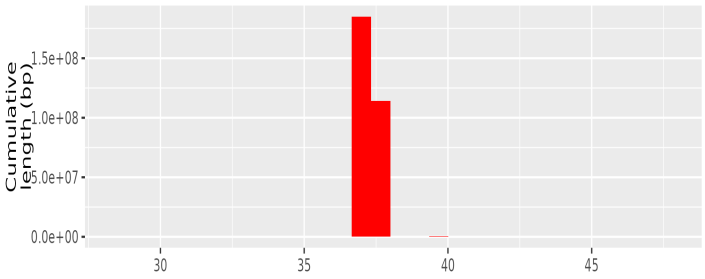


Distribution of k-mer counts per copy numbers found in asm

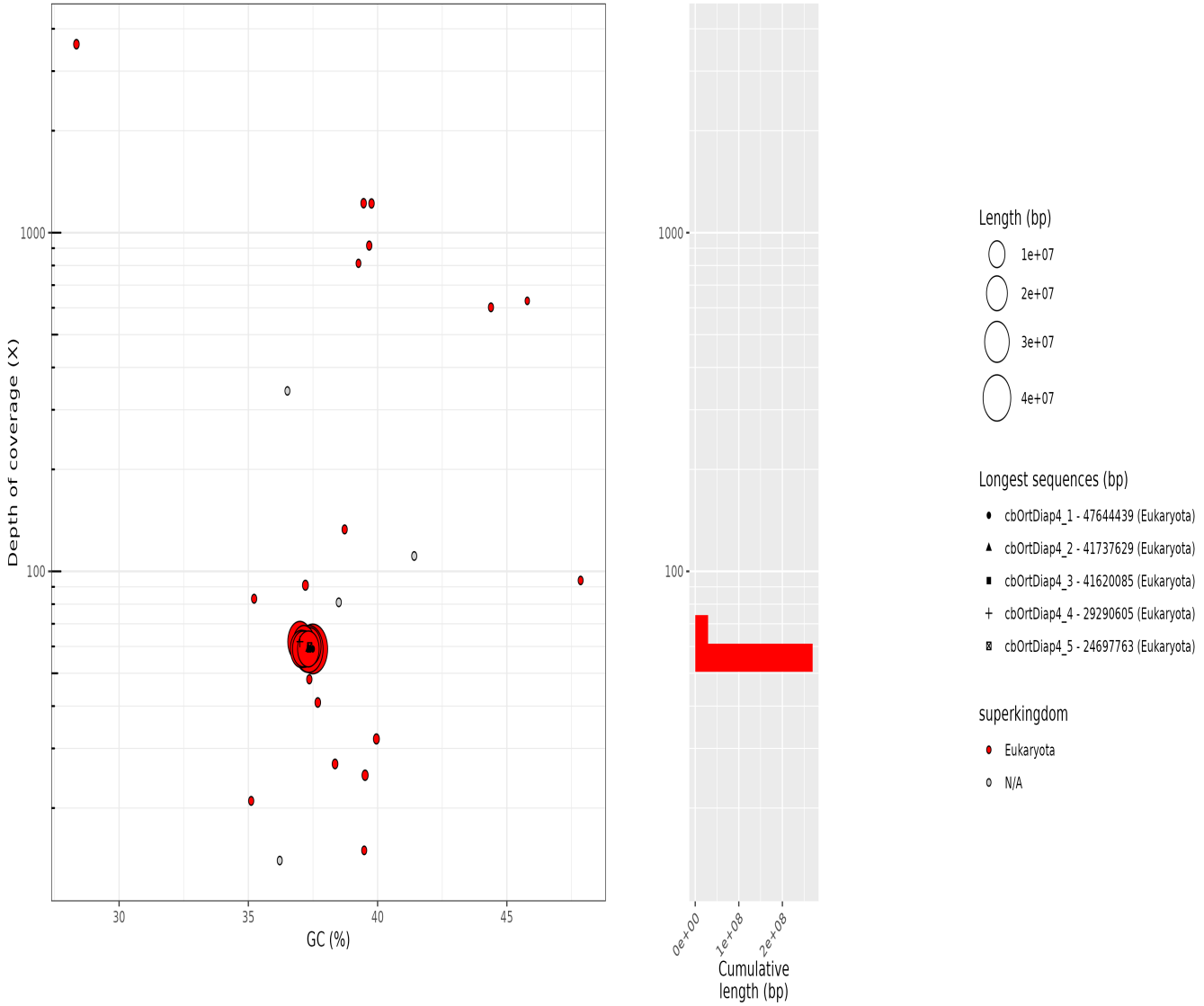


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



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	90	172

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: 2025-11-01 04:55:44 CET