

# ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	754874
ToLID	<b>drEbeCret1</b>
Species	Ebenus cretica
Class	Magnoliopsida
Order	Fabales

Genome Traits	Expected	Observed
Haploid size (bp)	988,901,153	942,902,264
Haploid Number	7 (source: direct)	6
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

## EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q59

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

### Curator notes

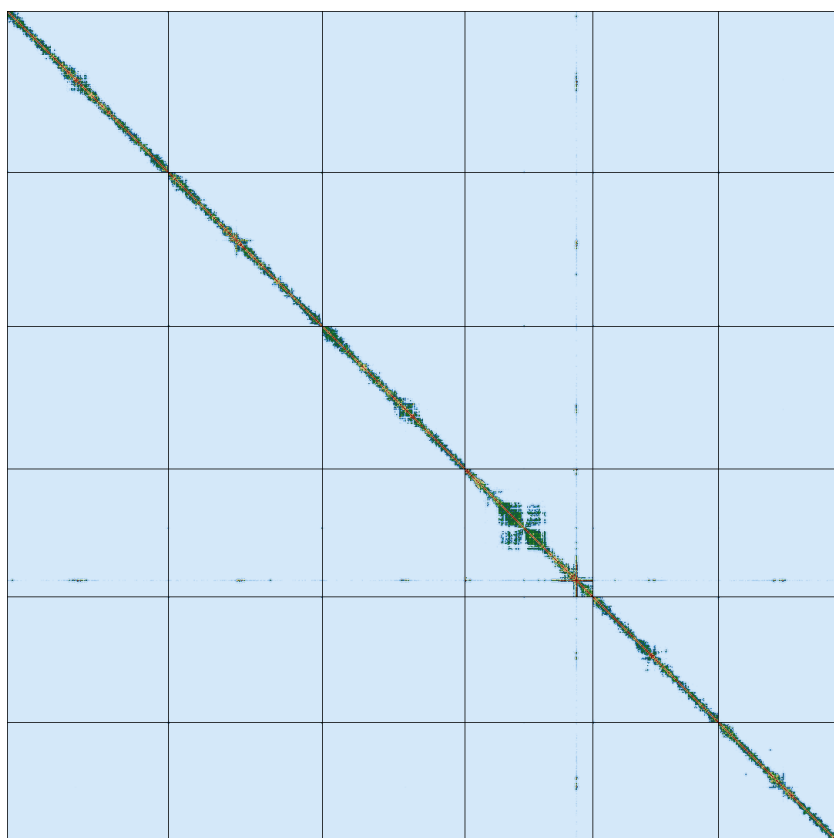
- . Interventions/Gb: 3
- . Contamination notes: ""
- . Other observations: "The assembly of Ebenus cretica (drEbeCret1.4) is based on 75X of PacBio data and Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, <https://www.erga-biodiversity.eu/>) via the Biodiversity Genomics Europe project (BGE, <https://biodiversitygenomics.eu/>). The assembly process included the following steps: initial PacBio assembly generation with Hifiiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge\_dups, and Hi-C-based scaffolding with YaHS. In total, 34 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 1.5 Mb (with the largest being 0.167 Mb). Additionally, 811 regions totaling 64 Mb were identified as haplotypic duplications and removed. Mitochondrial and chloroplast genomes was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 16 haplotypic regions were removed, totaling 2 Mb, respectively (with the largest being 0.8 Mb). 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	946,087,149	942,902,264
GC %	34.54	34.52
Gaps/Gbp	27.48	12.73
Total gap bp	2,600	1,300
Scaffolds	69	18
Scaffold N50	174,068,480	161,678,577
Scaffold L50	2	3
Scaffold L90	5	6
Contigs	95	30
Contig N50	92,227,207	92,227,207
Contig L50	4	4
Contig L90	10	10
QV	59.0722	59.3862
Kmer compl.	89.35	89.3305
BUSCO sing.	64.3%	64.4%
BUSCO dupl.	34.6%	34.5%
BUSCO frag.	0.1%	0.1%
BUSCO miss.	1.0%	1.0%

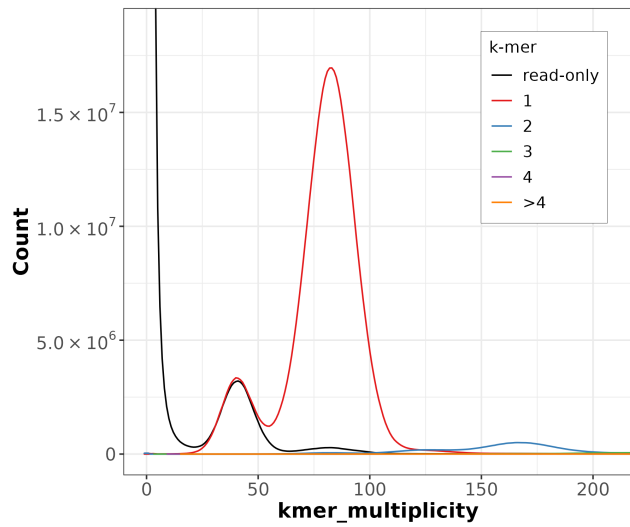
BUSCO: 5.4.3 (euk\_genome\_met, metaeuk) / Lineage: embryophyta\_odb10 (genomes:50, BUSCOs:1614)

# HiC contact map of curated assembly

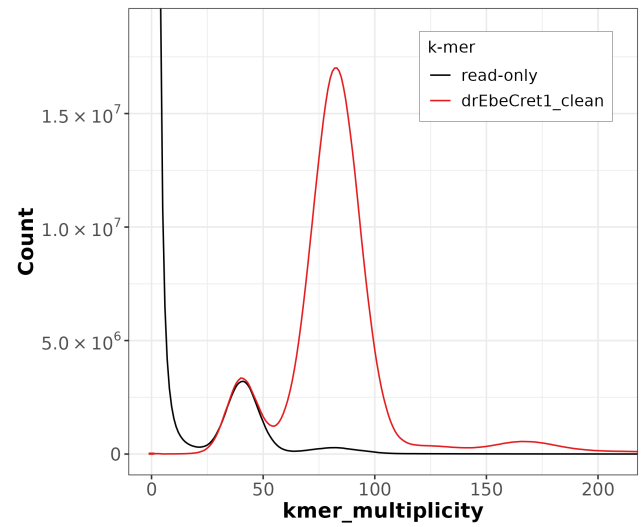


collapsed [\[LINK\]](#)

# K-mer spectra of curated assembly

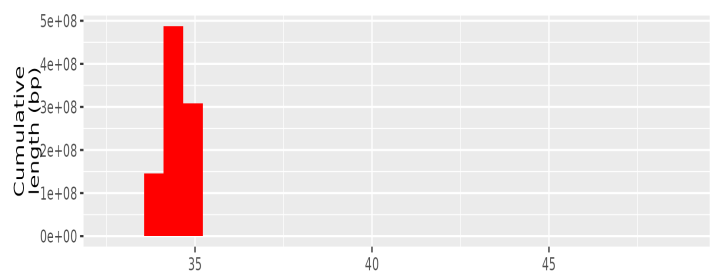


Distribution of k-mer counts per copy numbers found in asm

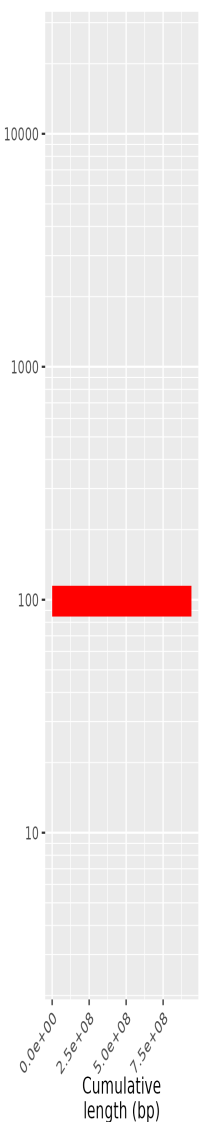
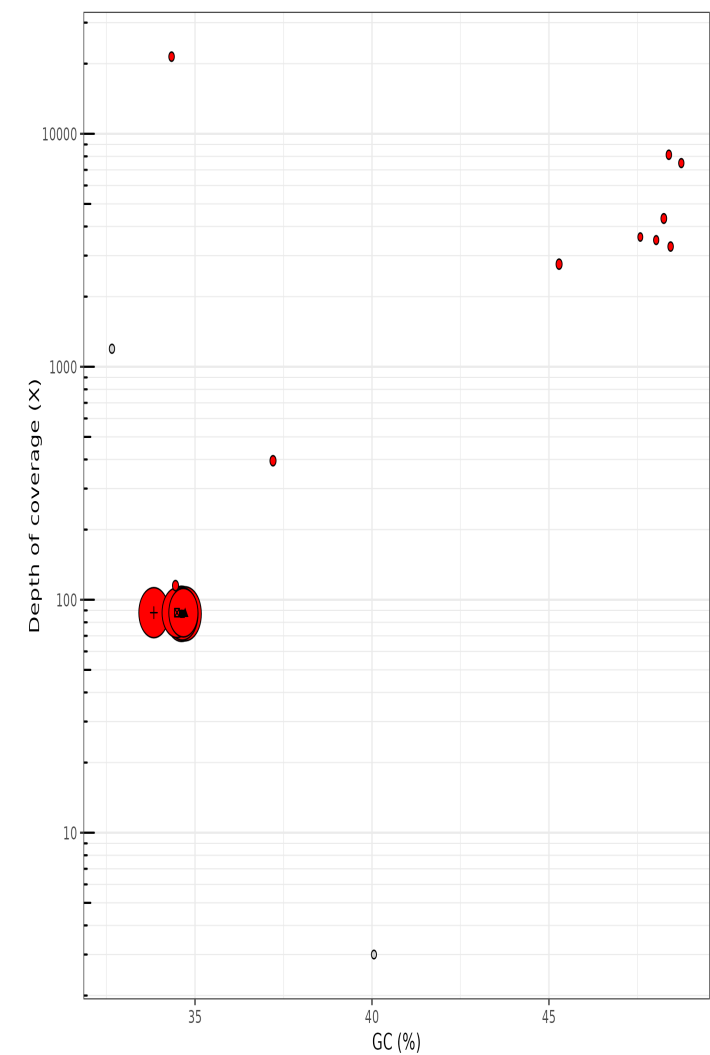


Distribution of k-mer counts coloured by their presence in reads/assemblies

# Post-curation contamination screening



TAPAs summary Graph



Longest sequences (bp)

- SUPER\_1 - 183158860 (Eukaryota)
- ▲ SUPER\_2 - 174068580 (Eukaryota)
- SUPER\_3 - 161678577 (Eukaryota)
- + SUPER\_4 - 145209119 (Eukaryota)
- ▣ SUPER\_5 - 142517981 (Eukaryota)

Length (bp)

- 5.0e+07
- 1.0e+08
- 1.5e+08

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	PACBIO Hifi	Omnic
Coverage	75	43

# 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: Adama Ndar

Affiliation: Genoscope

Date and time: 2025-01-23 23:16:25 CET