ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	317547
ToLID	**jaEunCavo1**
Species	Eunicella cavolini
Class	Anthozoa
Order	Malacalcyonacea

Genome Traits	Expected	Observed
Haploid size (bp)	411,546,873	489,969,558
Haploid Number	6 (source: ancestor)	16
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q61

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

Curator notes

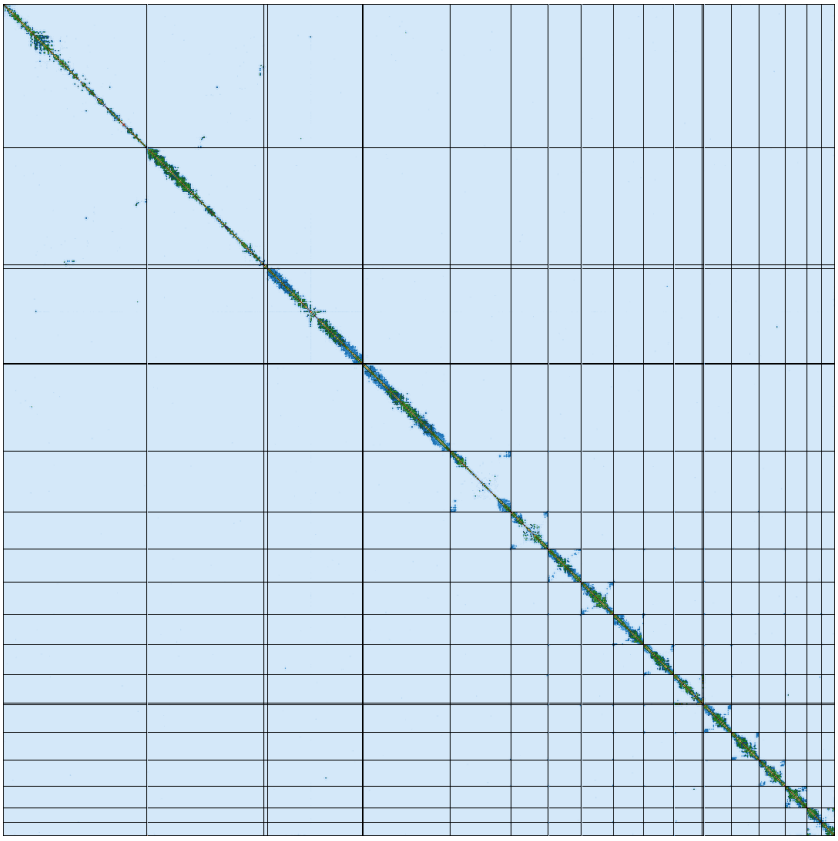
. Interventions/Gb: 516
. Contamination notes: ""
. Other observations: "The assembly of EUNICELLA CAVOLINI (jaEunCavo1) is based on 53X PacBio data and 130X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics 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, 37 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 1 Mb (with the largest being 88,287 pb). Additionally, 208 regions totaling 168 Mb were identified as haplotypic duplications and removed. However, all sequences larger than 1 Mb were added to the assembly, with purge_dups removing too many contigs needed for the scaffolding. The mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 12 haplotypic regions were removed, totaling 27 Mb (with the largest being 1.8 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	372,987,233	489,969,558
GC %	37.36	37.34
Gaps/Gbp	171.59	230.63
Total gap bp	6,400	15,300
Scaffolds	66	46
Scaffold N50	24,525,469	51,102,291
Scaffold L50	6	4
Scaffold L90	15	13
Contigs	130	159
Contig N50	8,615,778	7,689,143
Contig L50	13	20
Contig L90	50	69
QV	61.6191	61.5176
Kmer compl.	63.0875	77.4335
BUSCO sing.	79.7%	87.6%
BUSCO dupl.	1.3%	1.0%
BUSCO frag.	5.6%	5.6%
BUSCO miss.	13.4%	5.8%

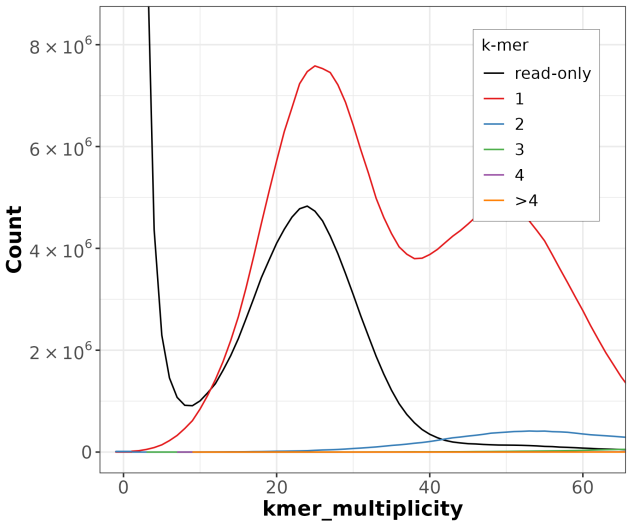
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: metazoa_odb10 (genomes:65, BUSCOs:954)

HiC contact map of curated assembly

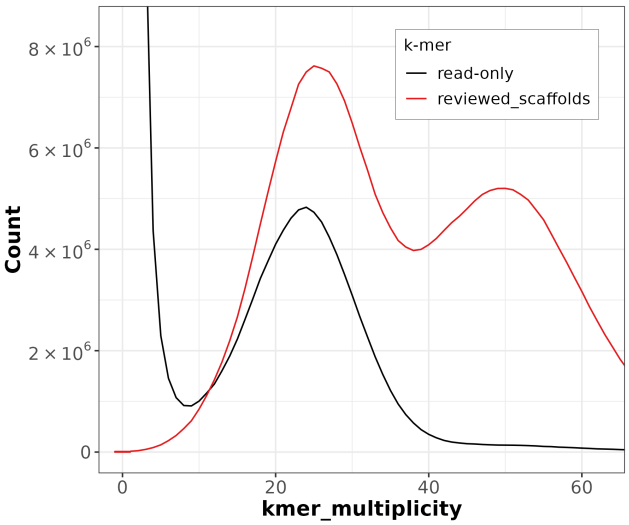


collapsed [LINK]

K-mer spectra of curated assembly

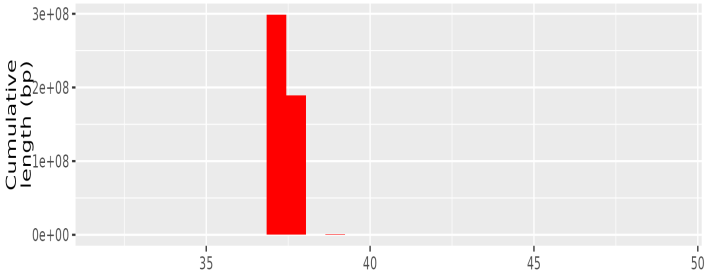


Distribution of k-mer counts per copy numbers found in asm

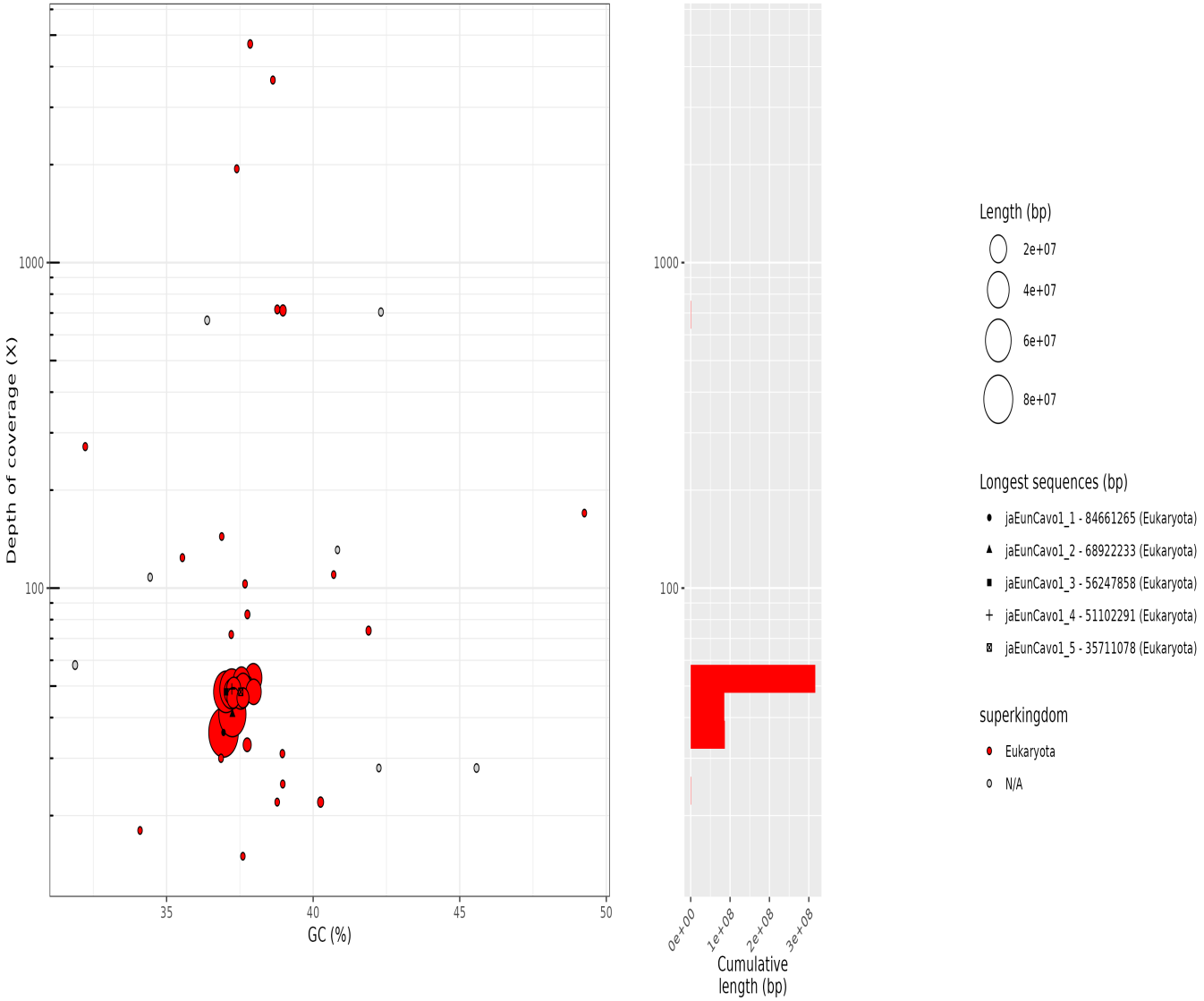


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	PACBIO Hifi	Arima
Coverage	53	142

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: 2025-02-28 00:46:58 CET