

# ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1502934
ToLID	<b>jaLepSarm1</b>
Species	Leptogorgia sarmentosa
Class	Anthozoa
Order	Malacalcyonacea

Genome Traits	Expected	Observed
Haploid size (bp)	195,449,275	210,376,152
Haploid Number	6 (source: ancestor)	36
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

## EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.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

### Curator notes

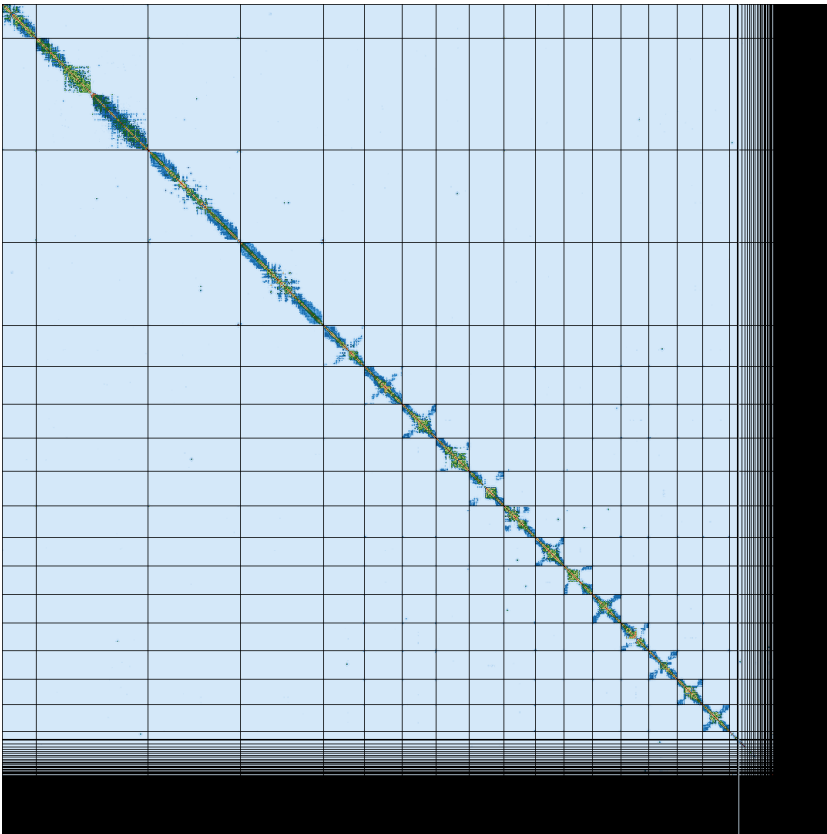
- . Interventions/Gb: 180
- . Contamination notes: ""
- . Other observations: "The assembly of *Leptogorgia sarmentosa* (jaLepSarm1) is based on 130X PacBio data and 200X Arima Hi-C data generated as part of the ATLASea programme (<https://www.atlasea.fr>). The assembly process included the following steps: generation of PacBio and Hi-C data assembly with Hifiiasm, removal of contaminant sequences using Context, and Hi-C-based scaffolding with YaHS. In total, 39 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 18 Mb (with the largest being 9 Mb). The mitochondrial genome was assembled using OATK. A manually phased assembly, based on the Hi-C signal, was generated from a jointly curated primary and alternate assembly using Pretext. For each pair of chromosomes, the longest was retained as the primary haplotype, and the shorter was designated as the alternate haplotype. 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	213,877,839	210,376,152
GC %	37.94	38.3
Gaps/Gbp	140.27	370.76
Total gap bp	3,000	10,400
Scaffolds	67	293
Scaffold N50	17,340,999	8,639,662
Scaffold L50	5	7
Scaffold L90	16	23
Contigs	97	371
Contig N50	4,682,911	4,255,000
Contig L50	15	14
Contig L90	40	70
QV	50.7543	59.8739
Kmer compl.	74.2773	70.2304
BUSCO sing.	86.9%	87.1%
BUSCO dupl.	1.8%	1.9%
BUSCO frag.	4.6%	4.8%
BUSCO miss.	6.7%	6.2%

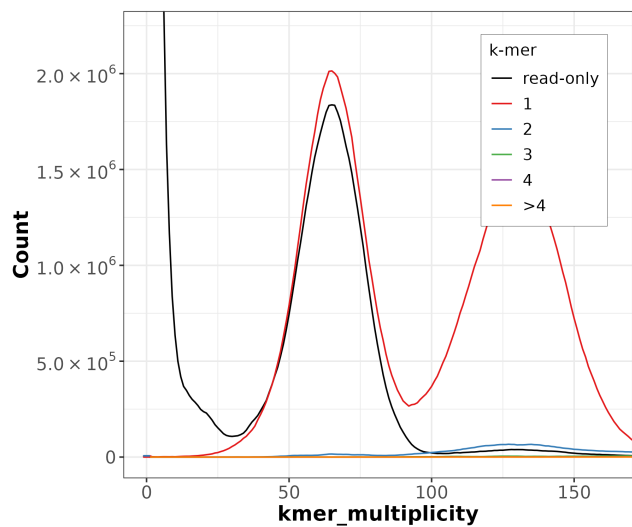
BUSCO: 5.4.3 (euk\_genome\_met, metaeuk) / Lineage: metazoa\_odb10 (genomes:65, BUSCOs:954)

# HiC contact map of curated assembly

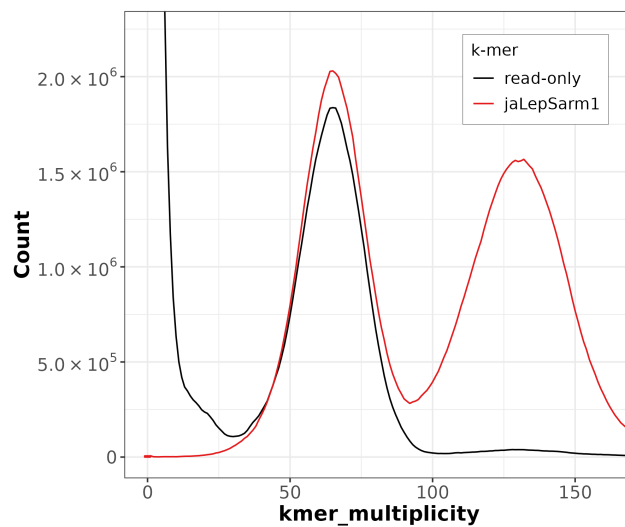


collapsed [\[LINK\]](#)

# K-mer spectra of curated assembly

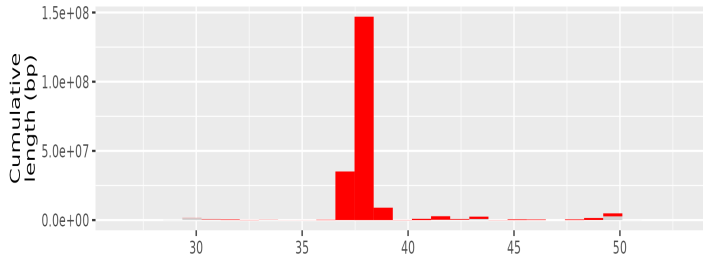


Distribution of k-mer counts per copy numbers found in asm



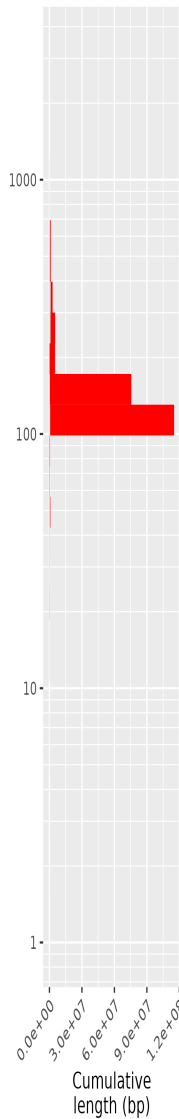
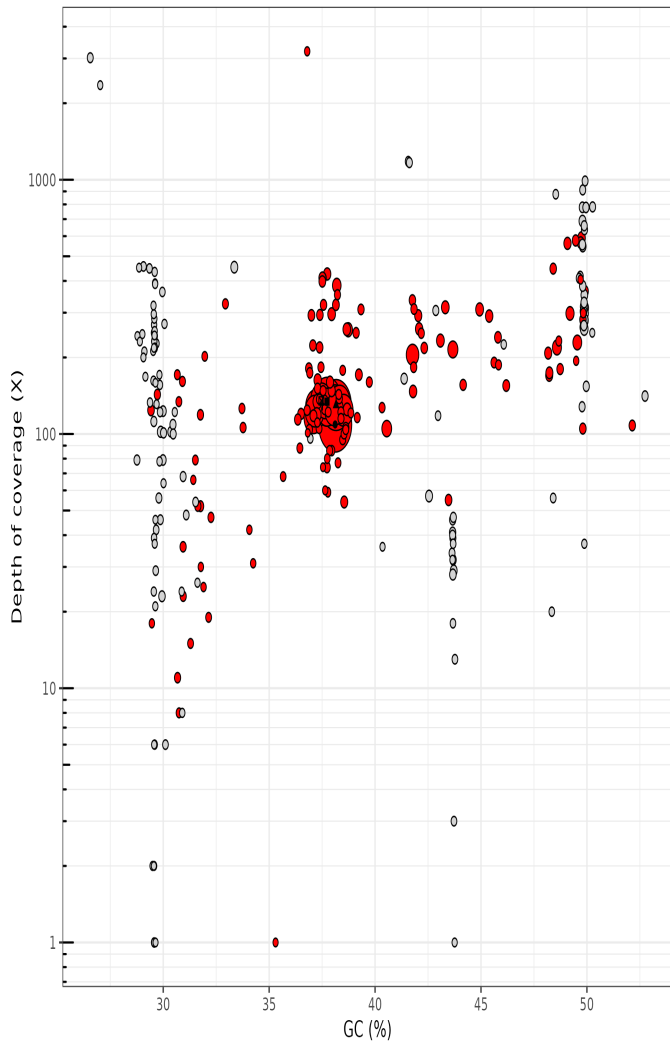
Distribution of k-mer counts coloured by their presence in reads/assemblies

# Post-curation contamination screening



## TAPAs summary Graph

(8 0X contigs have been hidden)



Length (bp)

○ 1e+07

○ 2e+07

superkingdom

● Eukaryota

○ N/A

Longest sequences (bp)

● SUPER\_4 - 28281438 (Eukaryota)

▲ SUPER\_5 - 23356100 (Eukaryota)

■ SUPER\_7 - 21093100 (Eukaryota)

+ SUPER\_9 - 10193100 (Eukaryota)

■ SUPER\_12 - 9714215 (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	PACBIO Hifi	Arima
Coverage	130	200

# 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: Sophie Mangenot

Affiliation: Genoscope

Date and time: 2024-12-17 15:38:54 CET