ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	429181
ToLID	**xbLaeCras1.1**
Species	Laevicardium crassum
Class	Bivalvia
Order	Cardiida

Genome Traits	Expected	Observed
Haploid size (bp)	1,367,916,027	1,507,413,752
Haploid Number	26 (source: ancestor)	19
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

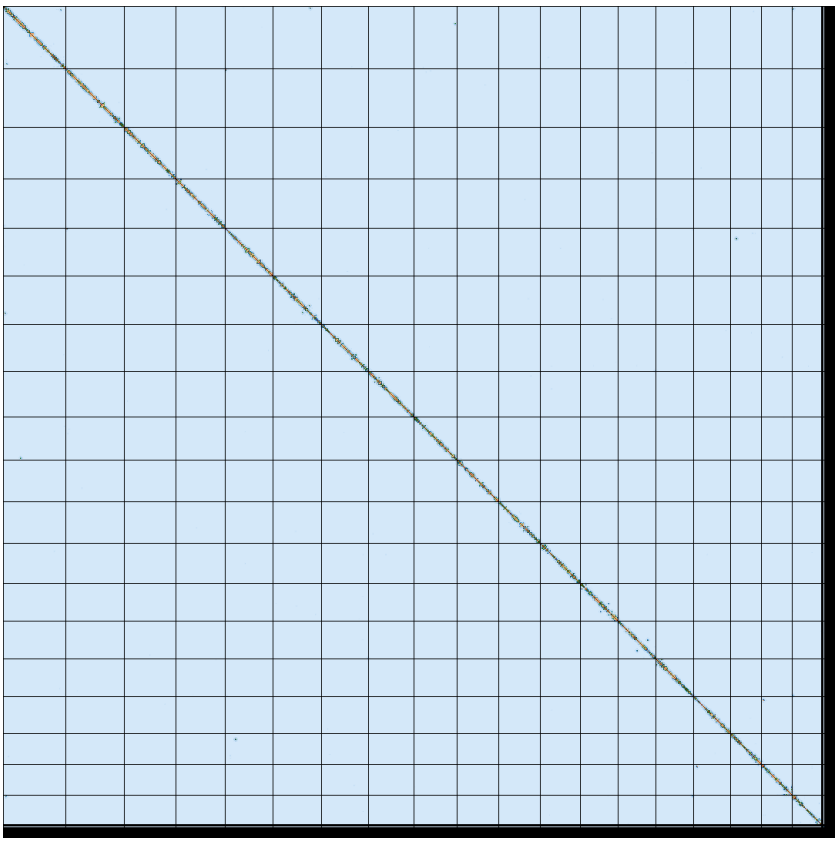
. Interventions/Gb: 41
. Contamination notes: ""
. Other observations: "The assembly of Laevicardium crassum (xbLaeCras1.1) is based on 59X PacBio data and Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio 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, 5 contigs were identified as contaminants (bacterial), totaling 489 Kb (with the largest being 287 Kb). Additionally, 563 regions totaling 125 Mb (with the largest being 4.1 Mb) 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, 66 haplotypic regions were removed, totaling 32.31 Mb (with the largest being 4.39 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	1,539,755,378	1,507,413,752
GC %	37.6	37.55
Gaps/Gbp	231.21	222.9
Total gap bp	35,600	34,400
Scaffolds	234	183
Scaffold N50	78,675,779	78,183,758
Scaffold L50	9	9
Scaffold L90	17	17
Contigs	590	519
Contig N50	6,777,341	7,549,063
Contig L50	62	55
Contig L90	228	206
QV	47.721	61.0195
Kmer compl.	68.9753	68.7821
BUSCO sing.	78.4%	78.6%
BUSCO dupl.	1.1%	0.8%
BUSCO frag.	4.0%	4.0%
BUSCO miss.	16.5%	16.6%

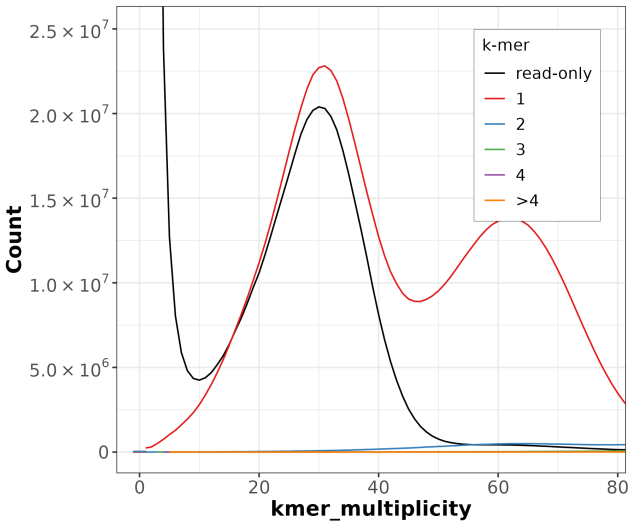
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: mollusca_odb10 (genomes:7, BUSCOs:5295)

HiC contact map of curated assembly

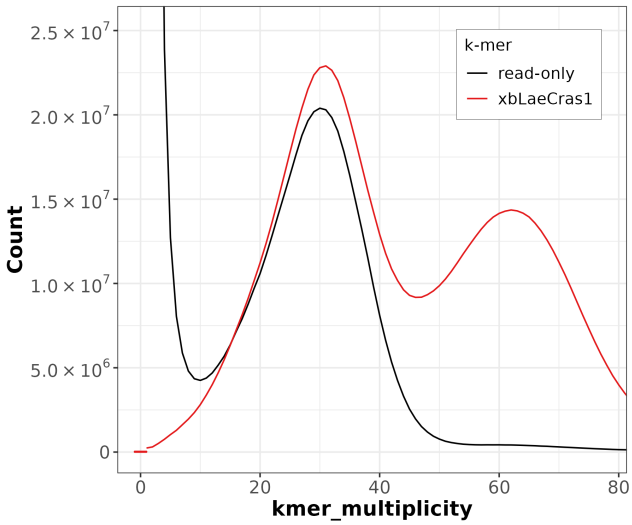


collapsed [LINK]

K-mer spectra of curated assembly

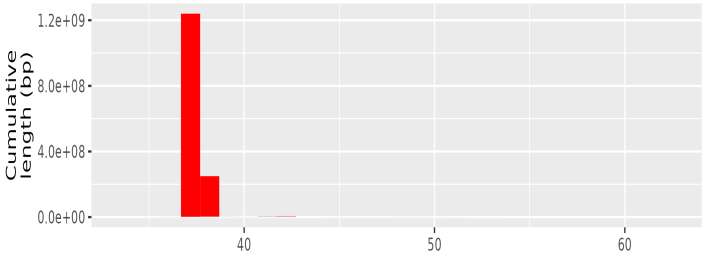


Distribution of k-mer counts per copy numbers found in asm

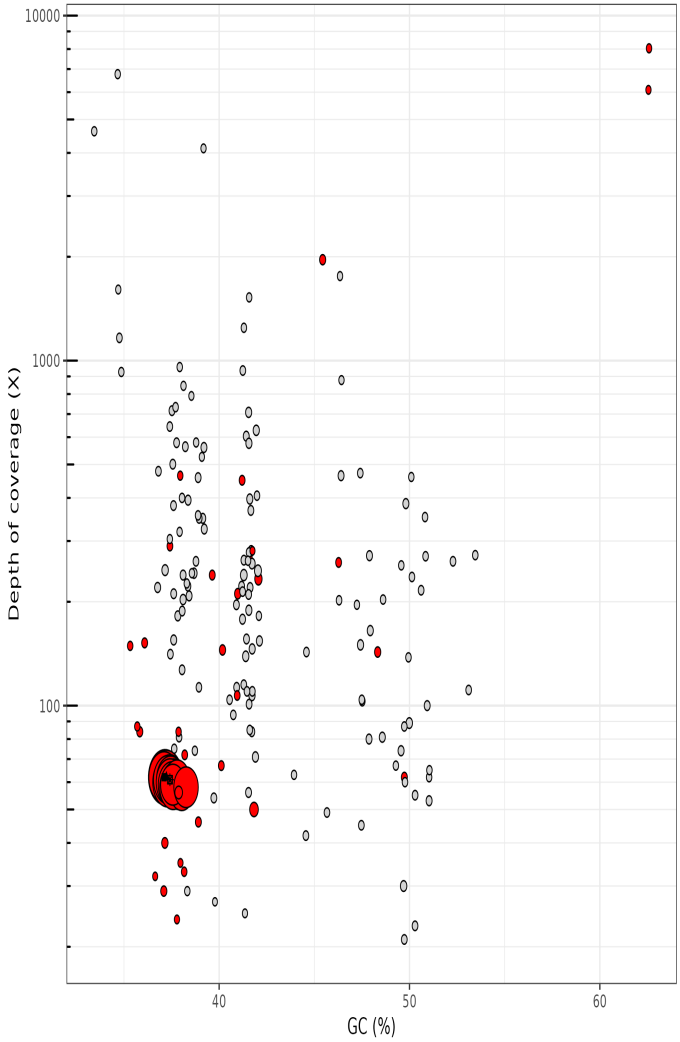


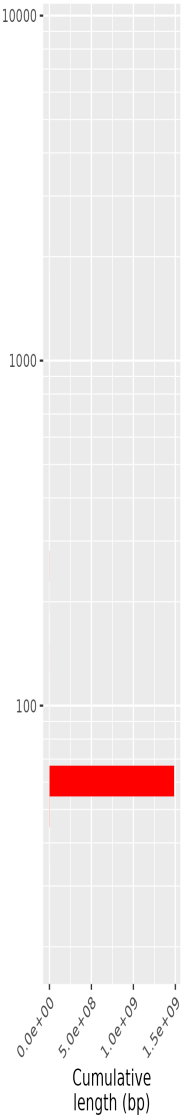
Distribution of k-mer counts coloured by their presence in reads/assemblies

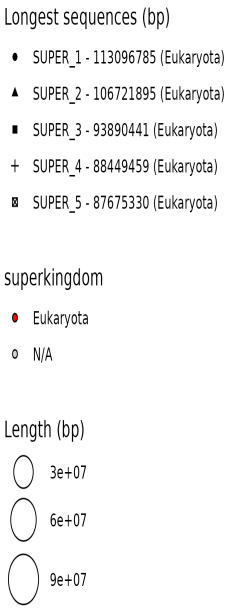
Post-curation contamination screening



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	59	115

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: S.Duprat
Affiliation: Genoscope

Date and time: 2024-12-20 12:12:16 CET