ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	120460
ToLID	**xbStrLact1**
Species	Striarca lactea
Class	Bivalvia
Order	Arcoida

Genome Traits	Expected	Observed
Haploid size (bp)	1,634,569,152	1,765,464,947
Haploid Number	14 (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.Q49

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
. Assembly length loss > 3% for collapsed

Curator notes

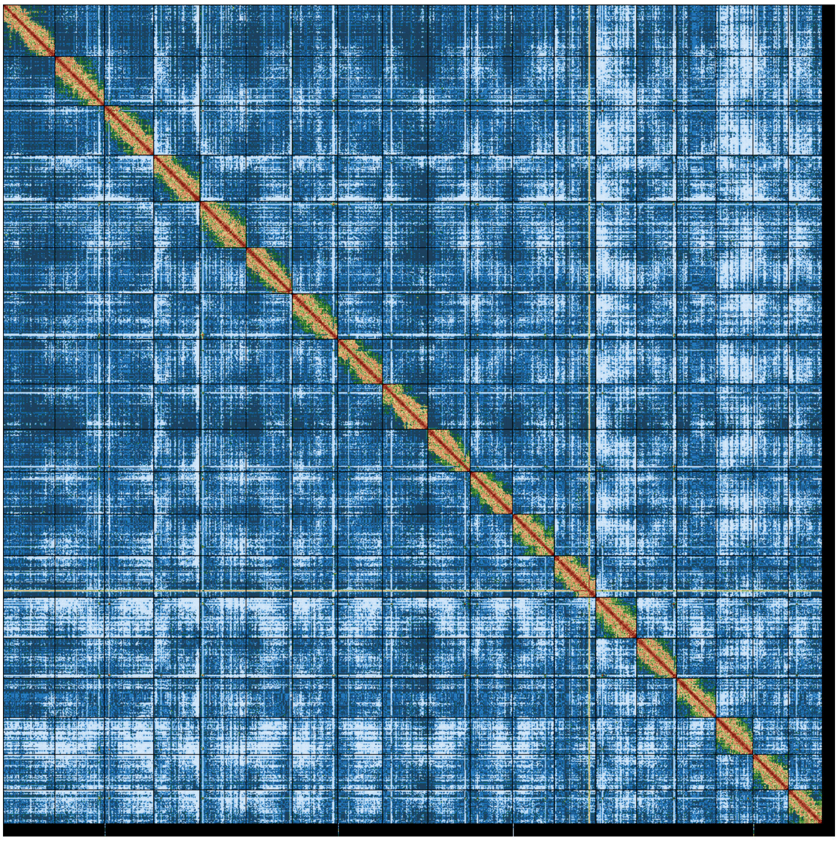
. Interventions/Gb: 174
. Contamination notes: ""
. Other observations: "The assembly of Striarca lactea (xbStrLact1) is based on 36X PacBio data and 113X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio phased assembly generation with Hifiasm using Hi-C data, removal of contaminant sequences using Context, and Hi-C-based scaffolding with YaHS. In total, 23 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 1.003 Mb (with the largest being 0.454 Mb). Additionally, 1034 regions totaling 517.205 Mb (with the largest being 8.567 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using OATK. A first round of curation was made on the combined map using Pretext followed by a curation made on each 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	2,607,375,634	1,765,464,947
GC %	35.23	35.26
Gaps/Gbp	124.26	261.12
Total gap bp	32,400	60,000
Scaffolds	117	635
Scaffold N50	117,109,666	95,426,836
Scaffold L50	10	9
Scaffold L90	23	17
Contigs	441	1,096
Contig N50	10,558,473	6,465,682
Contig L50	78	83
Contig L90	246	282
QV	49.6858	49.1421
Kmer compl.	81.5685	58.5068
BUSCO sing.	51.5%	90.0%
BUSCO dupl.	40.3%	0.7%
BUSCO frag.	6.4%	7.2%
BUSCO miss.	1.7%	2.1%

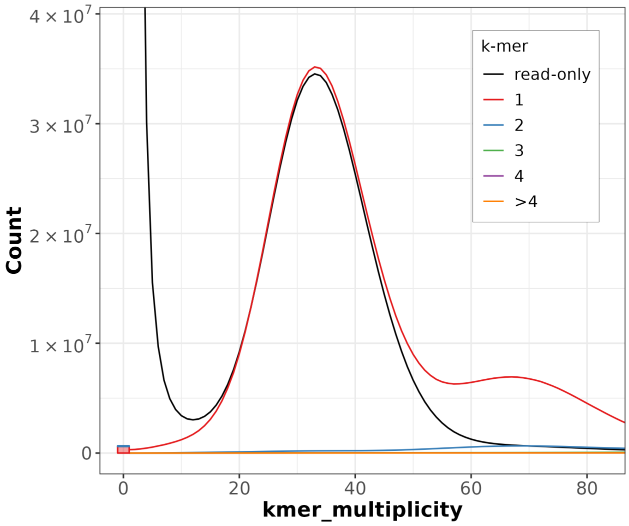
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: mollusca_odb12 (genomes:36, BUSCOs:4421)

HiC contact map of curated assembly

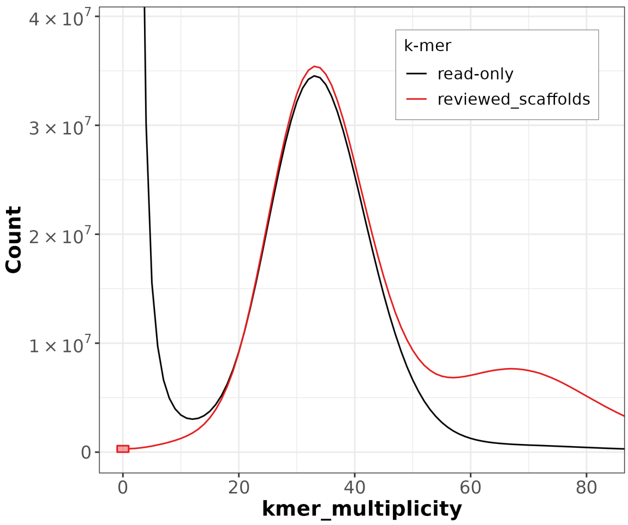


collapsed [LINK]

K-mer spectra of curated assembly

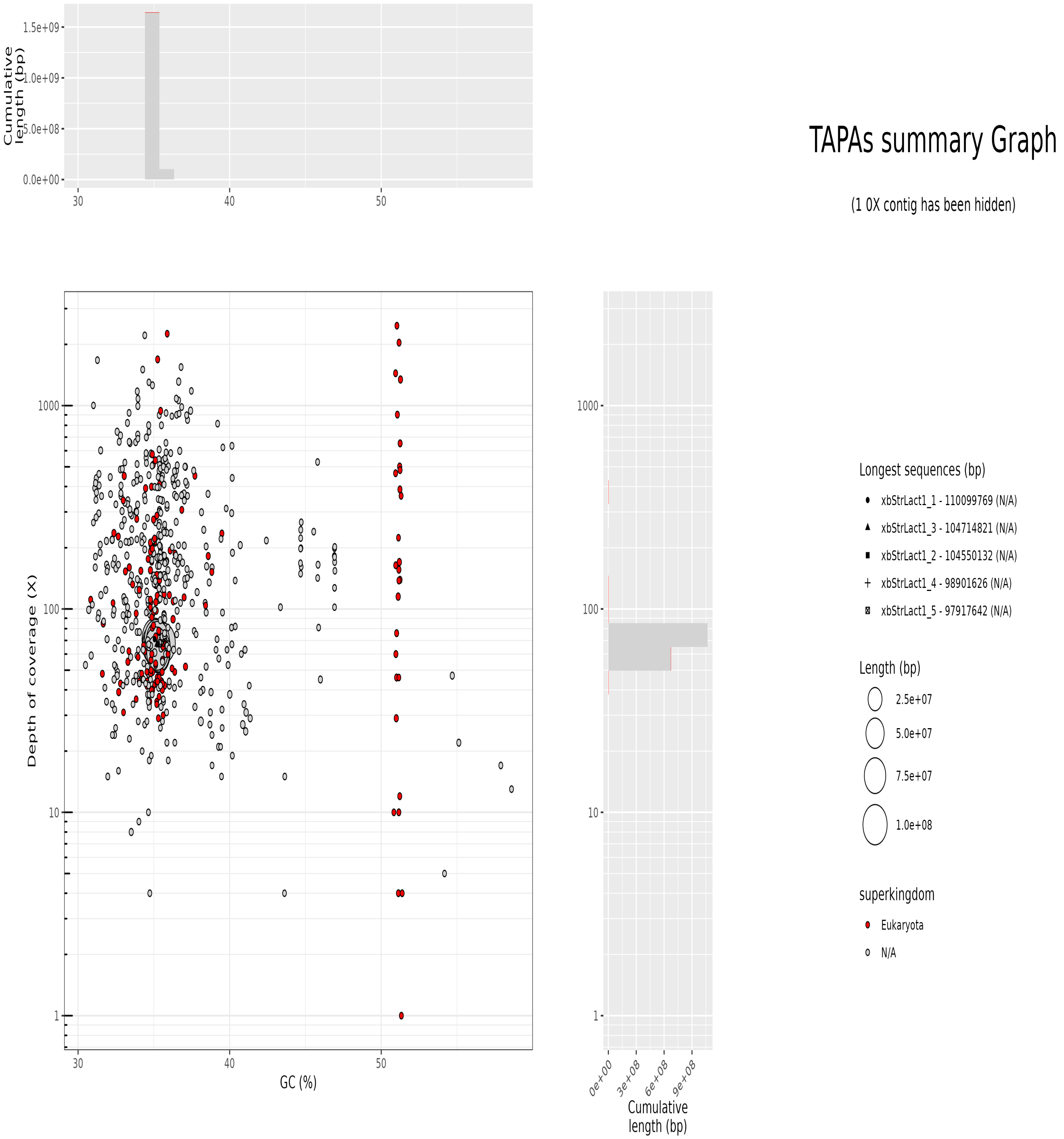


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	76	241

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: Emilie Teodori
Affiliation: Genoscope

Date and time: 2025-09-28 20:14:10 CEST