ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	483686
ToLID	**xgNerPlic1**
Species	Nerita plicata
Class	Gastropoda
Order	Cycloneritida

Genome Traits	Expected	Observed
Haploid size (bp)	982,018,156	1,179,884,630
Haploid Number	9 (source: ancestor)	12
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q40

The following metrics were automatically flagged as below EBP recommended standards or different from expected:

. Observed Haploid size (bp) has >20% difference with 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

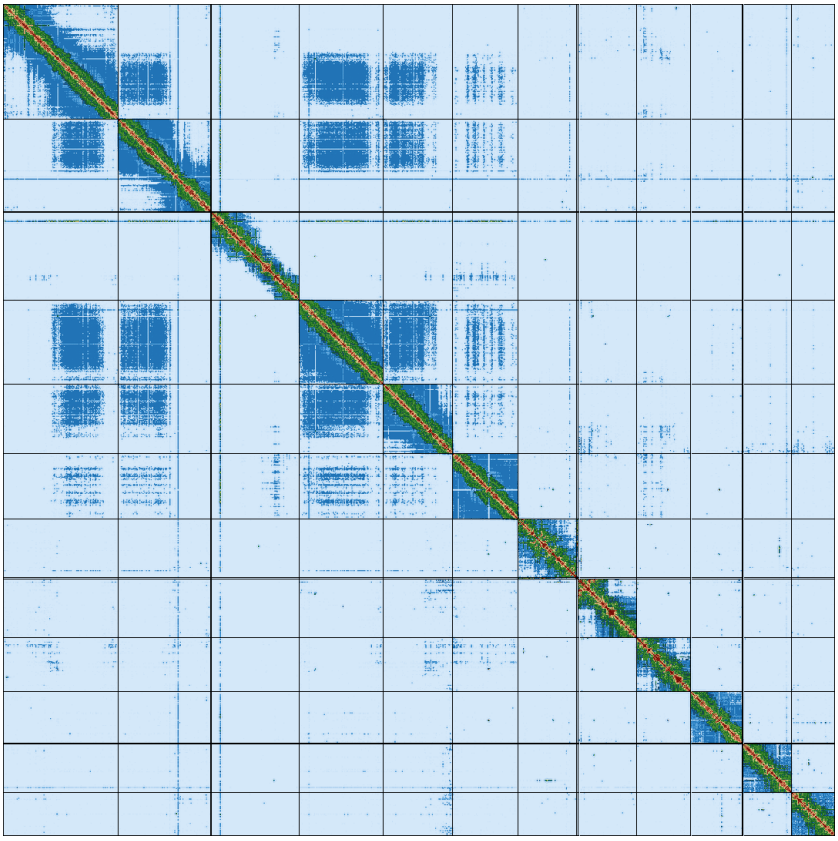
. Interventions/Gb: 299
. Contamination notes: ""
. Other observations: "The assembly of Nerita plicata (xgNerPlic1) is based on 41X ONT data and 202X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr).The assembly process included the following steps: initial ONT assembly generation with Hifiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. 15 contigs totaling 1.5 Mb (with the largest being 214 Kb) were identified as contaminants (bacterial, archaeal, or viral). Additionally, 624 regions totaling 129 Mb (with the largest being 1.3 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using ptGAUL. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 122 supplementary haplotypic regions totaling 164 Mb (with the largest being 6.8 Mb) were removed. 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	1,334,431,788	1,179,884,630
GC %	47.63	47.53
Gaps/Gbp	457.87	507.68
Total gap bp	61,100	74,800
Scaffolds	122	37
Scaffold N50	93,309,135	98,504,541
Scaffold L50	6	5
Scaffold L90	12	11
Contigs	733	636
Contig N50	3,282,898	3,648,574
Contig L50	95	76
Contig L90	407	325
QV	40.233	40.3683
Kmer compl.	71.39	65.4175
BUSCO sing.	84.6%	94.2%
BUSCO dupl.	11.5%	1.7%
BUSCO frag.	1.0%	1.1%
BUSCO miss.	2.8%	3.1%

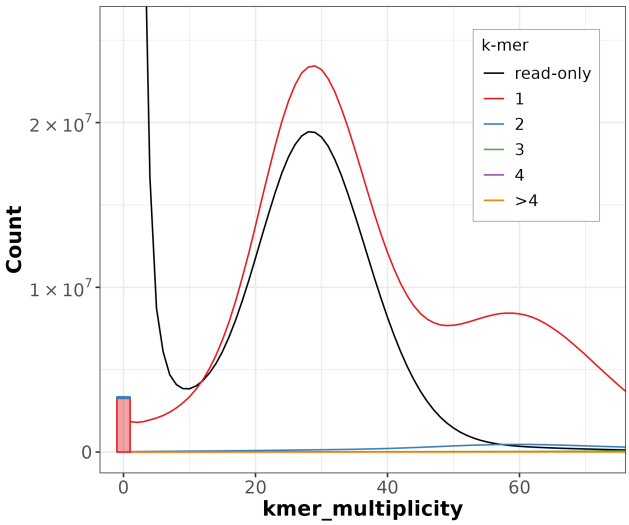
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: mollusca_odb12 (genomes:36, BUSCOs:4421)

HiC contact map of curated assembly

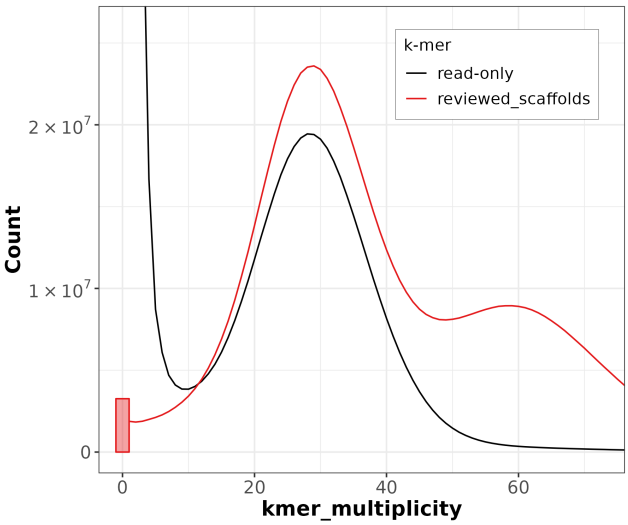


collapsed [LINK]

K-mer spectra of curated assembly

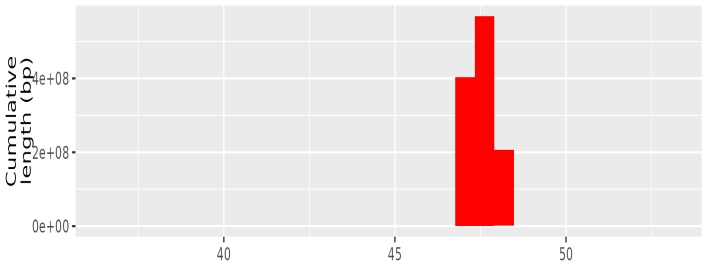


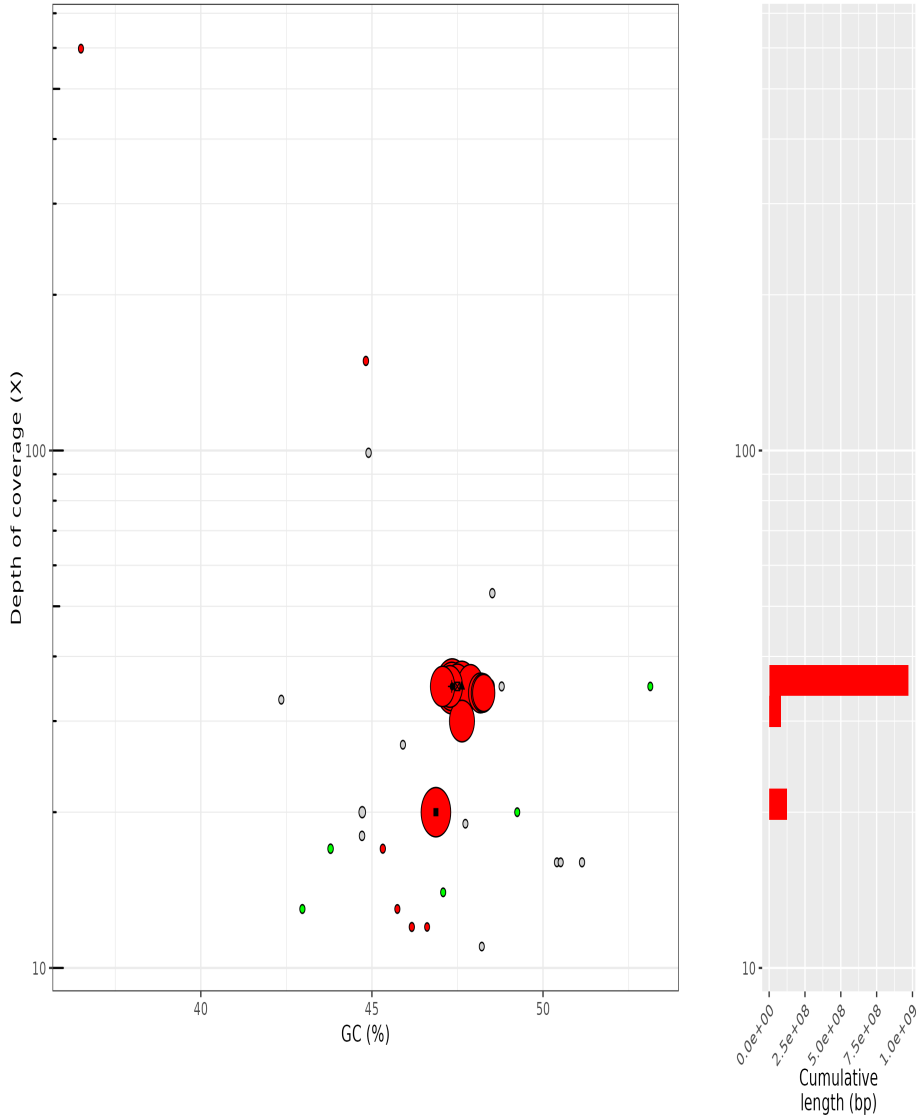
Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening





TAPAs summary Graph

(1 0X contig has been hidden)

superkingdom

- Bacteria
- Eukaryota
- N/A

Longest sequences (bp)

- xgNerPlic1_1 - 163226770 (Eukaryota)
- xgNerPlic1_2 - 131715174 (Eukaryota)
- xgNerPlic1_3 - 124913463 (Eukaryota)
- xgNerPlic1_4 - 119289339 (Eukaryota)
- xgNerPlic1_5 - 98504541 (Eukaryota)

Length (bp)

- 4.0e+07
- 8.0e+07
- 1.2e+08
- 1.6e+08

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	41	202

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: Caroline Menguy
Affiliation: Genoscope

Date and time: 2025-11-30 18:41:45 CET