ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3163778
ToLID	**xbAbrSegm1**
Species	Abra segmentum
Class	Bivalvia
Order	Cardiida

Genome Traits	Expected	Observed
Haploid size (bp)	1,624,760,083	1,704,618,433
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: 7.7.Q58

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

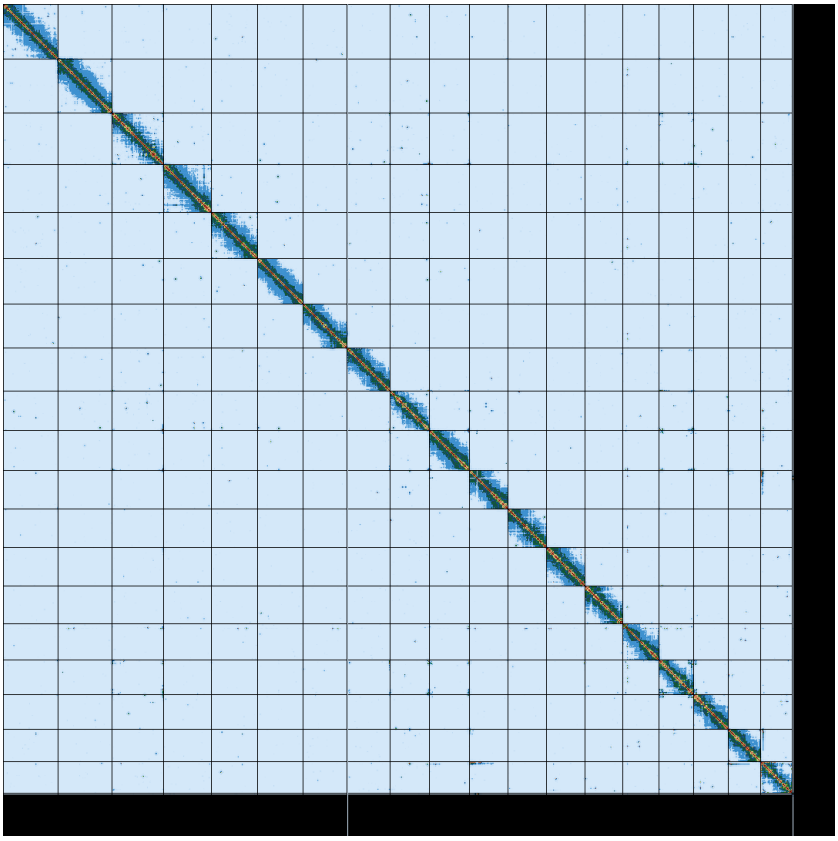
. Interventions/Gb: 54
. Contamination notes: ""
. Other observations: "The assembly of Abra segmentum (xbAbrSegm1) is based on 46X PacBio data and 136X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio 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, 209 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 8.343 Mb (with the largest being 0.379 Mb). Additionally, 585 regions totaling 105.694 Mb (with the largest being 7.245 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, 28 haplotypic regions, totaling 14.9Mb, (with the largest being 1.4Mb). 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	1,719,679,026	1,704,618,433
GC %	36.89	36.89
Gaps/Gbp	134.91	136.69
Total gap bp	23,200	26,500
Scaffolds	1,077	1,047
Scaffold N50	83,860,726	81,859,203
Scaffold L50	9	9
Scaffold L90	18	18
Contigs	1,309	1,280
Contig N50	9,710,108	10,647,745
Contig L50	52	48
Contig L90	181	170
QV	47.4785	58.8864
Kmer compl.	78.5468	78.6437
BUSCO sing.	78.6%	78.8%
BUSCO dupl.	1.4%	1.2%
BUSCO frag.	4.6%	4.6%
BUSCO miss.	15.4%	15.4%

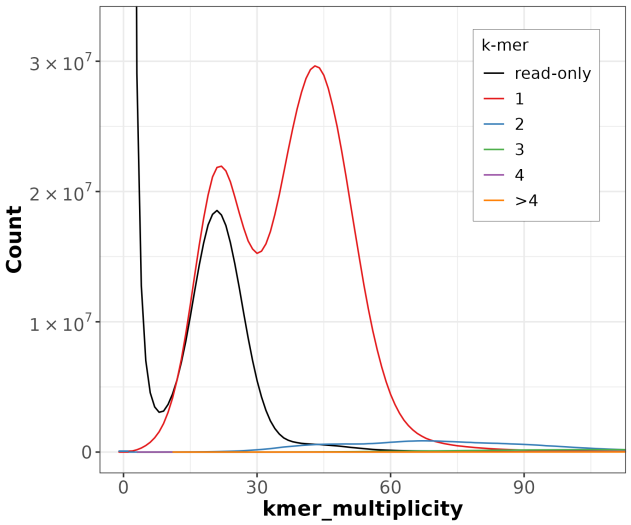
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: mollusca_odb10 (genomes:7, BUSCOs:5295)

HiC contact map of curated assembly

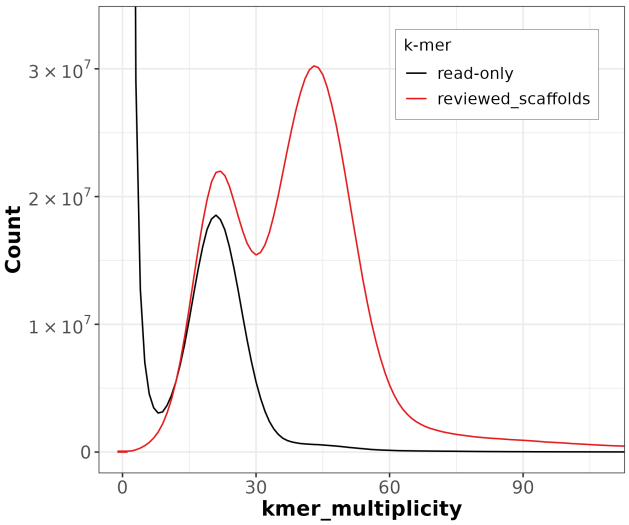


collapsed [LINK]

K-mer spectra of curated assembly

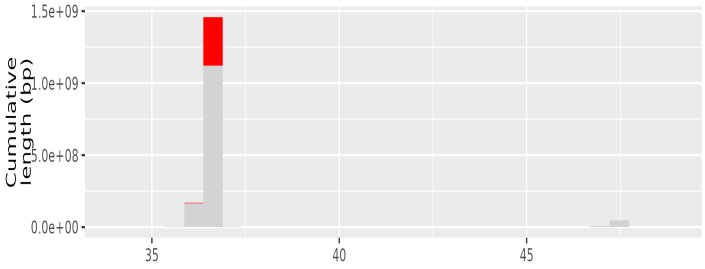


Distribution of k-mer counts per copy numbers found in asm



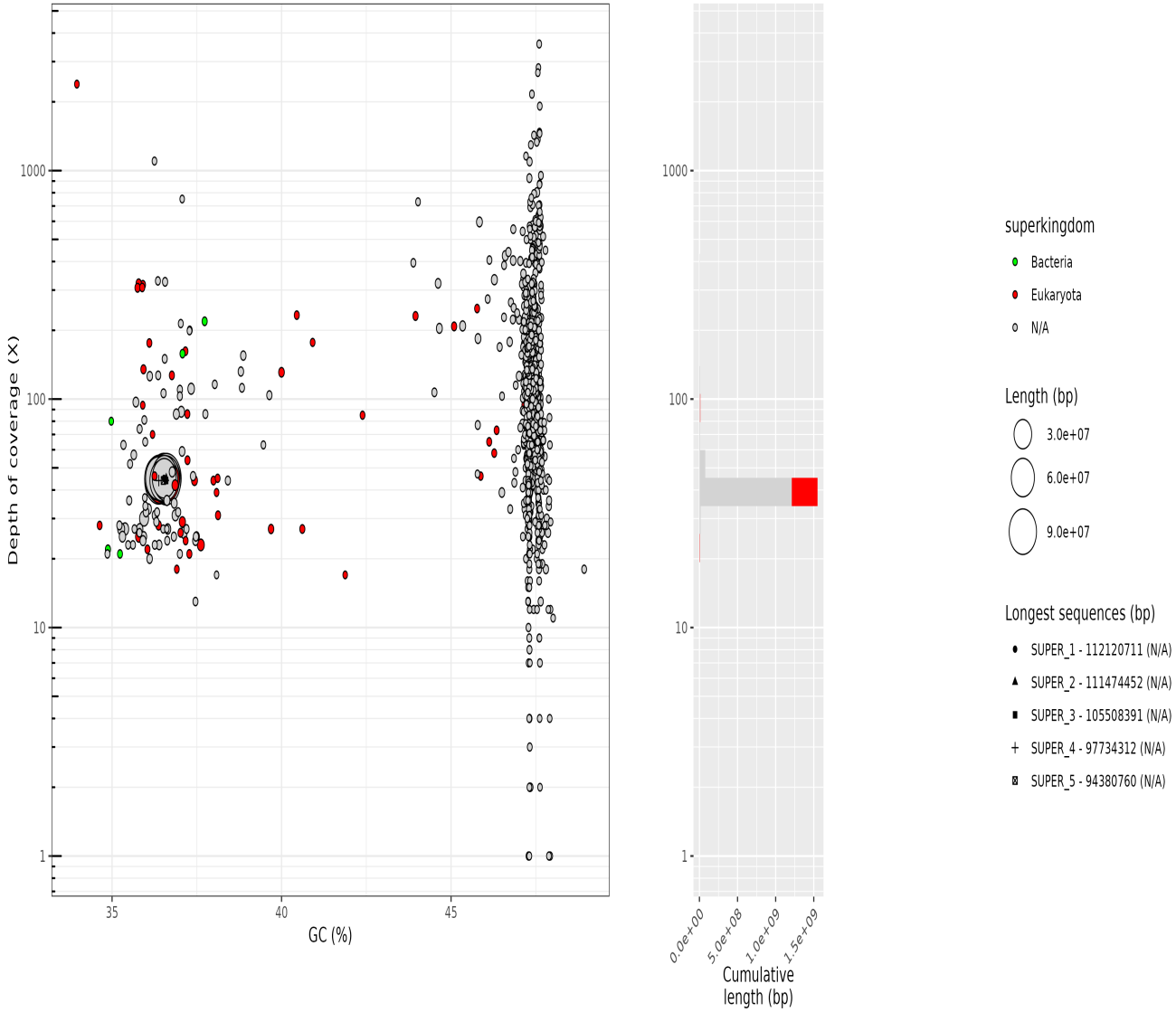
Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph

(14 0X contigs have been hidden)



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	46	150

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: 2025-04-05 17:03:46 CEST