ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3091327
ToLID	**xgThyAren1**
Species	Thylacodes arenarius
Class	Gastropoda
Order	Littorinimorpha

Genome Traits	Expected	Observed
Haploid size (bp)	2,917,874,933	2,944,024,542
Haploid Number	14 (source: direct)	17
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.8.Q40

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
. More than 1000 gaps/Gbp for collapsed

Curator notes

. Interventions/Gb: 103
. Contamination notes: ""
. Other observations: "The assembly of Thylacodes arenarius (xgThyAren1) is based on 42X PacBio data and 159X 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, 800 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 119.814 Mb (with the largest being 2.447 Mb). Additionally, 1996 regions totaling 180.829 Mb (with the largest being 0.648 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, 10 haplotypic regions and 3 contaminant sequences were removed, totaling 5.3 Mb and 0.133 Mb, respectively (with the largest being 0.91 Mb and 0.06 Mb). During manual curation, 11 haplotypic regions were removed, totaling 6.3Mb, (with the largest being 0.98Mb). Chromosome-scale

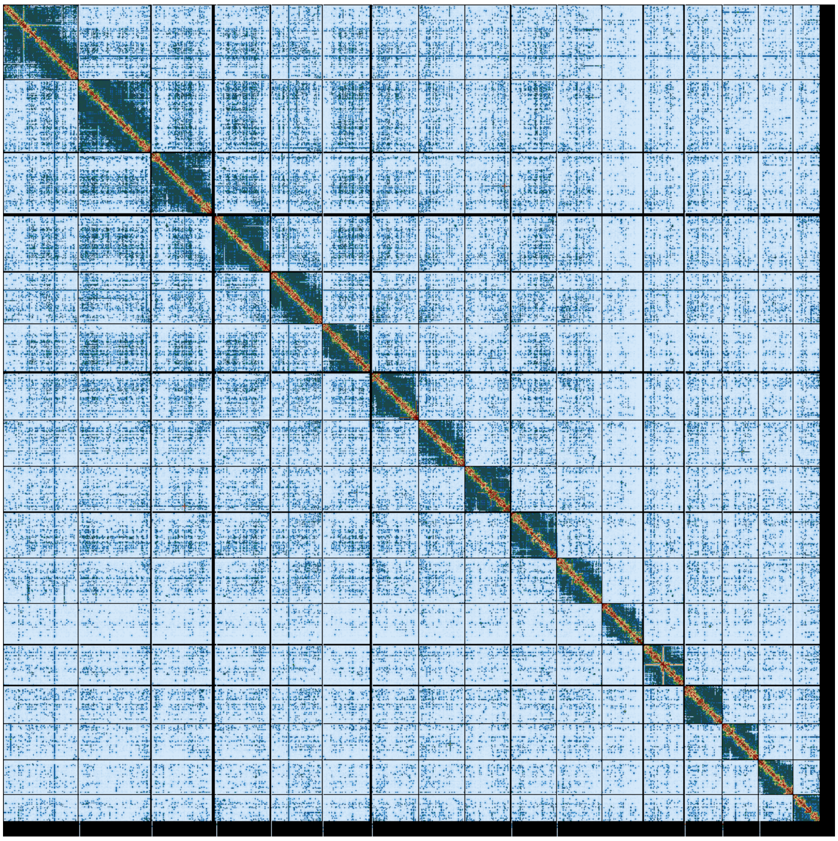
scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	2,951,239,468	2,944,024,542
GC %	43.05	43.05
Gaps/Gbp	1,267.26	1,299.58
Total gap bp	374,000	402,000
Scaffolds	969	527
Scaffold N50	161,936,748	163,651,496
Scaffold L50	7	8
Scaffold L90	15	16
Contigs	4,709	4,353
Contig N50	1,288,835	1,284,536
Contig L50	669	669
Contig L90	2,374	2,366
QV	40.2324	40.2933
Kmer compl.	88.3776	88.291
BUSCO sing.	84.7%	84.8%
BUSCO dupl.	1.7%	1.5%
BUSCO frag.	9.1%	9.1%
BUSCO miss.	4.5%	4.5%

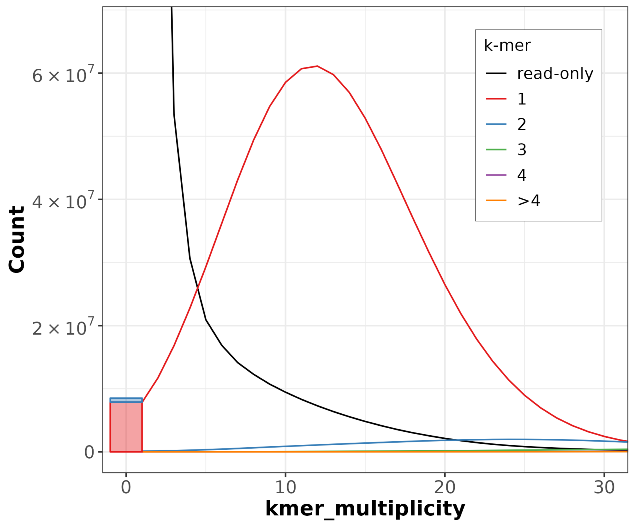
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: mollusca_odb12 (genomes:36, BUSCOs:4421)

HiC contact map of curated assembly

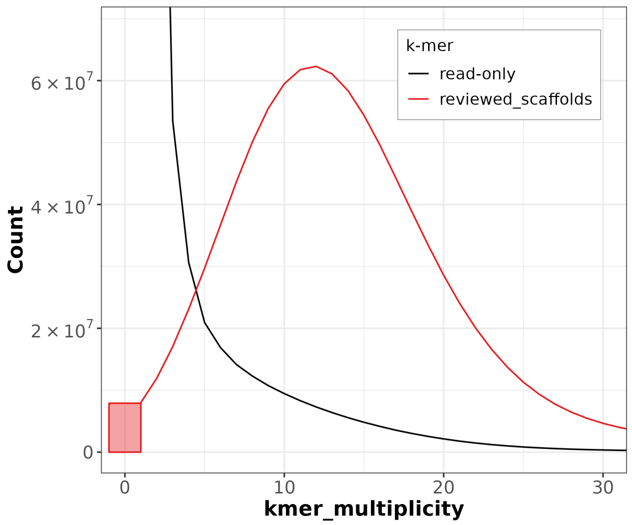


collapsed [LINK]

K-mer spectra of curated assembly

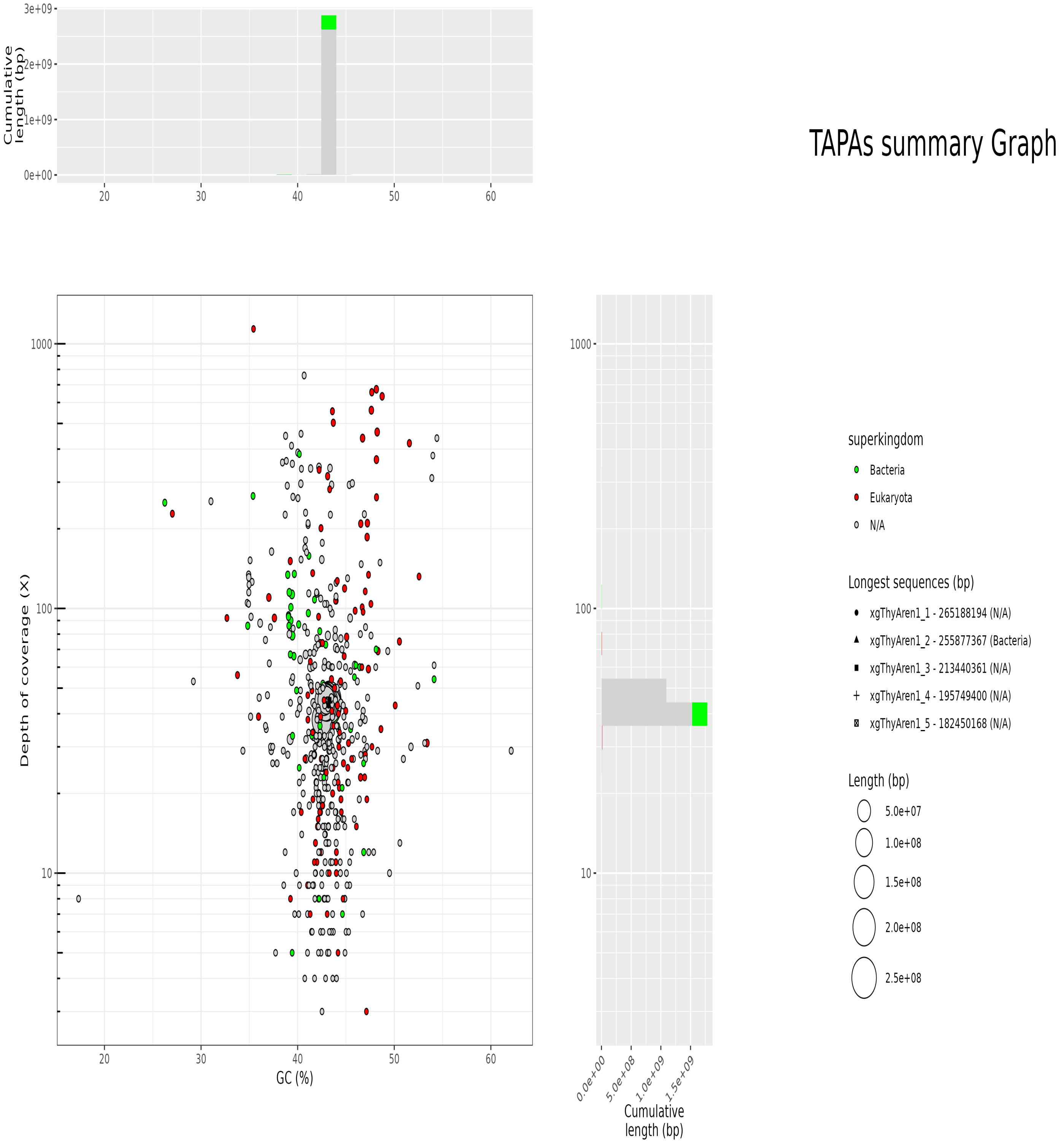


Distribution of k-mer counts per copy numbers found in asm



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	PACBIO Hifi	Arima
Coverage	41	158

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: 2025-08-10 09:54:12 CEST