ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	154630
ToLID	**xgThuHope1**
Species	Thuridilla hopei
Class	Gastropoda
Order	NA

Genome Traits	Expected	Observed
Haploid size (bp)	1,589,644,786	1,076,183,911
Haploid Number	16 (source: ancestor)	15
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q56

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
. BUSCO single copy value is less than 90% for collapsed
. More than 1000 gaps/Gbp for collapsed

Curator notes

. Interventions/Gb: 89
. Contamination notes: ""
. Other observations: "The assembly of Thuridilla hopei (xgThuHope1) is based on 46X PacBio data and 115X 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, 241 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 21.701 Mb (with the largest being 5.336 Mb). Additionally, 1984 regions totaling 164.557 Mb (with the largest being 1.215 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, totaling 2.8Mb, (with the largest being 0.65Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. Another decontamination step was performed after

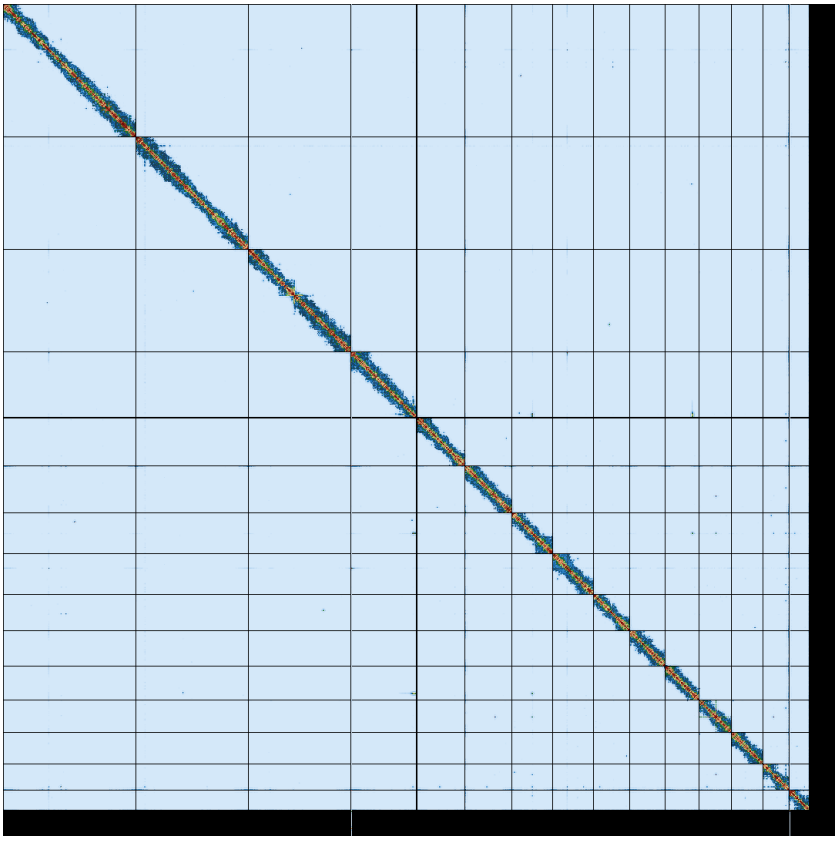
the manual curation, removing 495 scaffolds (8 bacterial scaffolds and 487 of green algae) for a total of 12.2 Mb. The sample was contaminated by a green algae, scaffolds identified as Viridiplantae (with alpha-majority>0.5) and scaffolds with valid pairs/Mb<1e4 were removed. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	1,091,832,850	1,076,183,911
GC %	38.62	38.48
Gaps/Gbp	1,532.29	1,573.15
Total gap bp	167,300	175,000
Scaffolds	1,119	485
Scaffold N50	64,228,433	61,942,397
Scaffold L50	5	5
Scaffold L90	14	13
Contigs	2,792	2,178
Contig N50	1,018,000	1,039,305
Contig L50	302	293
Contig L90	1,106	1,054
QV	24.4778	56.5266
Kmer compl.	63.4928	76.3543
BUSCO sing.	87.3%	87.6%
BUSCO dupl.	1.4%	1.0%
BUSCO frag.	2.3%	2.3%
BUSCO miss.	9.0%	9.1%

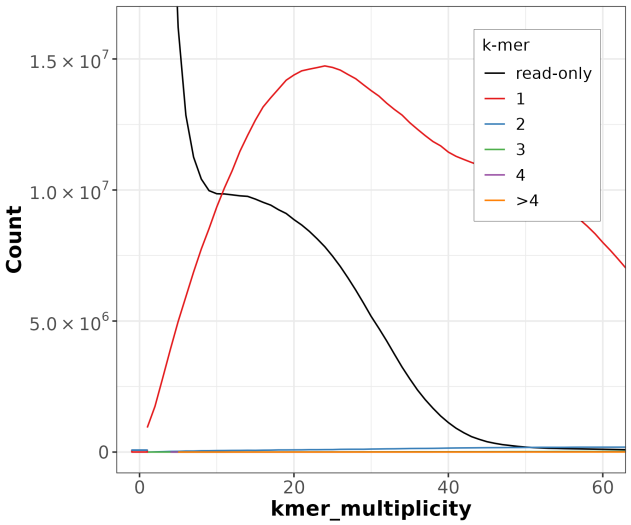
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: mollusca_odb10 (genomes:7, BUSCOs:5295)

HiC contact map of curated assembly

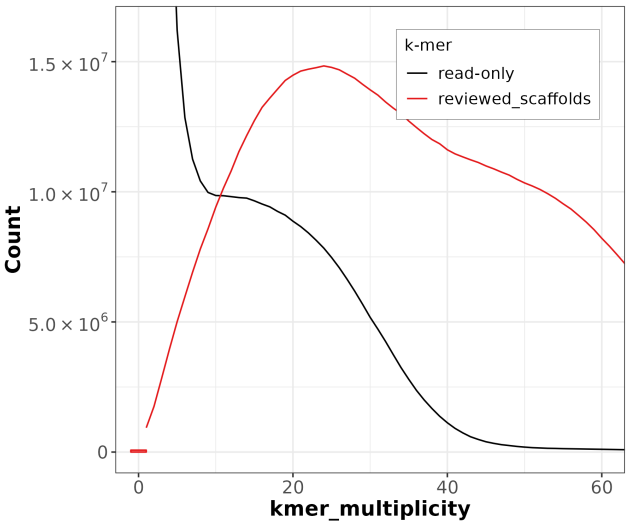


collapsed [LINK]

K-mer spectra of curated assembly

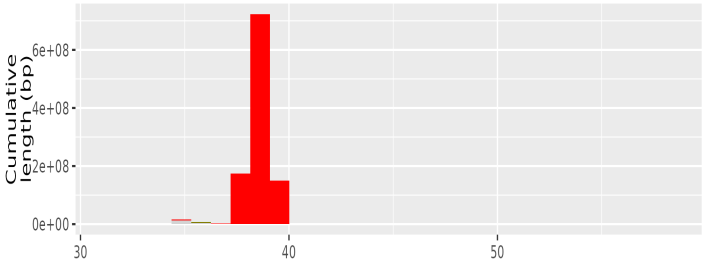


Distribution of k-mer counts per copy numbers found in asm

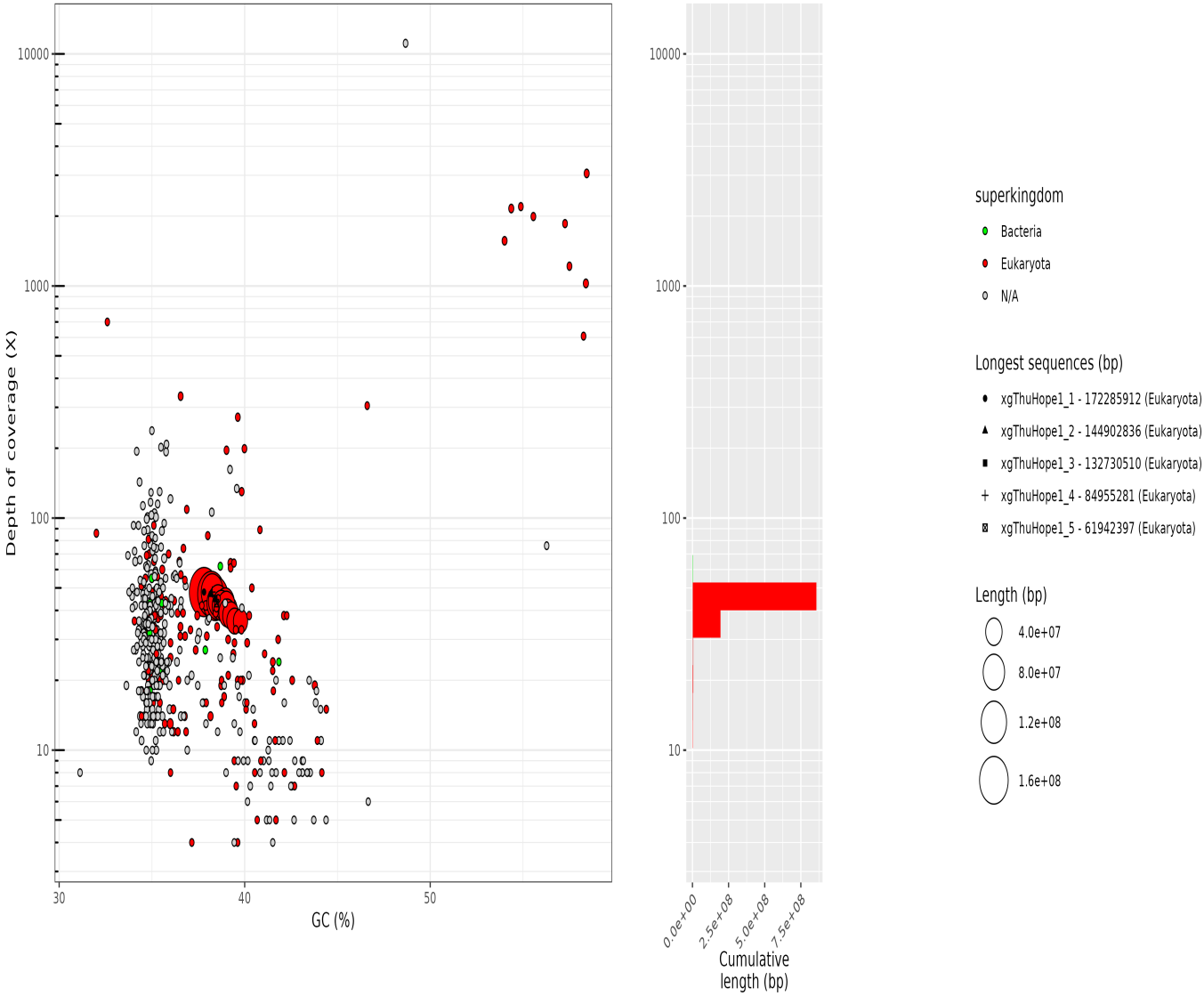


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	46	117

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: 2025-04-13 14:17:34 CEST