ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	179648
ToLID	**xgHexTrun1**
Species	Hexaplex trunculus
Class	Gastropoda
Order	Neogastropoda

Genome Traits	Expected	Observed
Haploid size (bp)	1,925,960,221	2,059,483,820
Haploid Number	12 (source: direct)	35
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q42

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
. BUSCO duplicated value is more than 5% for collapsed

Curator notes

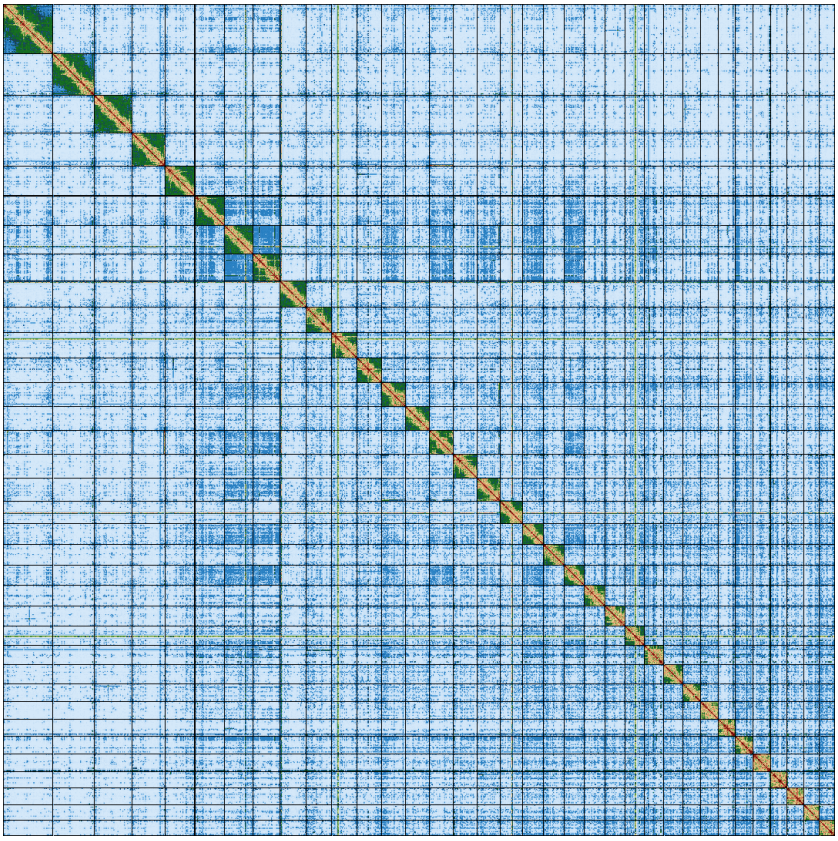
. Interventions/Gb: 63
. Contamination notes: ""
. Other observations: "The assembly of Hexaplex trunculus (xgHexTrun1) is based on 52X ONT data and 136X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial ONT assembly generation with Hifiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 27 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 14.372 Mb (with the largest being 10.011 Mb). Additionally, 251 regions totaling 68.164 Mb (with the largest being 4.783 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using OATK. During manual curation, 11 haplotypic regions were removed, totaling 27.8Mb (with the largest being 9.3Mb). 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	2,087,279,237	2,059,483,820
GC %	41.43	41.43
Gaps/Gbp	52.7	57.78
Total gap bp	11,000	14,500
Scaffolds	76	61
Scaffold N50	59,221,300	58,989,327
Scaffold L50	14	14
Scaffold L90	31	31
Contigs	186	180
Contig N50	24,760,430	24,966,295
Contig L50	28	27
Contig L90	87	84
QV	42.8556	42.8508
Kmer compl.	77.1769	76.8326
BUSCO sing.	78.1%	78.8%
BUSCO dupl.	19.2%	18.5%
BUSCO frag.	0.7%	0.8%
BUSCO miss.	2.0%	2.0%

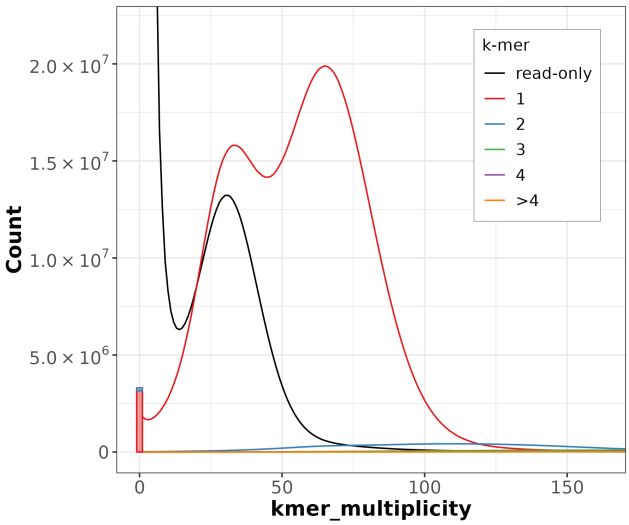
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: mollusca_odb12 (genomes:36, BUSCOs:4421)

HiC contact map of curated assembly

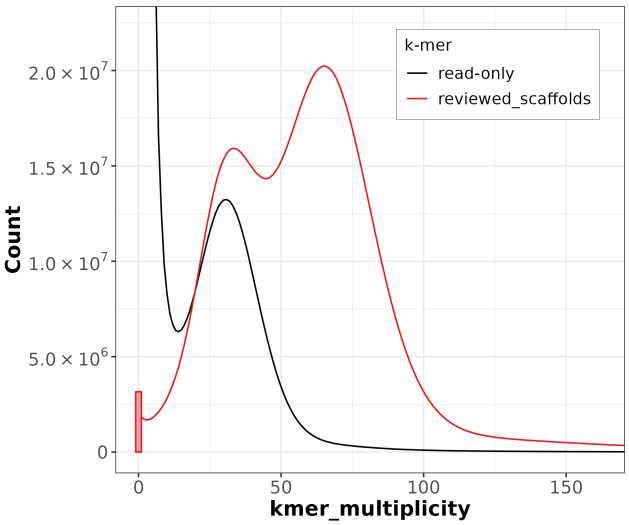


collapsed [LINK]

K-mer spectra of curated assembly

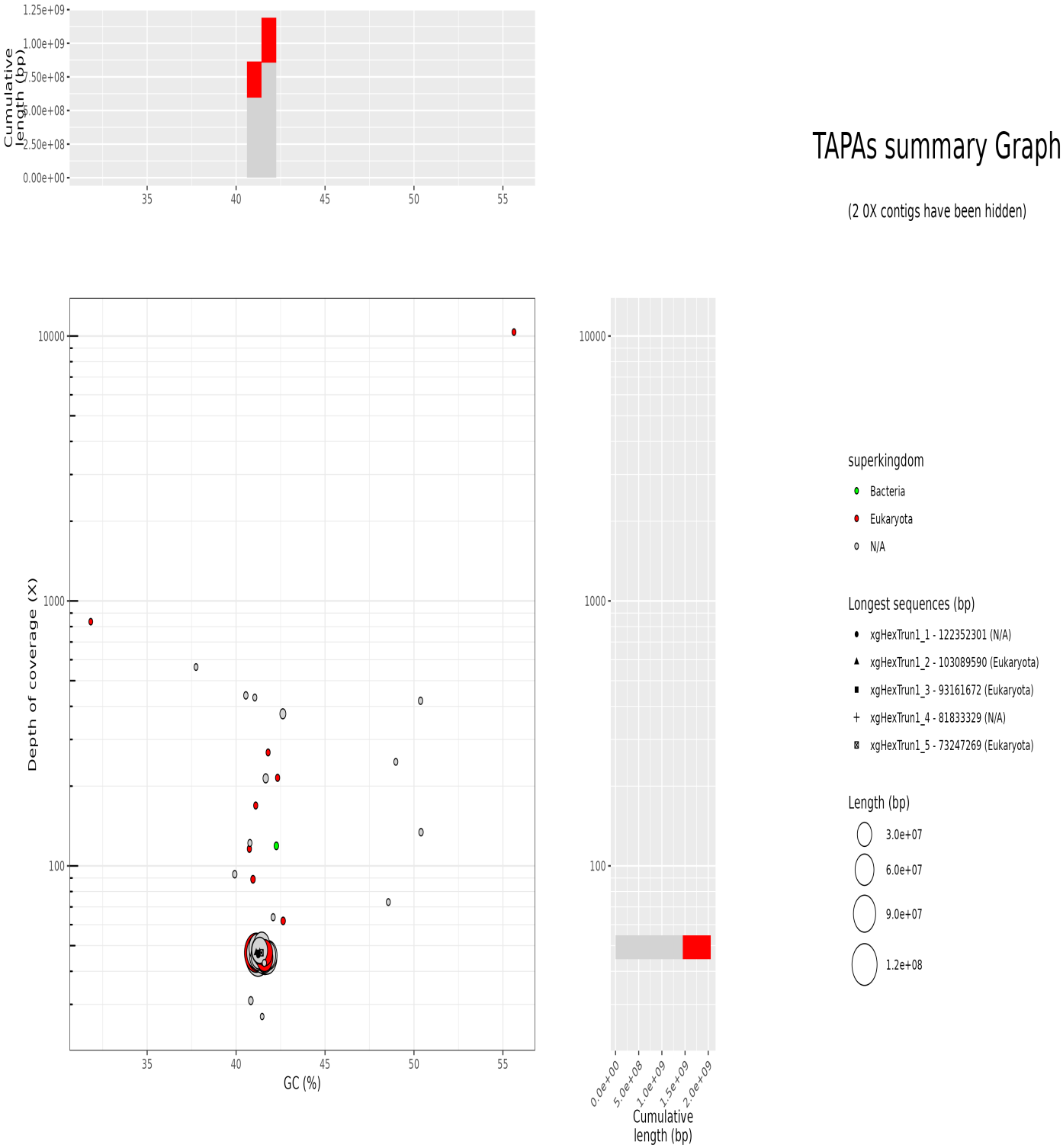


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	51	136

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: Benjamin Istace
Affiliation: Genoscope

Date and time: 2025-11-23 22:30:54 CET