ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1436025
ToLID	**qmDioPugi1**
Species	Diogenes pugilator
Class	Malacostraca
Order	Decapoda

Genome Traits	Expected	Observed
Haploid size (bp)	1,235,455,095	1,397,329,010
Haploid Number	12 (source: ancestor)	105
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.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

Curator notes

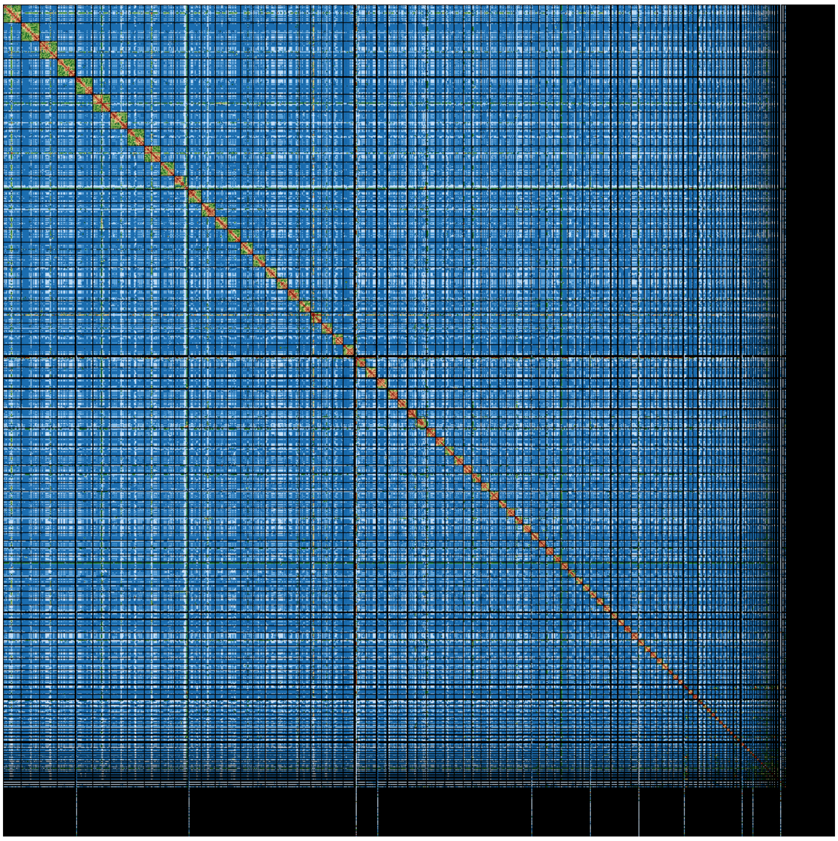
. Interventions/Gb: 227
. Contamination notes: ""
. Other observations: "The assembly of Diogenes pugilator (qmDioPugi1) is based on 34X ONT data and 584X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial ONT 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, 52 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 9.978 Mb (with the largest being 0.383 Mb). Additionally, 2216 regions totaling 340.821 Mb (with the largest being 0.816 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, 9 haplotypic regions, totaling 1.8Mb, (with the largest being 0.3Mb). 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	1,073,970,350	1,397,329,010
GC %	39.49	38.74
Gaps/Gbp	2,433.03	933.92
Total gap bp	261,300	149,700
Scaffolds	617	620
Scaffold N50	12,097,947	15,841,835
Scaffold L50	31	32
Scaffold L90	84	89
Contigs	3,230	1,925
Contig N50	504,336	1,353,000
Contig L50	405	327
Contig L90	1,950	1,037
QV	32.9491	42.4513
Kmer compl.	62.3626	67.6453
BUSCO sing.	91.3%	95.5%
BUSCO dupl.	1.4%	0.5%
BUSCO frag.	2.2%	1.5%
BUSCO miss.	5.1%	2.5%

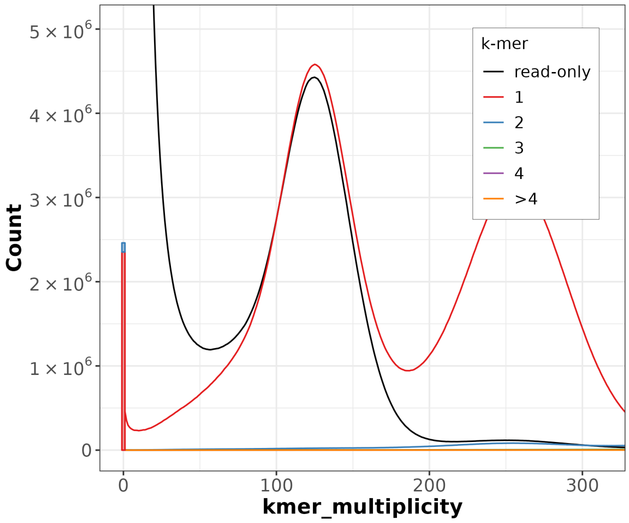
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: arthropoda_odb10 (genomes:90, BUSCOs:1013)

HiC contact map of curated assembly

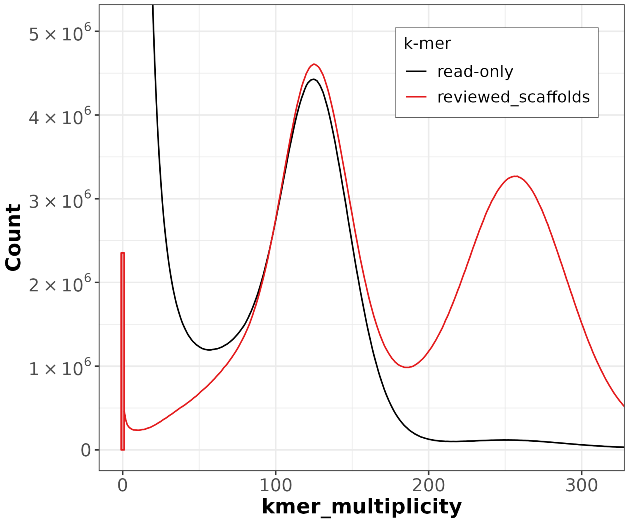


collapsed [LINK]

K-mer spectra of curated assembly

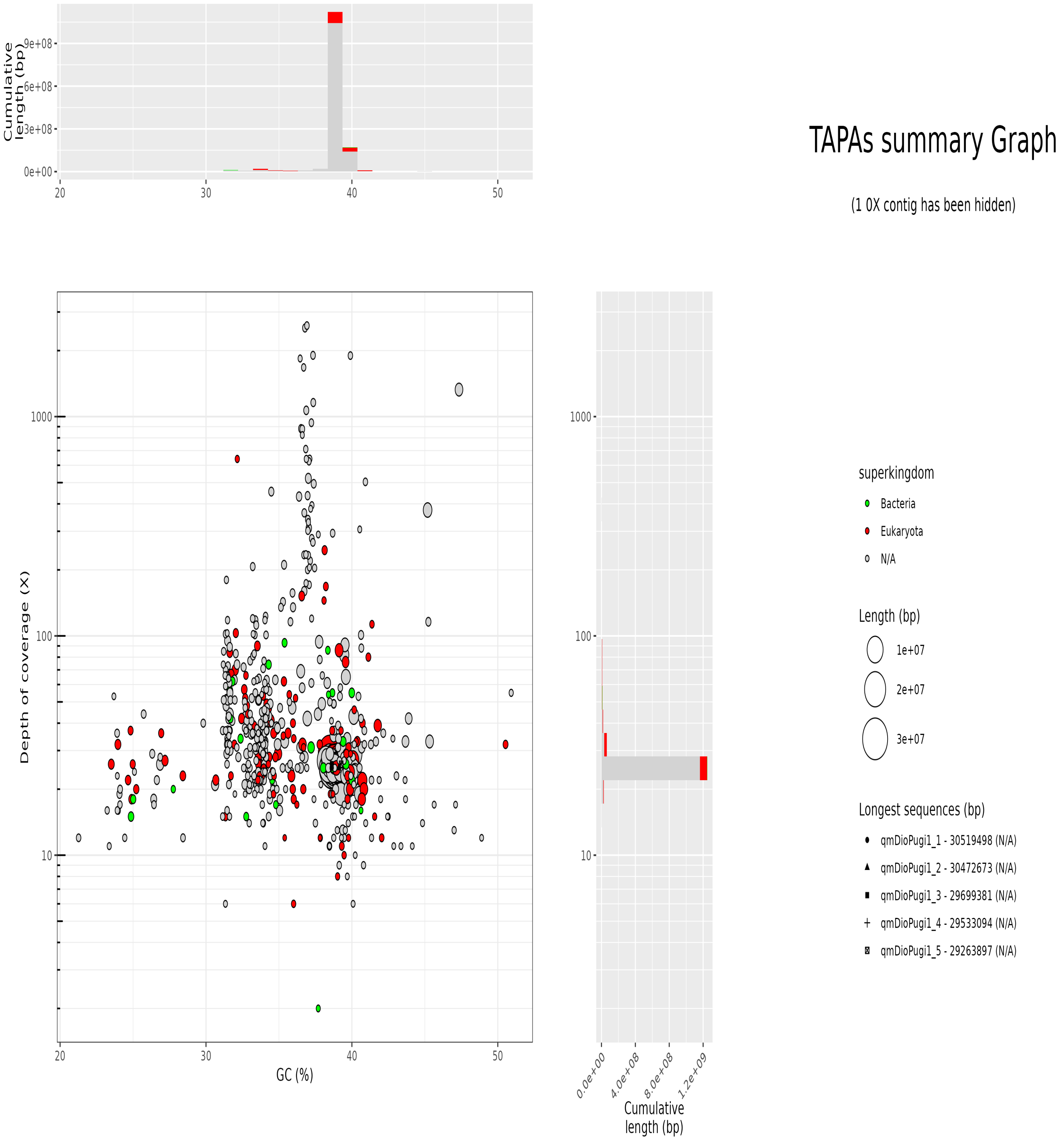


Distribution of k-mer counts per copy numbers found in asm



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	33	658

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: 2025-05-22 06:43:05 CEST