ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	2820131
ToLID	**qmDioCurv1**
Species	Diogenes curvimanus
Class	Malacostraca
Order	Decapoda

Genome Traits	Expected	Observed
Haploid size (bp)	1,530,297,999	2,000,766,069
Haploid Number	12 (source: ancestor)	106
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.7.Q42

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
. Assembly length loss > 3% for collapsed
. More than 1000 gaps/Gbp for collapsed
. Not 90% of assembly in chromosomes for collapsed

Curator notes

. Interventions/Gb: 389
. Contamination notes: ""
. Other observations: "The assembly of Diogenes curvimanus (qmDioCurv1.1) is based on 37X ONT data and 231X 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, 50 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 9.418 Mb (with the largest being 1.769 Mb). Additionally, 3293 regions totaling 174.142 Mb (with the largest being 0.385 Mb) were identified as haplotypic duplications and removed.The mitochondrial genome was assembled using ptGAUL. A linear sequence of 35Kb was obtained. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 300 haplotypic regions and 5 contaminant sequences were removed, totaling 67.859 Mb and 0.111 Mb,

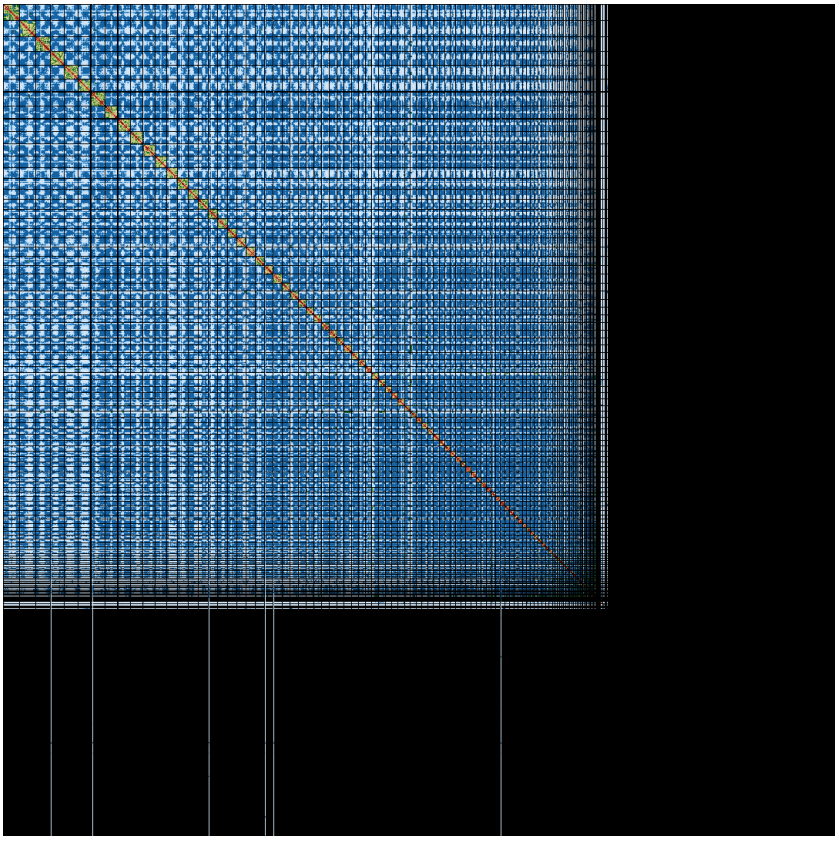
respectively (with the largest being 0.955 Mb and 0.032 Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	2,069,360,393	2,000,766,069
GC %	38.98	38.97
Gaps/Gbp	1,626.59	1,636.37
Total gap bp	336,600	361,600
Scaffolds	3,571	3,170
Scaffold N50	12,339,892	13,908,947
Scaffold L50	50	44
Scaffold L90	927	829
Contigs	6,937	6,444
Contig N50	610,196	619,108
Contig L50	914	873
Contig L90	3,718	3,547
QV	42.4611	42.5009
Kmer compl.	81.0272	79.5478
BUSCO sing.	92.7%	97.5%
BUSCO dupl.	6.9%	2.0%
BUSCO frag.	0.1%	0.1%
BUSCO miss.	0.3%	0.4%

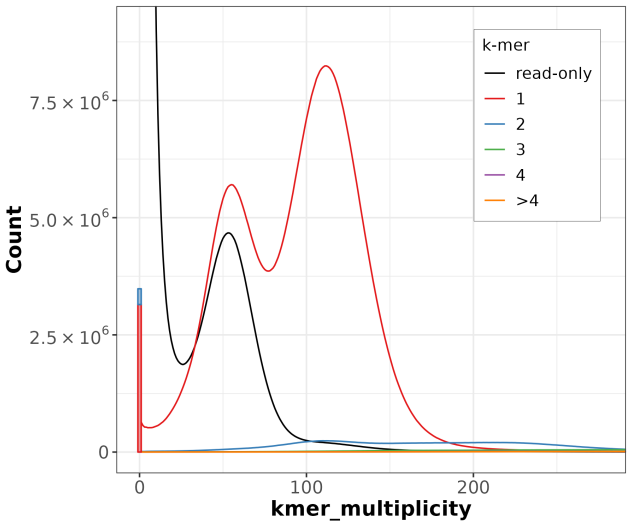
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: crustacea_odb12 (genomes:25, BUSCOs:1536)

HiC contact map of curated assembly

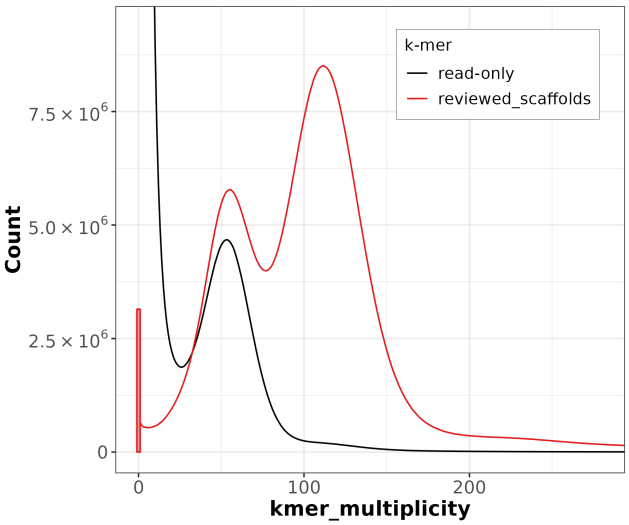


collapsed [LINK]

K-mer spectra of curated assembly

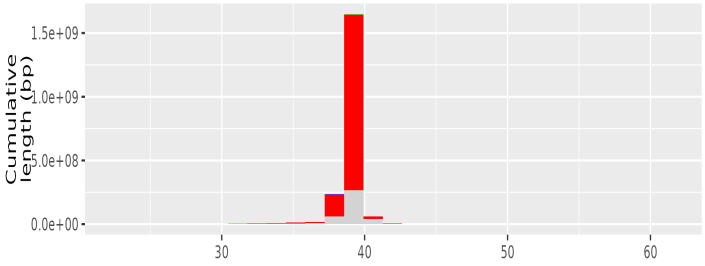


Distribution of k-mer counts per copy numbers found in asm



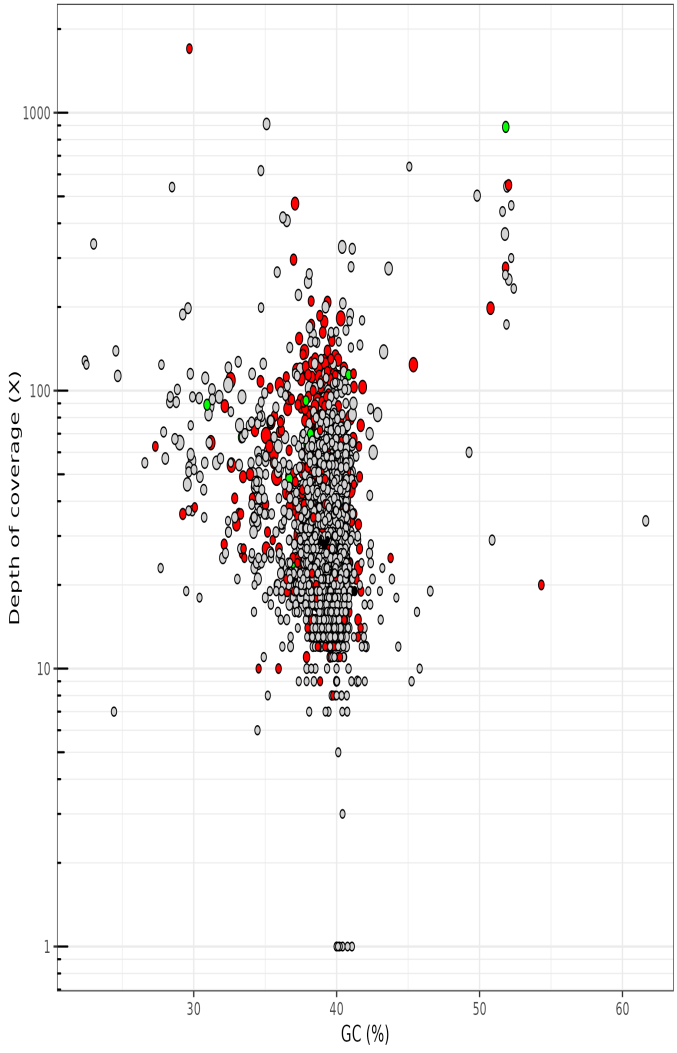
Distribution of k-mer counts coloured by their presence in reads/assemblies

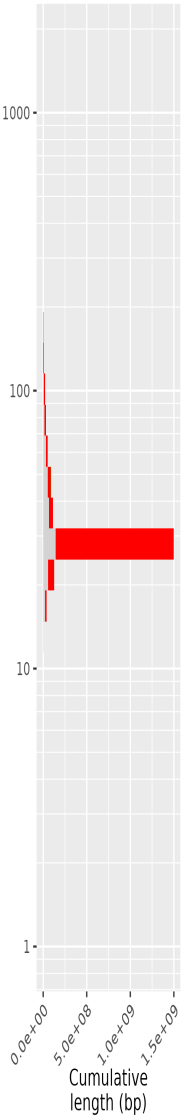
Post-curation contamination screening

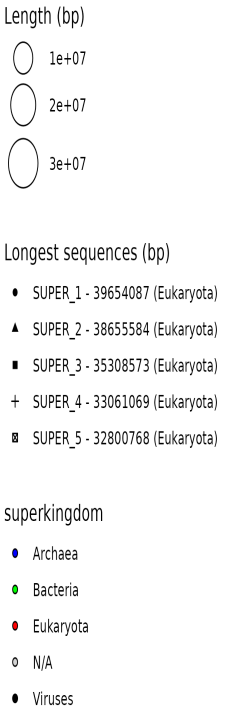


TAPAs summary Graph

(22 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	Long reads	Arima
Coverage	37	231

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: Adama Ndar
Affiliation: Genoscope

Date and time: 2025-12-14 19:17:05 CET