ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	344925
ToLID	**wsDitArie1**
Species	Ditrupa arietina
Class	Polychaeta
Order	Sabellida

Genome Traits	Expected	Observed
Haploid size (bp)	1,224,994,191	1,355,772,114
Haploid Number	14 (source: ancestor)	10
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.8.Q43

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
. Assembly length loss > 3% for collapsed

Curator notes

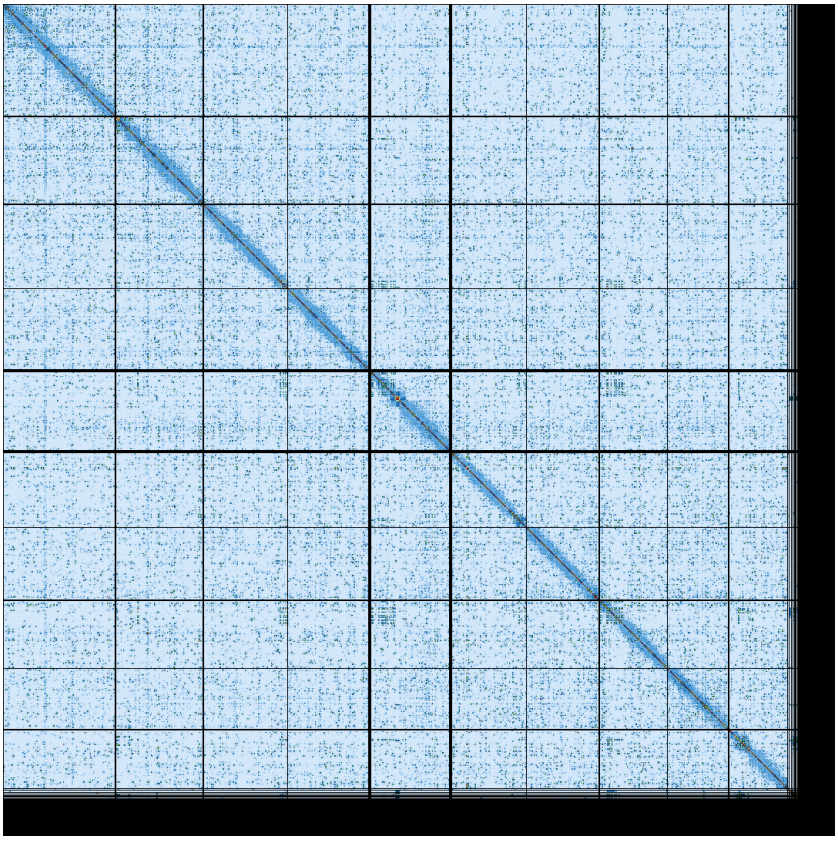
. Interventions/Gb: 30
. Contamination notes: ""
. Other observations: "The assembly of Ditrupa arietina (wsDitArie1) is based on 56.51X PacBio HiFi data from an amplified DNA library and 298.32X Arima and OmniC Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly 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, 1128 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 36.665 Mb (with the largest being 1.656 Mb). Additionally, 2300 regions totaling 149.764 Mb (with the largest being 1.401 Mb) were identified as haplotypic duplications and removed. Mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 12 haplotypic regions were removed, totaling 3.501Mb (with the largest being 0.543Mb). The telomeric pattern AATGACACTAG was identified with TelFinder and used to generate the Pretext Telomeres tracks. 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	1,429,416,692	1,355,772,114
GC %	36.09	36.11
Gaps/Gbp	758.35	784.79
Total gap bp	108,400	120,800
Scaffolds	807	748
Scaffold N50	125,447,008	129,088,172
Scaffold L50	6	5
Scaffold L90	13	10
Contigs	1,891	1,812
Contig N50	2,705,000	2,726,246
Contig L50	157	148
Contig L90	549	505
QV	43.193	43.1184
Kmer compl.	69.6187	67.2221
BUSCO sing.	88.5%	90.7%
BUSCO dupl.	8.8%	2.4%
BUSCO frag.	1.2%	5.2%
BUSCO miss.	1.5%	1.8%

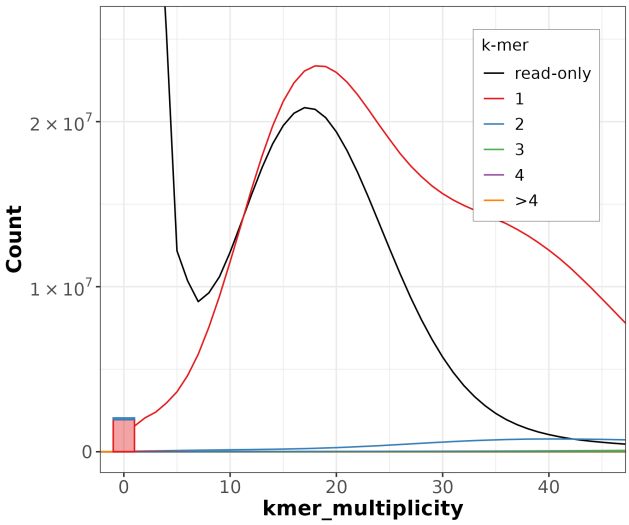
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: lophotrochozoa_odb12 (genomes:75, BUSCOs:1252)
BUSCO: 5.8.2 (euk_genome_min, miniprot) / Lineage: lophotrochozoa_odb12 (genomes:75, BUSCOs:1252)

HiC contact map of curated assembly

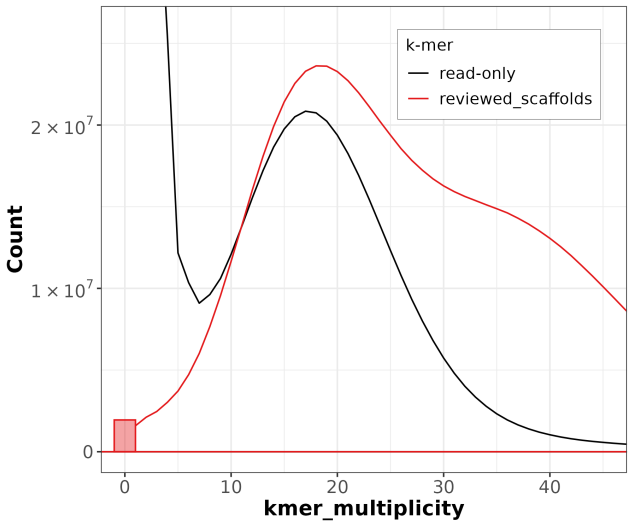


collapsed [LINK]

K-mer spectra of curated assembly

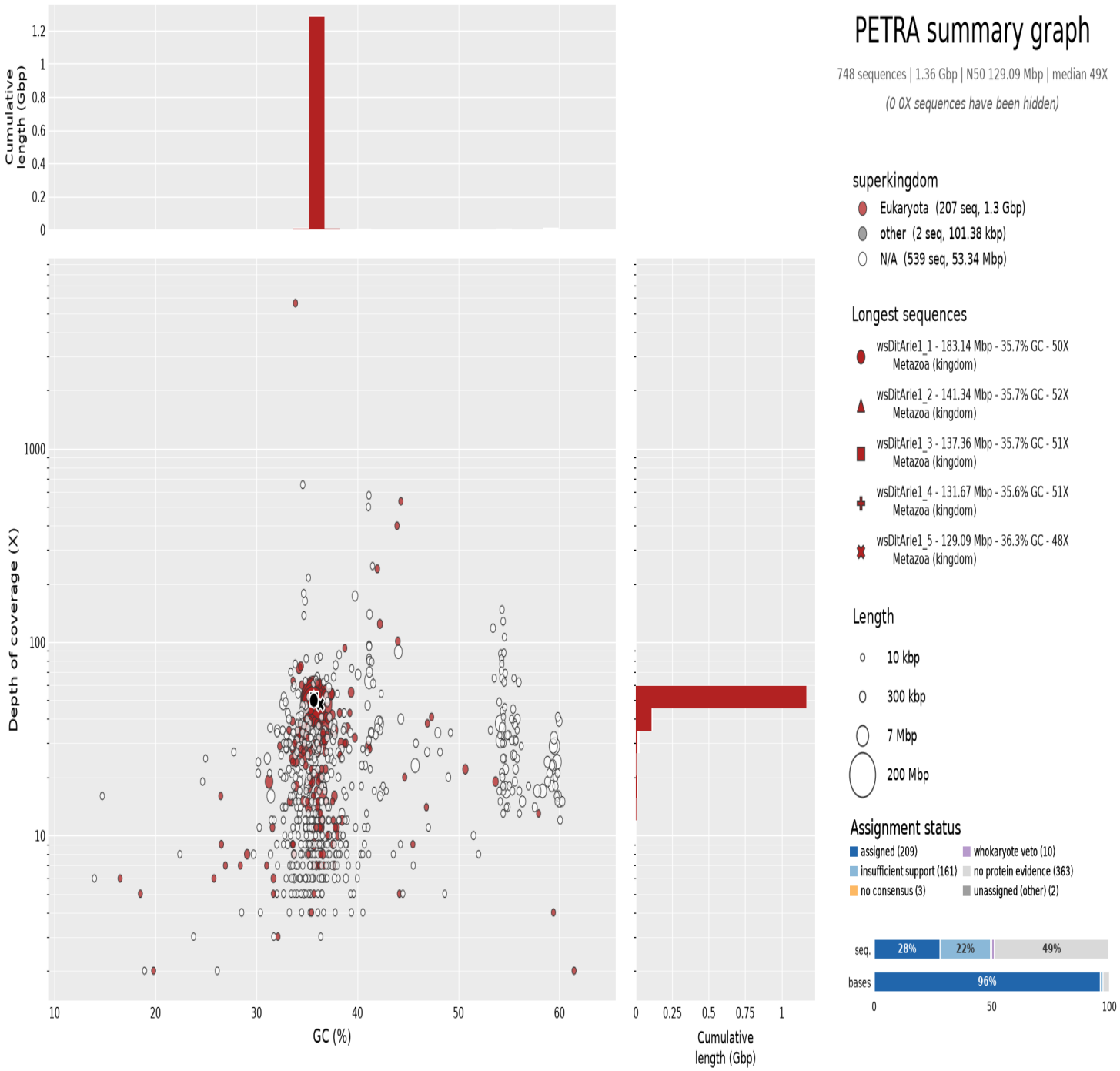


Distribution of k-mer counts per copy numbers found in asm



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	Omnic,Arima
Coverage	59	314

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:* 0.2.4
  |_ *key param:* NA
- **PretextView**
  |_ *ver:* 1.0.7
  |_ *key param:* NA

Submitter: Eleanore Lacoste
Affiliation: Genoscope

Date and time: 2026-09-29 12:03:09 CEST