ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	3034287
ToLID	**xgDerLasi**
Species	Deroceras lasithionense
Class	Gastropoda
Order	Stylommatophora

Genome Traits	Expected	Observed
Haploid size (bp)	1,014,304,119	1,318,353,577
Haploid Number	16 (source: ancestor)	31
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q55

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

Curator notes

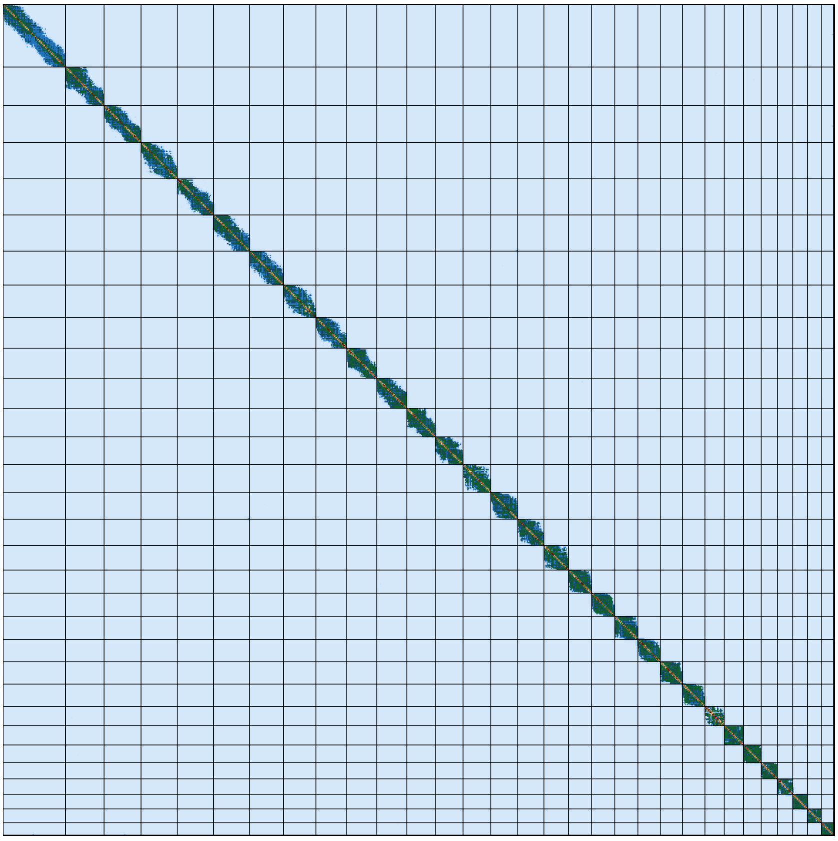
. Interventions/Gb: 73
. Contamination notes: ""
. Other observations: "The assembly of Deroceras lasithionense (xgDerLasi3) is based on 45X PacBio data and Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe project (BGE, https://biodiversitygenomics.eu/). 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, 3 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 1.2 Mb (with the largest being 1.2 Mb). Additionally, 1.208 regions totaling 342 Mb (with the largest being 3.1 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, 26 haplotypic regions and 1 mitochondrial sequence were removed, totaling 68 Mb and 65 Kb, respectively (with the largest being 47 Mb and 65 Kb). 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	1,330,276,266	1,318,353,577
GC %	41.44	41.45
Gaps/Gbp	798.33	811.62
Total gap bp	106,200	109,500
Scaffolds	90	65
Scaffold N50	45,757,421	45,757,421
Scaffold L50	12	12
Scaffold L90	26	26
Contigs	1,152	1,135
Contig N50	2,037,934	2,037,218
Contig L50	190	188
Contig L90	630	628
QV	55.0595	55.055
Kmer compl.	77.1708	76.6606
BUSCO sing.	98.0%	98.0%
BUSCO dupl.	1.6%	1.6%
BUSCO frag.	0.0%	0.0%
BUSCO miss.	0.4%	0.4%

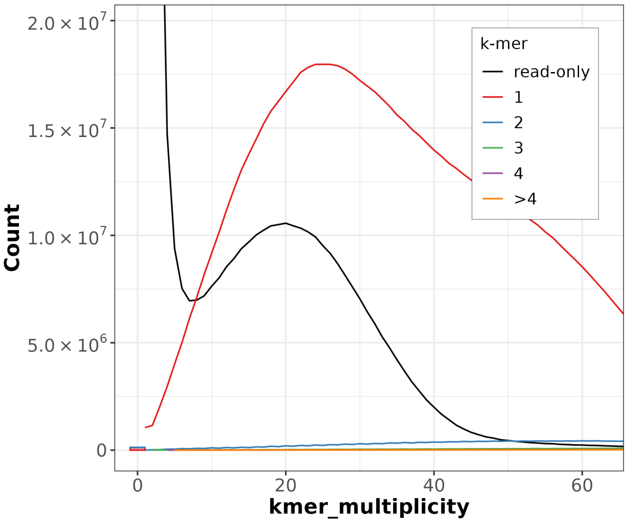
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: eukaryota_odb10 (genomes:70, BUSCOs:255)

HiC contact map of curated assembly

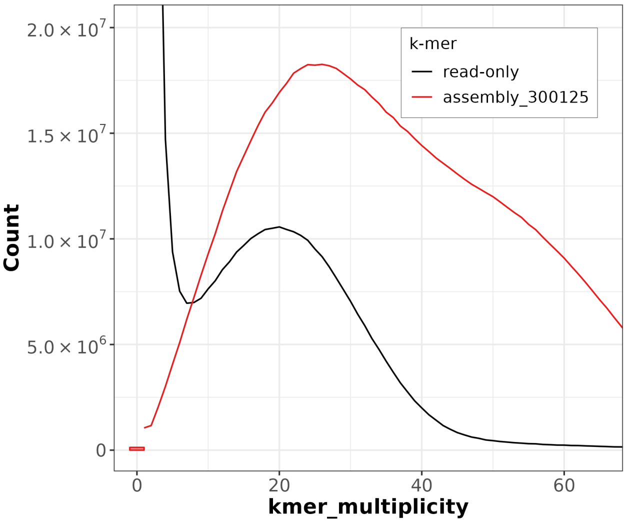


collapsed [LINK]

K-mer spectra of curated assembly

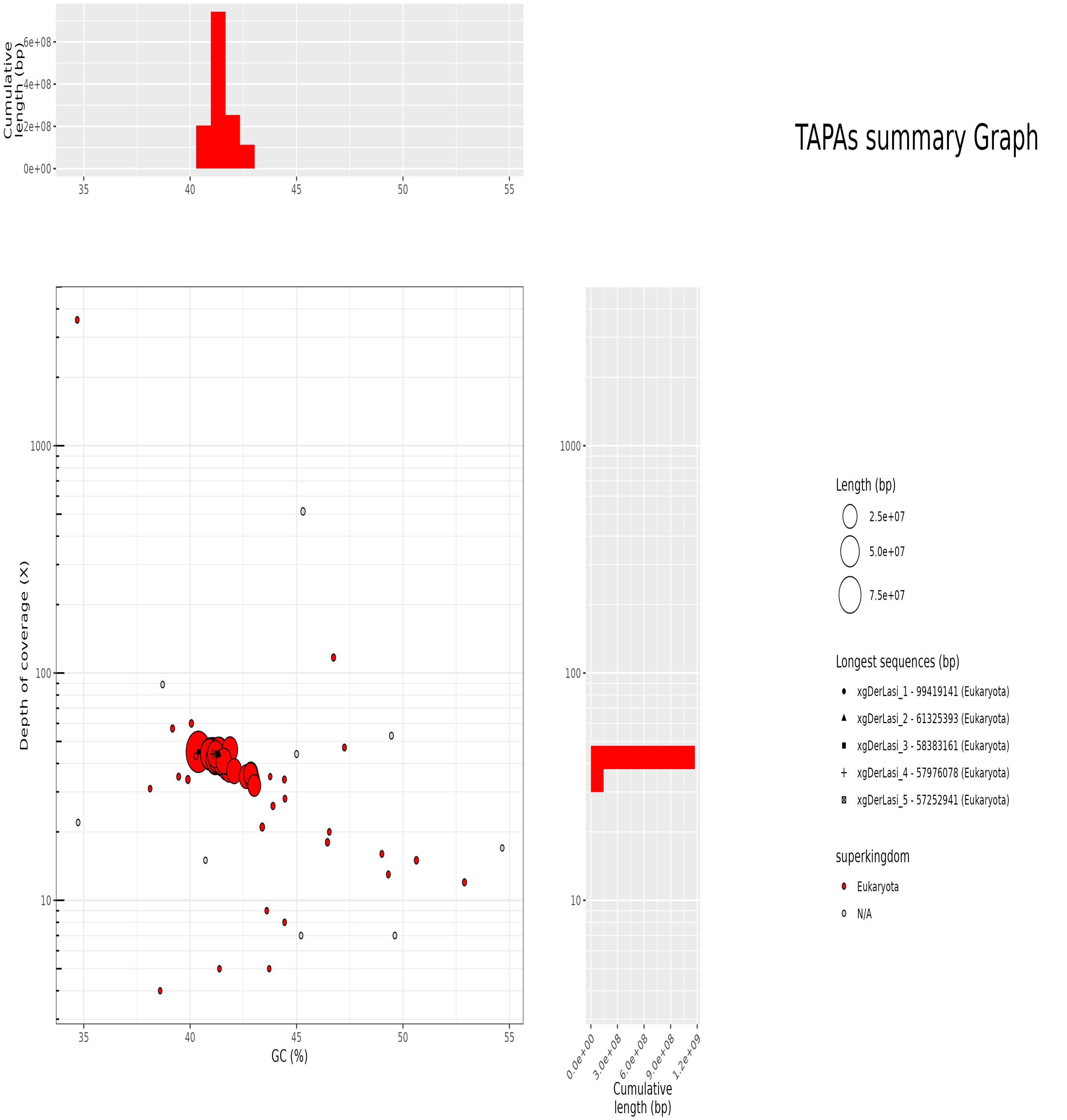


Distribution of k-mer counts per copy numbers found in asm



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	45	42

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: Caroline Belser
Affiliation: Genoscope

Date and time: 2025-01-31 09:45:58 CET