ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	2712997
ToLID	**icDenFora10**
Species	Dendarus foraminosus
Class	Insecta
Order	Coleoptera

Genome Traits	Expected	Observed
Haploid size (bp)	557,593,036	593,991,149
Haploid Number	10 (source: ancestor)	11
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q62

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
. Not 90% of assembly in chromosomes for collapsed

Curator notes

. Interventions/Gb: 15
. Contamination notes: ""
. Other observations: "The assembly of Dendarus foraminosus (icDenFora10.1) is based on 58X PacBio data and 82X OmniC 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, 13 contigs were identified as contaminants (bacterial), totaling 840 Kb (with the largest being 358 Kb). Additionally, 255 regions totaling 30.5 Mb (with the largest being 1.6 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using OATK ; two circular forms were assembled and named mitochondrion_1, mitochondrion_2. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, no regions were tagged as allelic duplications nor contaminants ; Scaffold_10 and Scaffold_11 with low coverage, and homology with Tenebrio molitor (X,Y) chromosomes, were renamed as X and Y respectively. Telomeric

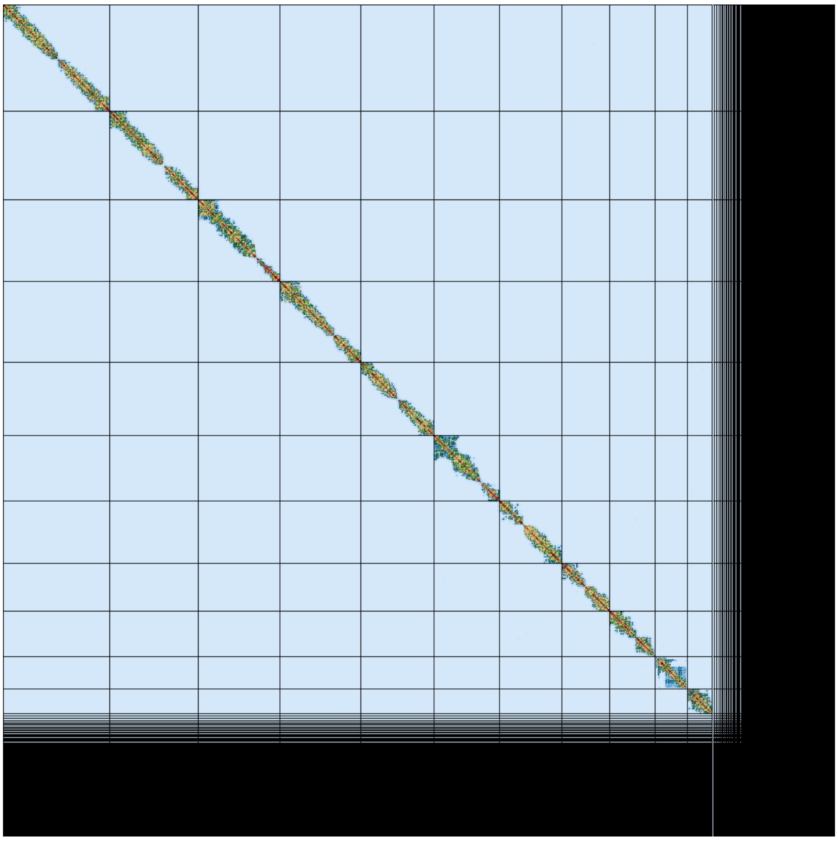
repeat pattern found is TCGGG. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	596,929,933	593,991,149
GC %	35.94	35.92
Gaps/Gbp	21.78	25.25
Total gap bp	1,300	2,000
Scaffolds	444	417
Scaffold N50	41,704,000	51,873,200
Scaffold L50	6	5
Scaffold L90	53	49
Contigs	457	432
Contig N50	23,820,166	24,404,000
Contig L50	9	9
Contig L90	65	63
QV	49.8095	62.293
Kmer compl.	66.5019	66.7811
BUSCO sing.	96.6%	97.0%
BUSCO dupl.	0.7%	0.6%
BUSCO frag.	1.7%	0.6%
BUSCO miss.	1.0%	1.8%

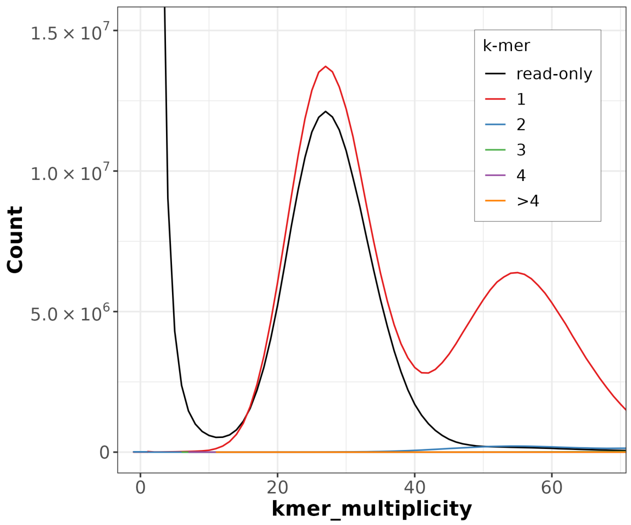
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 5.8.2 (euk_genome_min, miniprot) / Lineage: polyphaga_odb12 (genomes:60, BUSCOs:4010)
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: polyphaga_odb12 (genomes:60, BUSCOs:4010)

HiC contact map of curated assembly

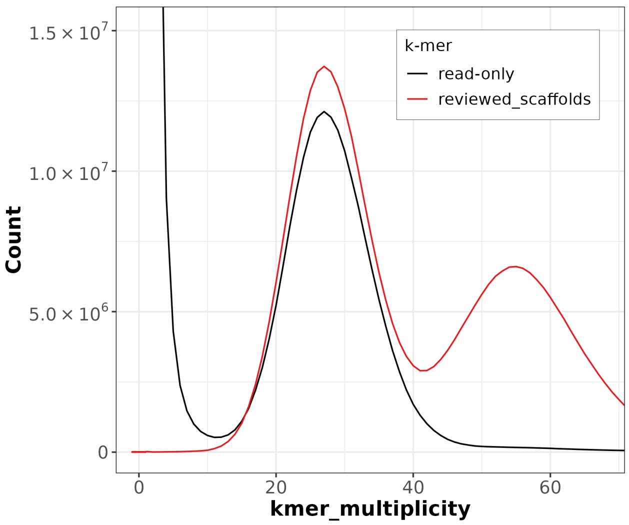


collapsed [LINK]

K-mer spectra of curated assembly

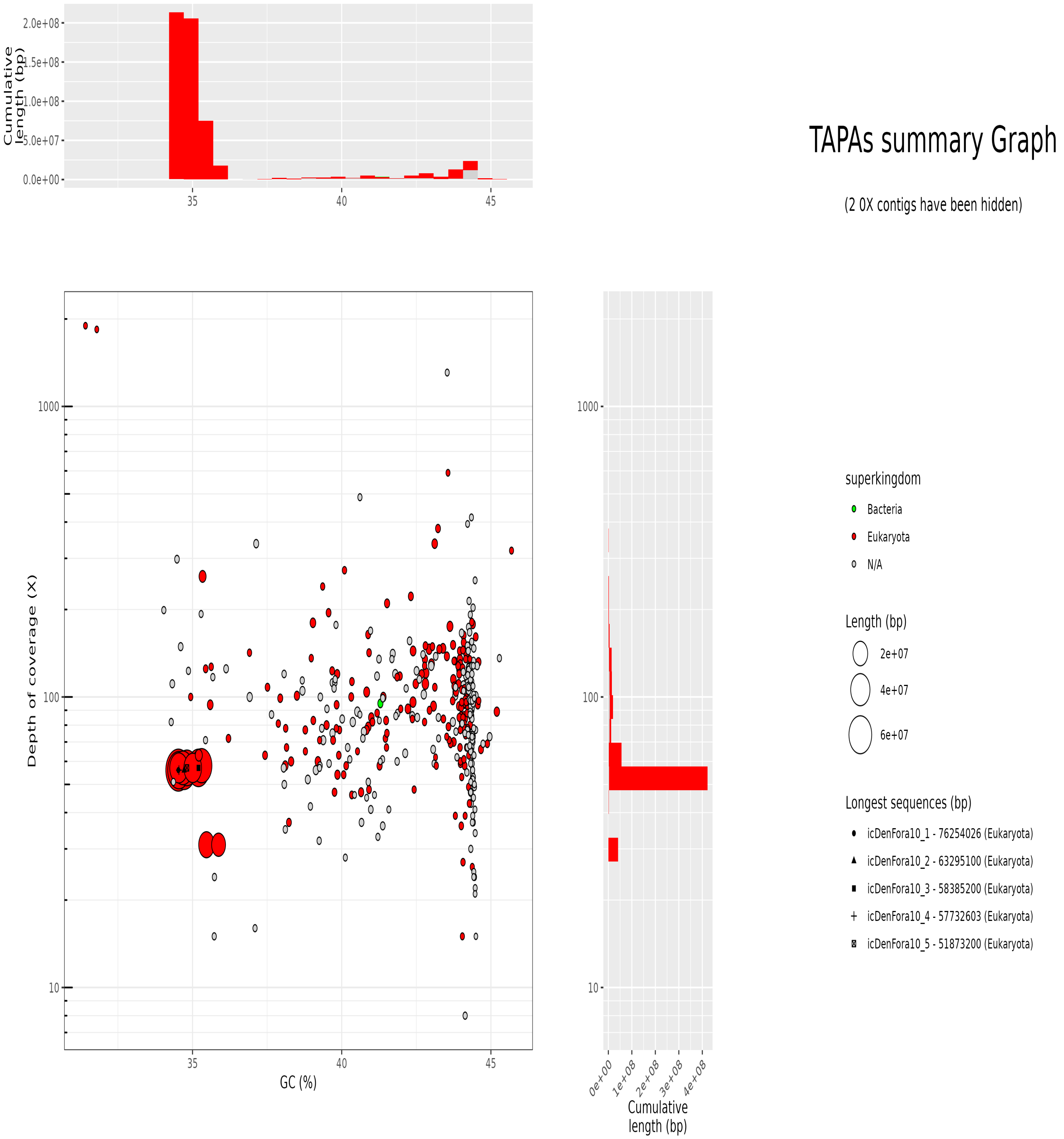


Distribution of k-mer counts per copy numbers found in asm



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	Omnic,Arima
Coverage	54	78

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: Simone Duprat
Affiliation: Genoscope

Date and time: 2025-04-10 23:02:28 CEST