ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	538483
ToLID	**daCarDiae1**
Species	Carlina diae
Class	Magnoliopsida
Order	Asterales

Genome Traits	Expected	Observed
Haploid size (bp)	2,130,449,475	4,151,721,093
Haploid Number	10 (source: direct)	10
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q72

The following metrics were automatically flagged as below EBP recommended standards or different from expected:

. Observed Haploid size (bp) has >20% difference with Expected

. BUSCO duplicated value is more than 5% for collapsed

Curator notes

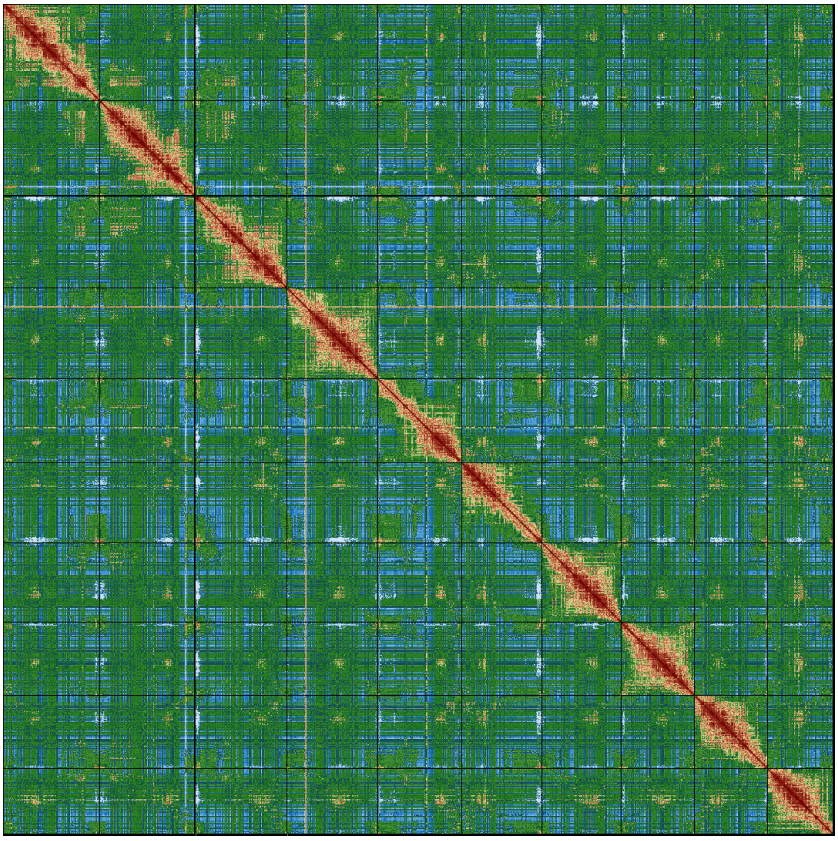
. Interventions/Gb: 1
. Contamination notes: ""
. Other observations: "The assembly of CARLINA DIAE (daCarDiae1) is based on 61X PacBio data and 210X Omni-C Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe 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. No contigs were found to be contaminants (bacterial, archaeal, or viral). But, 1,107 regions totaling 94 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, no regions were tagged as allelic duplications or contaminants. The telemeric track is based on TAGTAG, because the usual pattern was not found. 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	4,151,810,759	4,151,721,093
GC %	39.27	39.27
Gaps/Gbp	84.3	85.27
Total gap bp	35,000	36,400
Scaffolds	125	105
Scaffold N50	416,577,109	416,577,109
Scaffold L50	5	5
Scaffold L90	9	9
Contigs	475	459
Contig N50	23,495,317	23,495,317
Contig L50	55	55
Contig L90	191	191
QV	72.5913	72.7263
Kmer compl.	98.1121	98.1119
BUSCO sing.	91.1%	91.1%
BUSCO dupl.	6.8%	6.8%
BUSCO frag.	0.4%	0.4%
BUSCO miss.	1.7%	1.7%

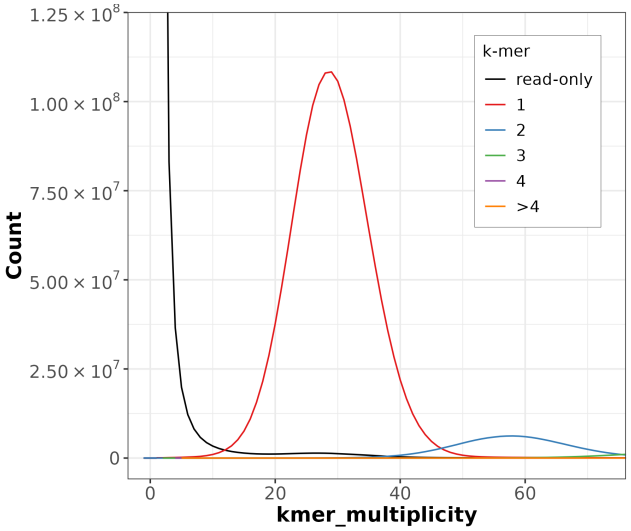
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: embryophyta_odb10 (genomes:50, BUSCOs:1614)

HiC contact map of curated assembly

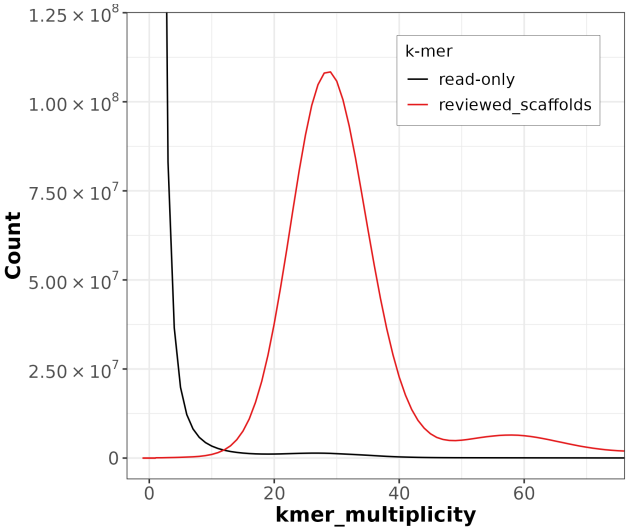


collapsed [LINK]

K-mer spectra of curated assembly

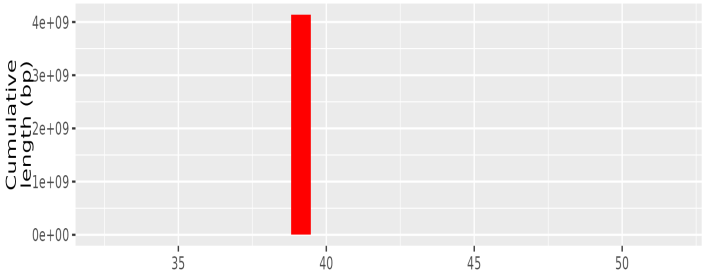


Distribution of k-mer counts per copy numbers found in asm

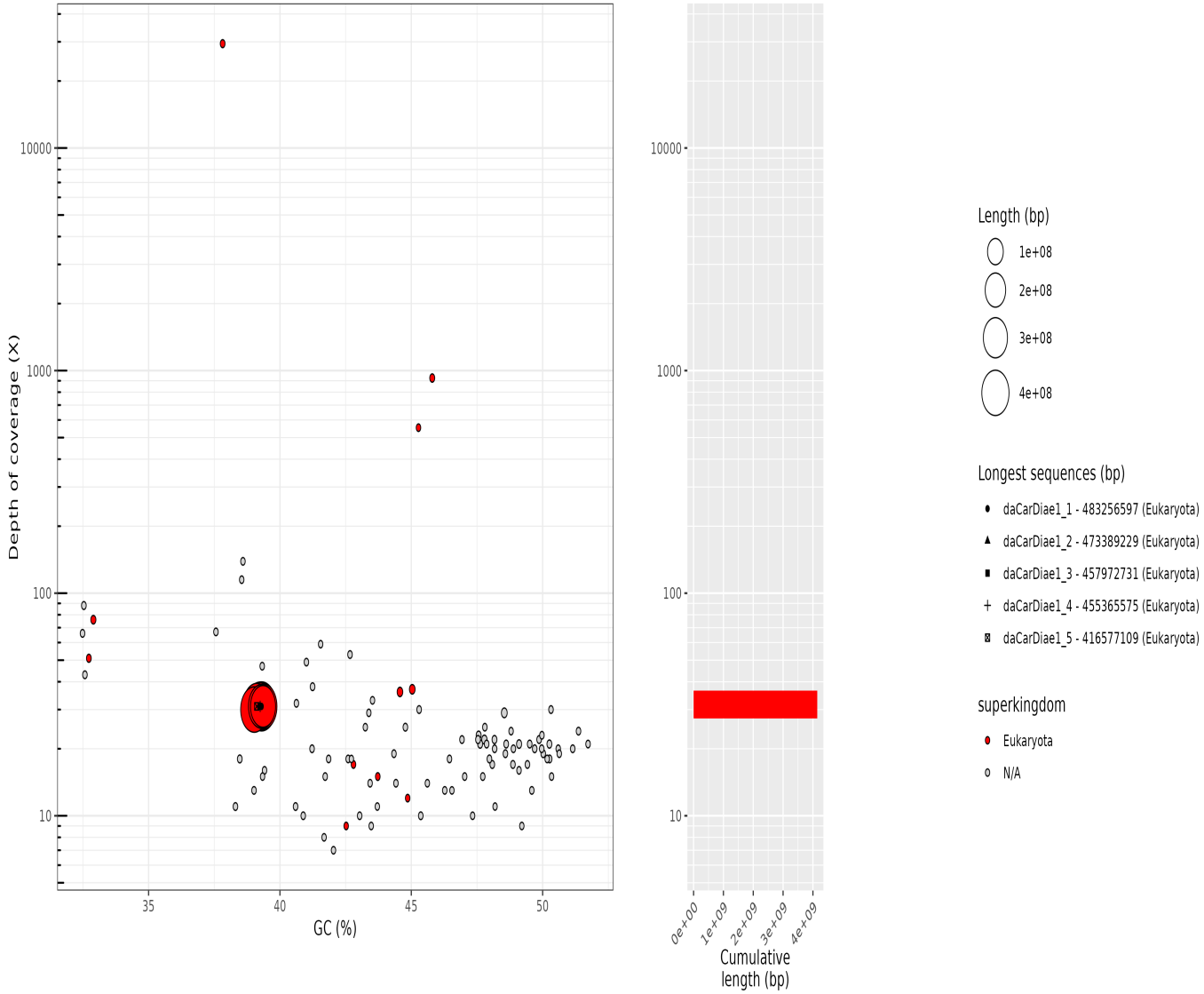


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



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
Coverage	61	273

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: Lola Demirdjian
Affiliation: Genoscope

Date and time: 2025-03-02 12:18:58 CET