ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	335185
ToLID	**ddCisCris1.1**
Species	Cistus crispus
Class	Magnoliopsida
Order	Malvales

Genome Traits	Expected	Observed
Haploid size (bp)	1,635,821,345	1,618,429,185
Haploid Number	9 (source: direct)	9
Ploidy	2 (source: direct)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 8.8.Q68

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

. Kmer completeness value is less than 90 for collapsed

Curator notes

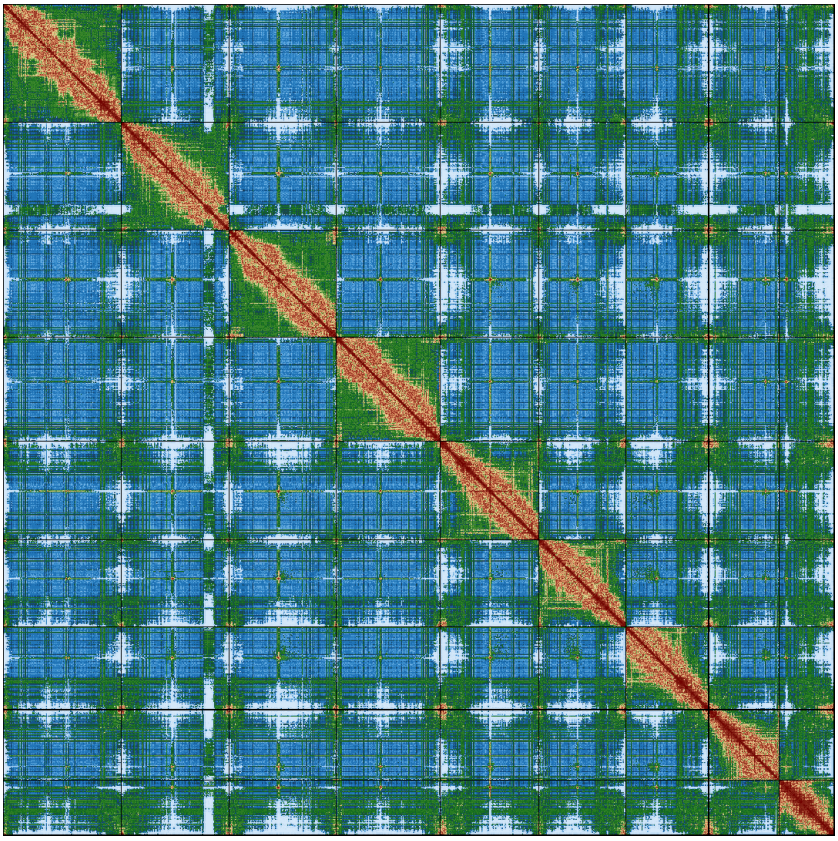
. Interventions/Gb: 18
. Contamination notes: ""
. Other observations: "The assembly of CISTUS CRISPUS (ddCisCris1) is based on 54X PacBio data and 137X Omni-C Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe 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. In total, 128 contigs were identified as contaminants (bacterial, archaeal, or viral) totaling 5.4 Mb, scaffolds classified as Arthropoda or Metazoa by Context were removed, as were those without taxonomic assignment but which aligned, even slightly, with an arthropod in blastn. Additionally, 699 regions totaling 135 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 as contaminants. 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	1,640,629,704	1,618,429,185
GC %	38.85	38.83
Gaps/Gbp	13.41	9.89
Total gap bp	2,200	2,200
Scaffolds	173	37
Scaffold N50	201,128,670	201,400,918
Scaffold L50	4	4
Scaffold L90	8	8
Contigs	195	53
Contig N50	128,040,706	128,040,706
Contig L50	5	5
Contig L90	13	13
QV	66.5688	68.0869
Kmer compl.	88.1678	87.5541
BUSCO sing.	91.1%	91.2%
BUSCO dupl.	4.9%	4.8%
BUSCO frag.	0.4%	0.4%
BUSCO miss.	3.6%	3.6%

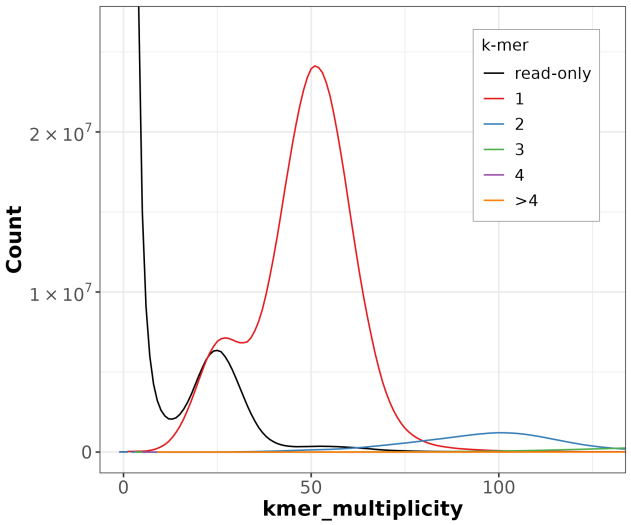
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: embryophyta_odb10 (genomes:50, BUSCOs:1614)

HiC contact map of curated assembly

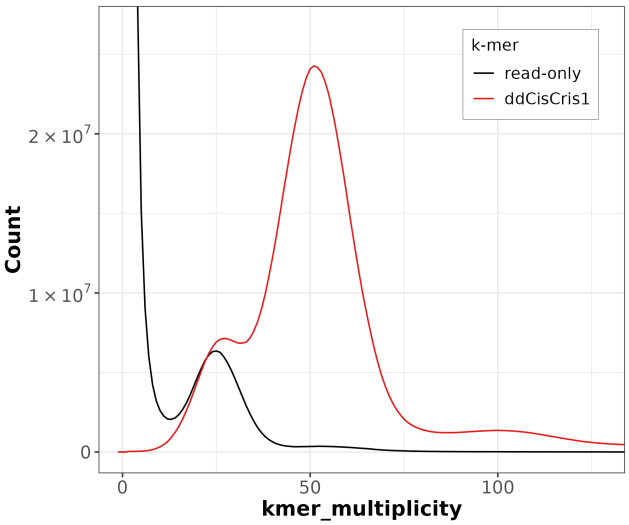


collapsed [LINK]

K-mer spectra of curated assembly

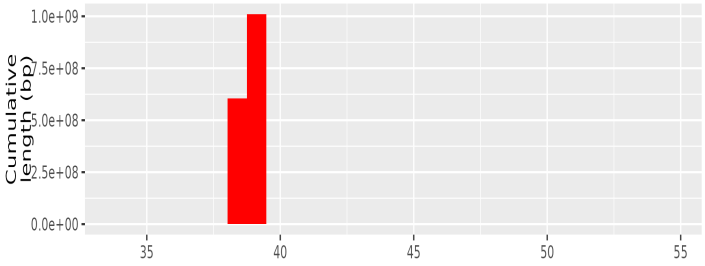


Distribution of k-mer counts per copy numbers found in asm

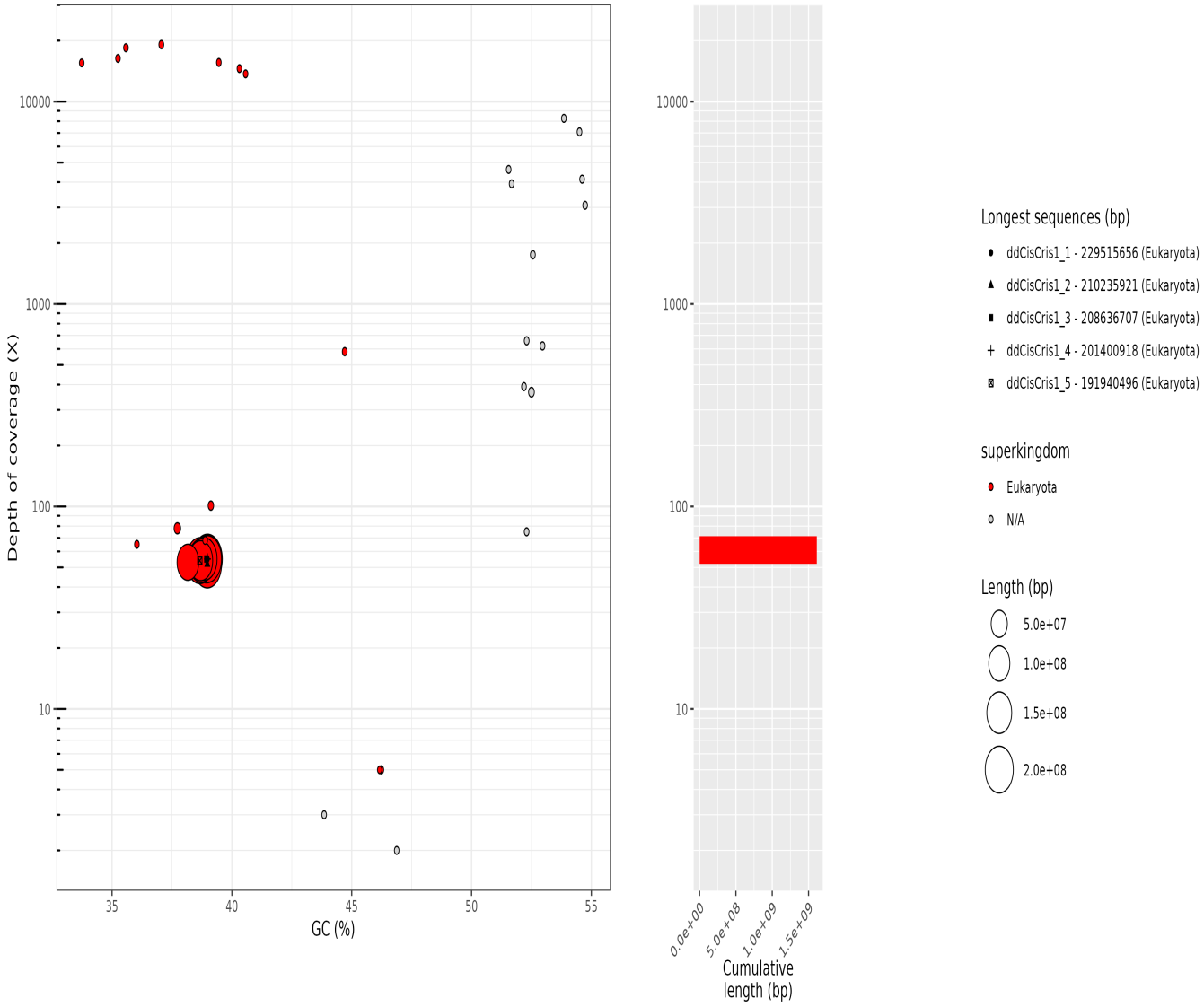


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	54	137

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: 2025-03-03 12:15:57 CET