ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	3116505
ToLID	**icCedAzor3**
Species	Cedrorum azoricus
Class	Insecta
Order	Coleoptera

Genome Traits	Expected	Observed
Haploid size (bp)	729,929,118	567,877,672
Haploid Number	17 (source: ancestor)	16
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q65

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
. Not 90% of assembly in chromosomes for collapsed

Curator notes

. Interventions/Gb: 292
. Contamination notes: ""
. Other observations: "The assembly of CEDRORUM AZORICUS (icCedAzor3) is based on 43X PacBio data and 189X 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 (and no contig correction). In total, 9 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 11 Mb (with the largest being 2 Mb). Additionally, 730 regions totaling 427 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. Manual curation was performed using the normal resolution map, and the genome was reviewed on the BGE GitHub with a normal resolution q0 map (find in the EAR.pdf). Every effort was made to improve this genome, but full

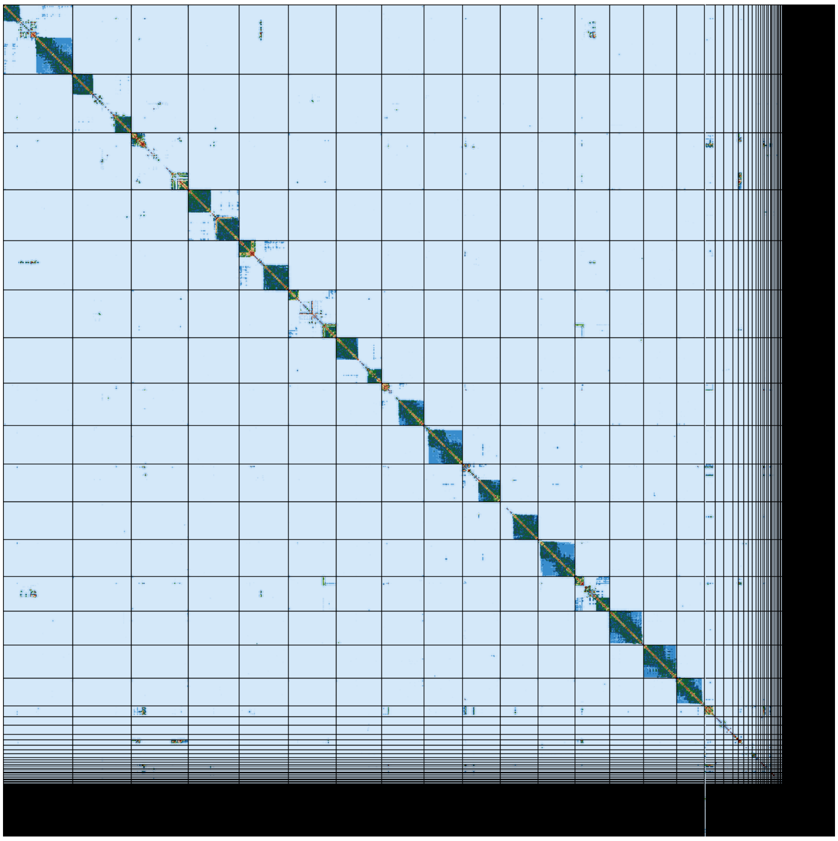
confidence has not been reached. During manual curation, 5 haplotypic regions were removed, totaling 416,211 pb (with the largest being 156,079 pb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	568,266,605	567,877,672
GC %	28.76	28.75
Gaps/Gbp	22.88	105.66
Total gap bp	1,300	11,100
Scaffolds	504	302
Scaffold N50	12,493,000	28,853,378
Scaffold L50	15	8
Scaffold L90	120	23
Contigs	517	362
Contig N50	8,810,000	10,971,072
Contig L50	20	16
Contig L90	132	69
QV	65.7835	65.6817
Kmer compl.	84.2012	84.1826
BUSCO sing.	98.4%	98.4%
BUSCO dupl.	0.5%	0.5%
BUSCO frag.	0.4%	0.4%
BUSCO miss.	0.7%	0.7%

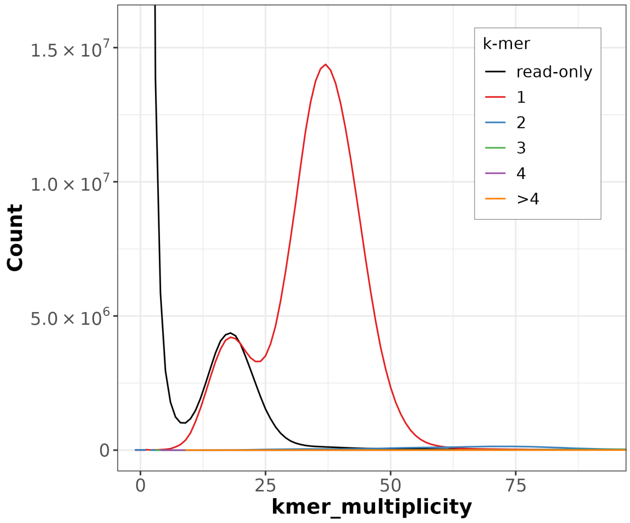
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: endopterygota_odb10 (genomes:56, BUSCOs:2124)

HiC contact map of curated assembly

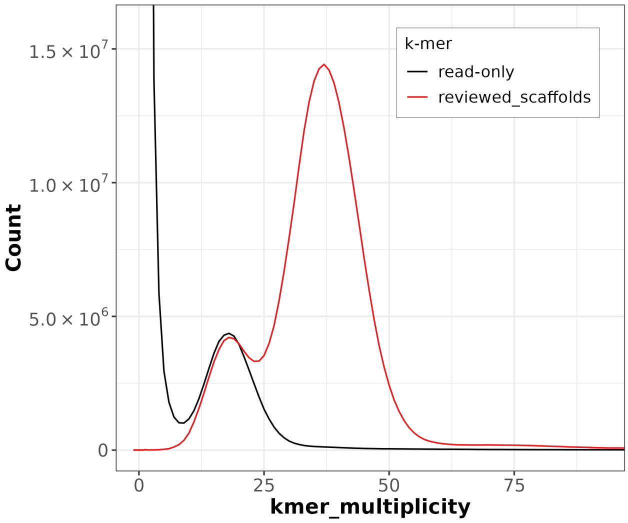


collapsed [LINK]

K-mer spectra of curated assembly

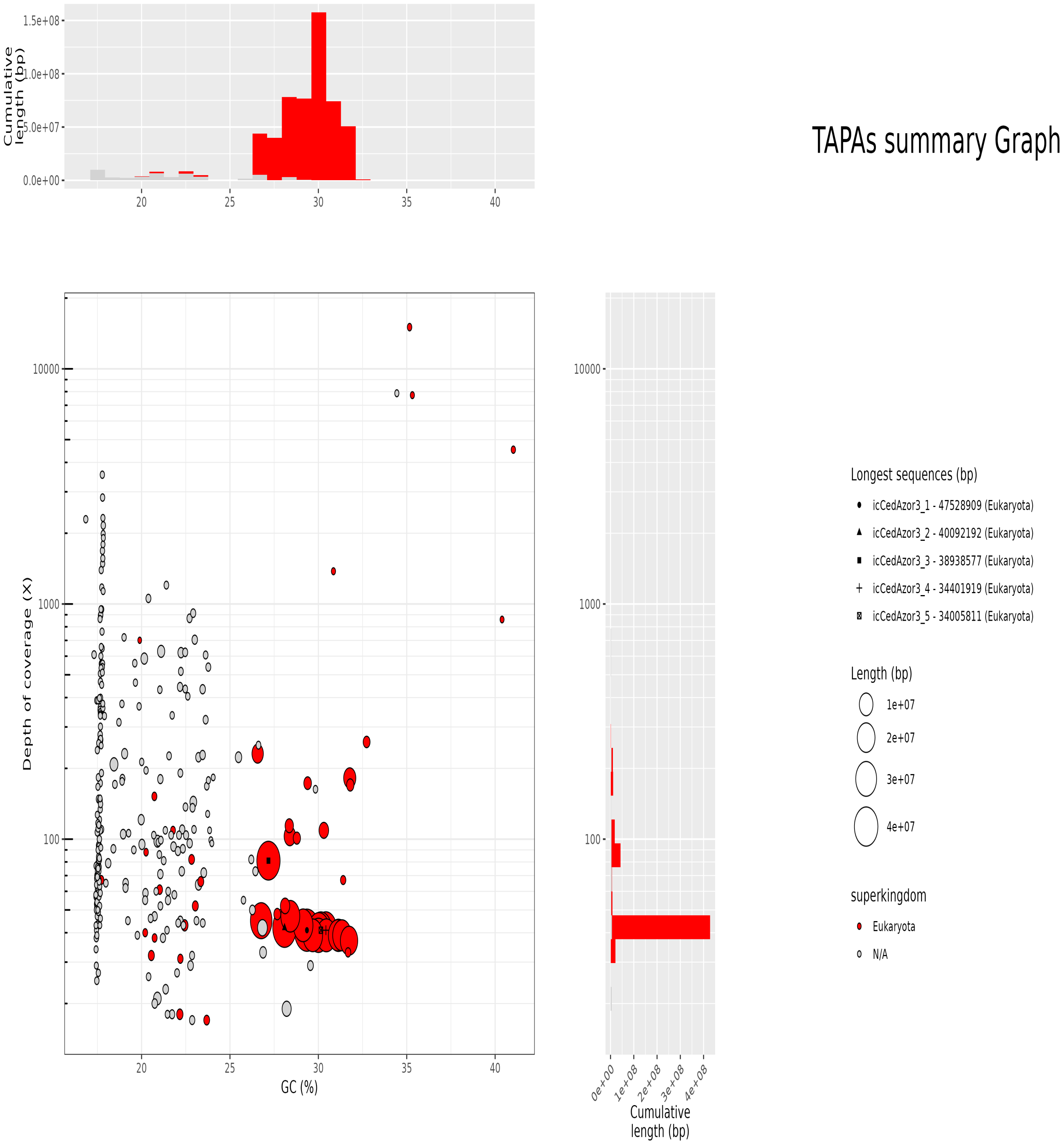


Distribution of k-mer counts per copy numbers found in asm



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	43	189

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: 2025-05-28 06:32:16 CEST