ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	3116505
ToLID	**icCedAzor3.1**
Species	Cedrorum azoricus
Class	Insecta
Order	Coleoptera

Genome Traits	Expected	Observed
Haploid size (bp)	729,929,118	567,690,156
Haploid Number	19 (source: ancestor)	21
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.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: 283
. Contamination notes: ""
. Other observations: "The assembly of CEDRORUM AZORICUS (icCedAzor3) is based on 44X PacBio data and 244X 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, 9 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 11 Mb (with the largest being 2 Mb).There are also two scaffolds that are classified as viruses, but they are "Errantiviruses", i.e. probably transposons and/or viruses integrated into the genome. 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. During manual curation, 12 haplotypic regions were removed, totaling 593,827 pb (with the largest

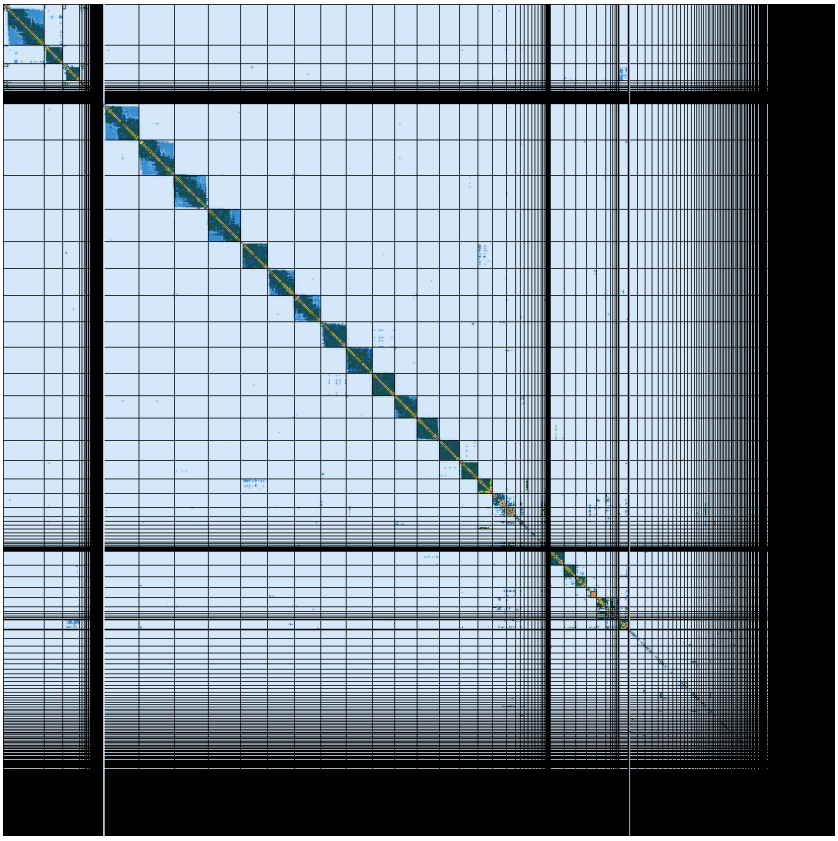
being 86,711 pb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	568,265,305	567,690,156
GC %	28.76	28.75
Gaps/Gbp	0	21.14
Total gap bp	0	1,200
Scaffolds	358	505
Scaffold N50	10,971,072	12,493,000
Scaffold L50	16	15
Scaffold L90	66	120
Contigs	358	517
Contig N50	10,971,072	9,054,958
Contig L50	16	20
Contig L90	66	131
QV	65.7836	65.5244
Kmer compl.	84.2012	84.1588
BUSCO sing.	98.4%	98.3%
BUSCO dupl.	0.5%	0.5%
BUSCO frag.	0.4%	0.4%
BUSCO miss.	0.7%	0.8%

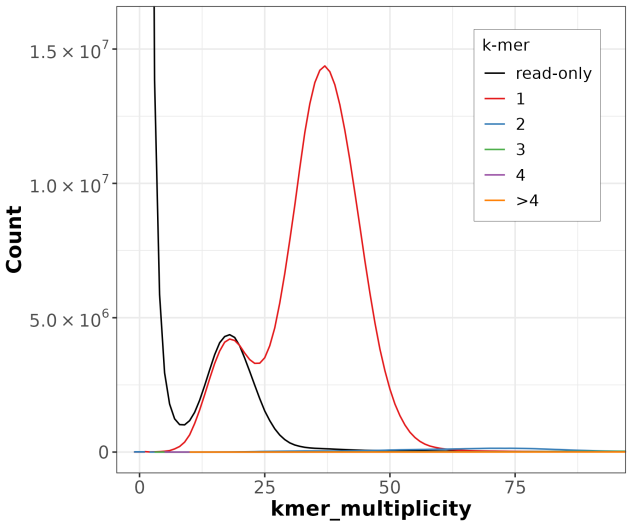
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: endopterygota_odb10 (genomes:56, BUSCOs:2124)

HiC contact map of curated assembly

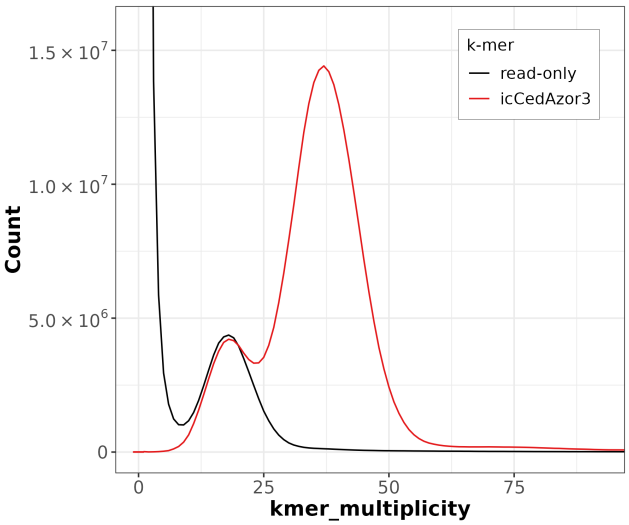


collapsed [LINK]

K-mer spectra of curated assembly

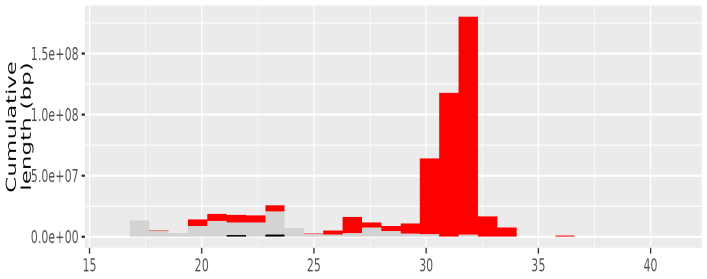


Distribution of k-mer counts per copy numbers found in asm

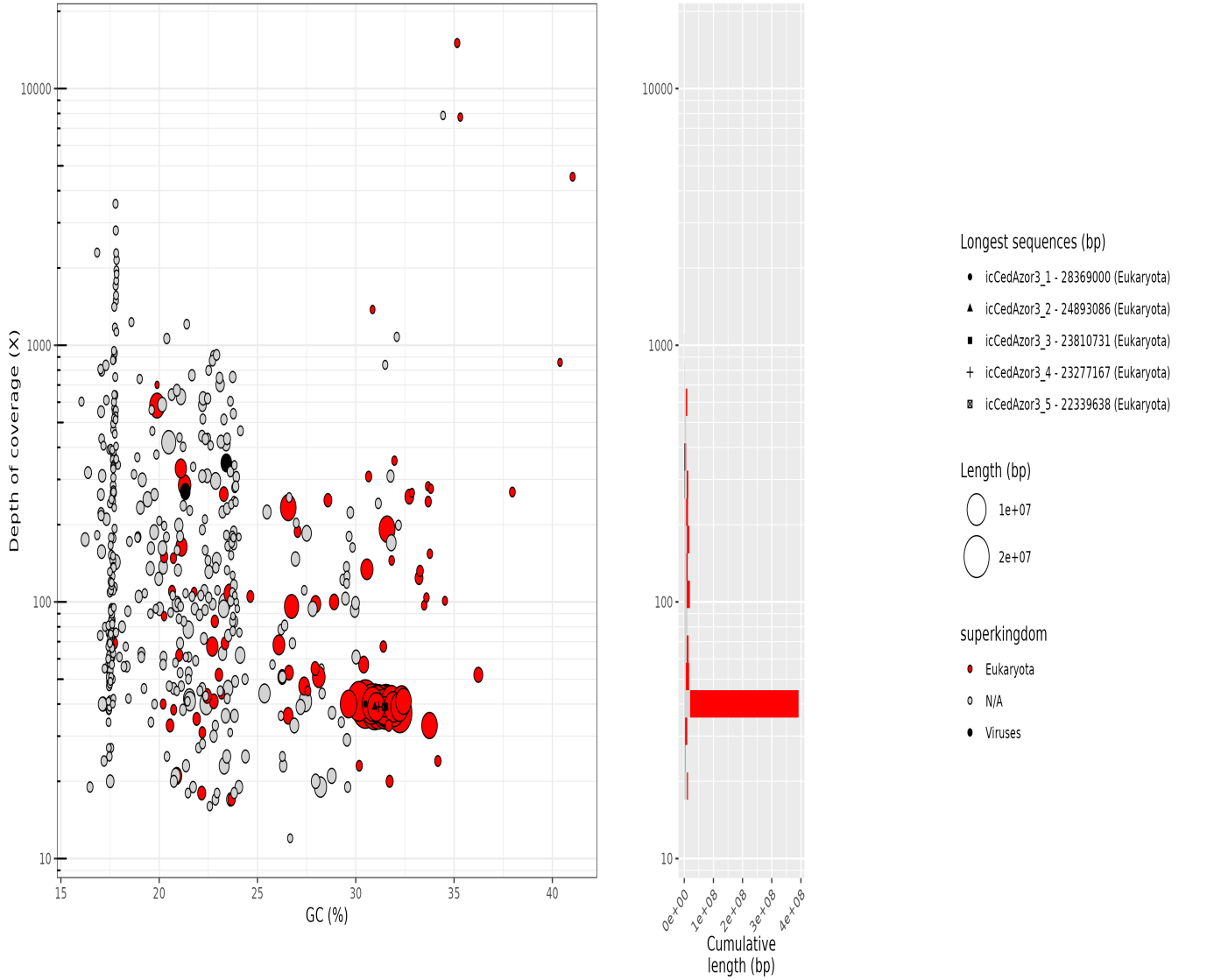


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



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	Arima
Coverage	44	244

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: 2025-02-14 14:35:52 CET