ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	997549
ToLID	**ilCalTrid2**
Species	Calamia tridens
Class	Insecta
Order	Lepidoptera

Genome Traits	Expected	Observed
Haploid size (bp)	696,390,245	694,406,941
Haploid Number	30 (source: ancestor)	30
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q72

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

. Kmer completeness value is less than 90 for collapsed
. Assembly length loss > 3% for collapsed

Curator notes

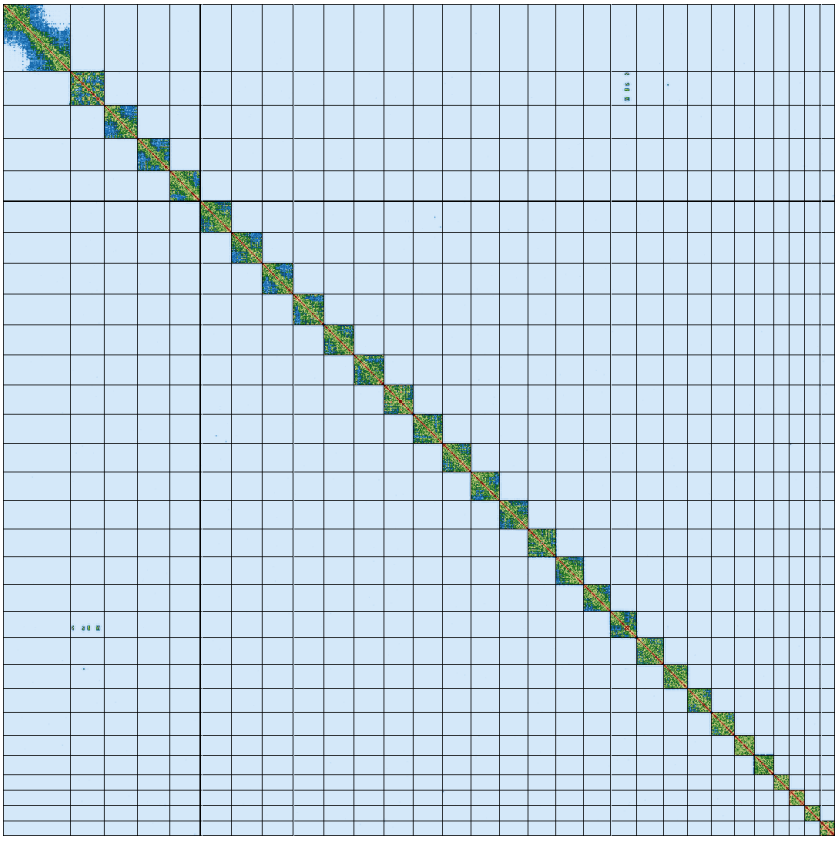
. Interventions/Gb: 46
. Contamination notes: ""
. Other observations: "The assembly of CALAMIA TRIDENS (ilCalTrid2) is based on 68X PacBio data and 132X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe project (BGE, https://biodiversitygenomics.eu/). The assembly process included the following steps: initial PacBio assembly generation with Hifiasm, removal of contaminant sequences using Context and Hi-C-based scaffolding with YaHS. In total, 2 contigs were identified as contaminants (bacterial), totaling 64,299 pb (with the largest being 52,155 pb). We did not use purge_dups to remove haplotype duplications, as it did not work properly on this genome. The mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual selection, 4 haplotype regions were removed, totaling 13,402,842 bp (the largest being 6,868,000 bp) and 8 scaffolds were identified as contaminants and removed. 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	718,006,249	694,406,941
GC %	37.6	37.85
Gaps/Gbp	9.75	10.08
Total gap bp	700	1,200
Scaffolds	448	54
Scaffold N50	24,184,673	24,184,673
Scaffold L50	13	13
Scaffold L90	27	25
Contigs	455	61
Contig N50	23,752,865	23,752,865
Contig L50	13	13
Contig L90	28	26
QV	64.2062	72.0684
Kmer compl.	71.3141	70.6023
BUSCO sing.	96.4%	97.9%
BUSCO dupl.	1.9%	0.4%
BUSCO frag.	0.7%	0.8%
BUSCO miss.	1.0%	1.0%

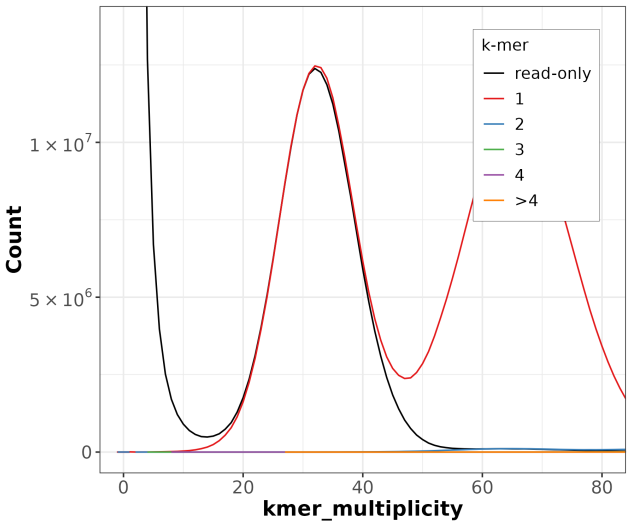
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: lepidoptera_odb12 (genomes:79, BUSCOs:5760)

HiC contact map of curated assembly

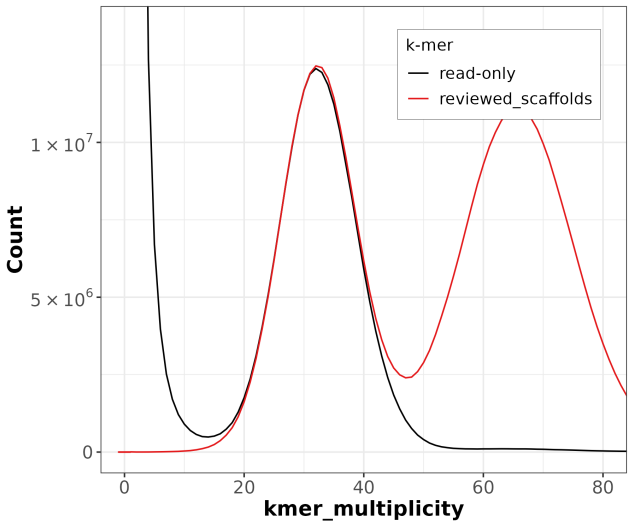


collapsed [LINK]

K-mer spectra of curated assembly

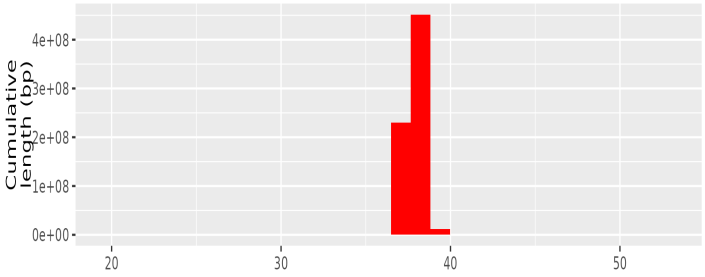


Distribution of k-mer counts per copy numbers found in asm

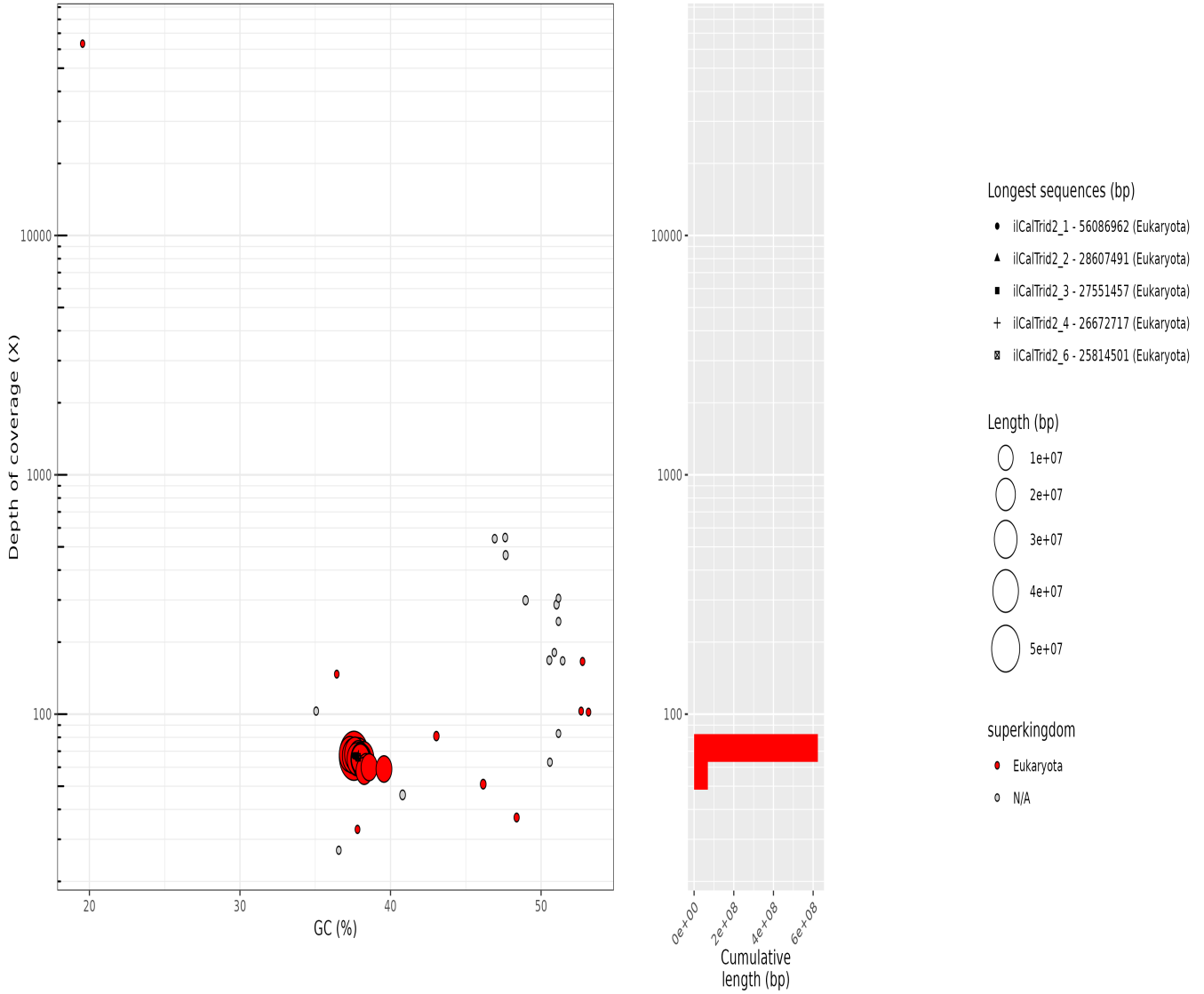


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	Long reads	Arima
Coverage	68	132

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: 2025-09-25 07:57:32 CEST