ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1734902
ToLID	**ilCatNymp1**
Species	Catocala nymphagoga
Class	Insecta
Order	Lepidoptera

Genome Traits	Expected	Observed
Haploid size (bp)	570,200,512	628,081,482
Haploid Number	31 (source: ancestor)	33
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.7.Q51

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

. Observed Haploid Number is different from Expected

. Kmer completeness value is less than 90 for collapsed
. More than 1000 gaps/Gbp for collapsed

Curator notes

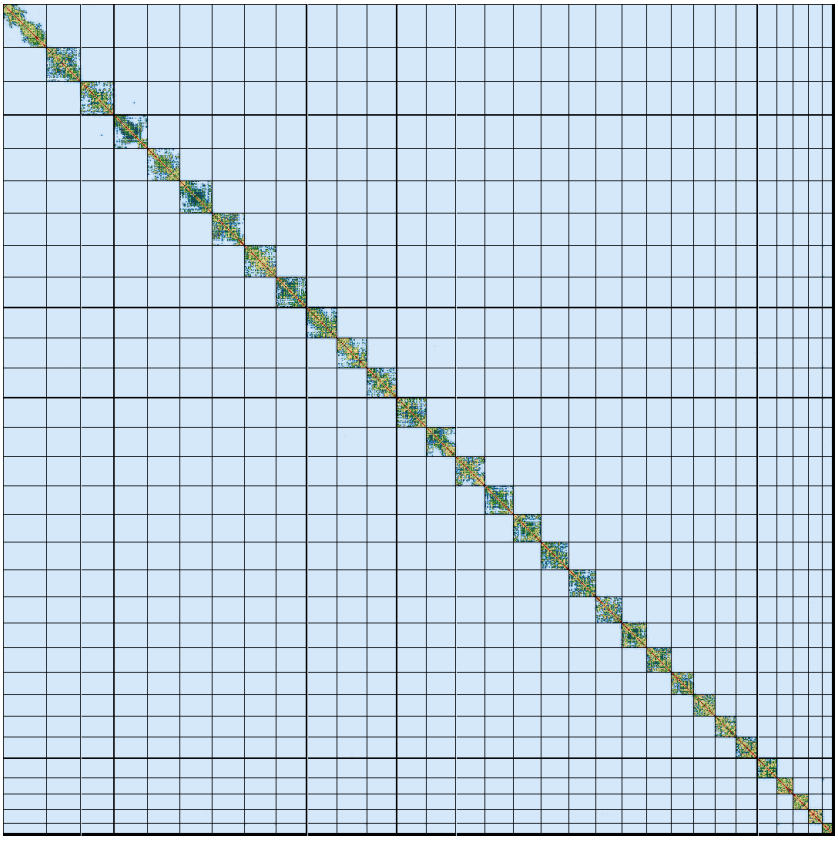
. Interventions/Gb: 189
. Contamination notes: ""
. Other observations: "The assembly of Catocala nymphagoga (ilCatNymp1) is based on 61X PacBio data and 410X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics 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, 146 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 4.9 Mb (with the largest being 327Kb). Additionally, 565 regions totaling 22 Mb (with the largest being 0.293 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, 2 haplotypic regions were removed, totaling 536 Kb(with the largest being 0.303 Kb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. Z chromosome was identified by comparison with the Catocala promissa genome assembly "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	628,680,820	628,081,482
GC %	36.59	36.59
Gaps/Gbp	2,223.7	2,286.33
Total gap bp	151,900	165,100
Scaffolds	209	145
Scaffold N50	22,011,387	22,258,276
Scaffold L50	13	13
Scaffold L90	27	26
Contigs	1,588	1,581
Contig N50	913,285	913,285
Contig L50	203	203
Contig L90	773	773
QV	51.1531	51.1661
Kmer compl.	77.4125	77.3907
BUSCO sing.	93.1%	93.1%
BUSCO dupl.	0.6%	0.6%
BUSCO frag.	3.0%	3.0%
BUSCO miss.	3.3%	3.3%

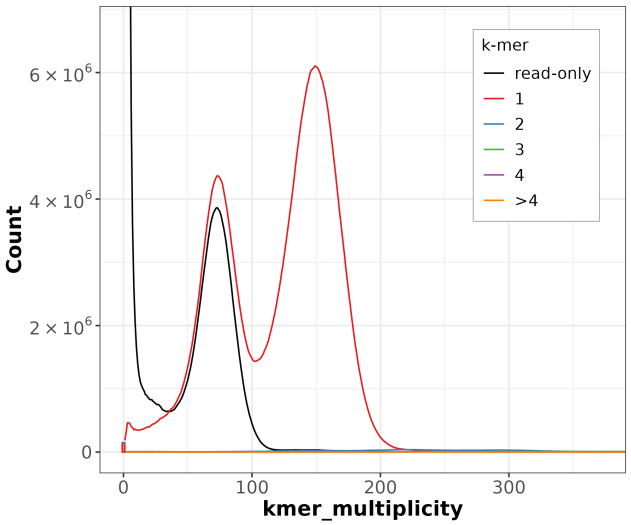
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: lepidoptera_odb12 (genomes:79, BUSCOs:5760)

HiC contact map of curated assembly

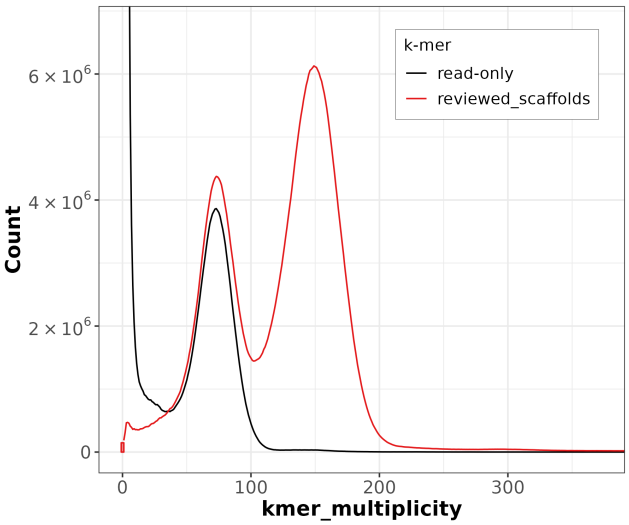


collapsed [LINK]

K-mer spectra of curated assembly

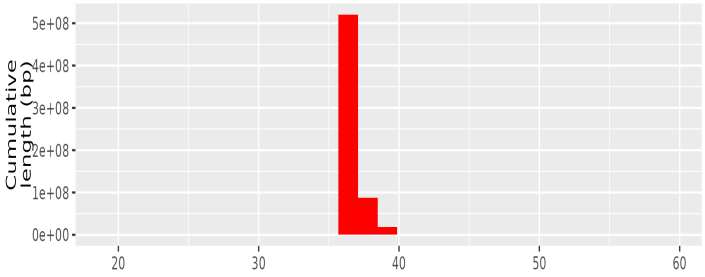


Distribution of k-mer counts per copy numbers found in asm

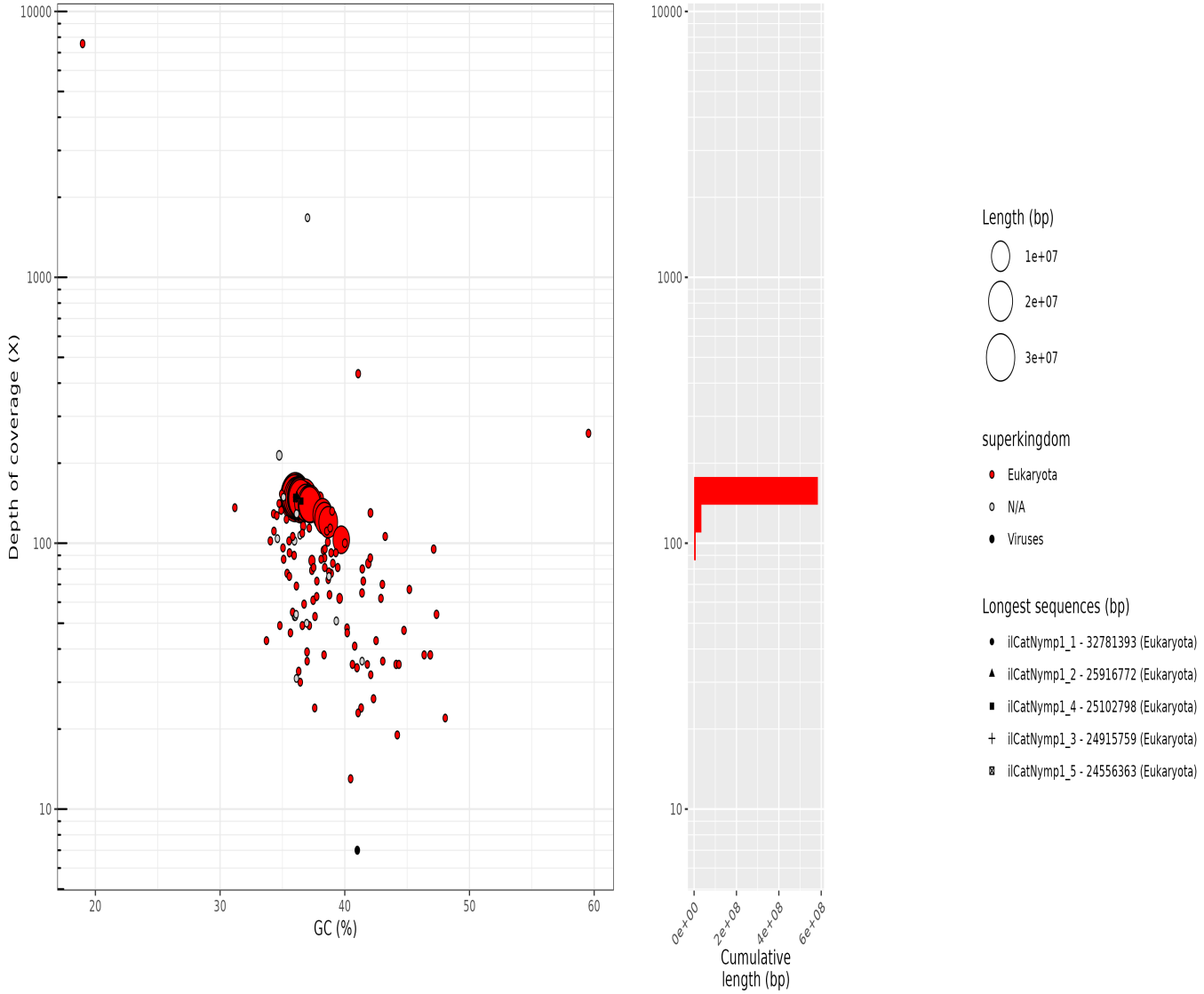


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	158	410

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: Caroline Menguy
Affiliation: Genoscope

Date and time: 2025-11-28 09:42:54 CET