ERGA Assembly Report
v24.10.15

Tags: ERGA-BGE

TxID	1086284
ToLID	**ilDolEleg1**
Species	Dolbina elegans
Class	Insecta
Order	Lepidoptera

Genome Traits	Expected	Observed
Haploid size (bp)	443,096,965	431,447,688
Haploid Number	28 (source: ancestor)	27
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q64

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

. Observed Haploid Number is different from Expected

Curator notes

. Interventions/Gb: 4
. Contamination notes: ""
. Other observations: "The assembly of DOLBINA ELEGANS (ilDolEleg1) is based on 42X PacBio data and 246X 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, 270 regions totaling 6,430,096 pb (with the largest being 164,404 pb) 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 region were removed, totaling 609,538 pb (with the largest being 164,404 pb) and 2 scaffolds were identified as contaminants and removed. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. The sex chromosome Z was identified by alignment with ilSphPina1.1. Lepidoptera Z chromosomes are well preserved, allowing their identification by synteny. Furthermore, it can be inferred that the specimen is male, given that all chromosomes are uniformly covered. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	432,271,758	431,447,688
GC %	38.46	38.47
Gaps/Gbp	18.51	16.22
Total gap bp	1,300	1,300
Scaffolds	44	32
Scaffold N50	16,422,477	16,422,477
Scaffold L50	12	12
Scaffold L90	24	24
Contigs	50	39
Contig N50	16,128,841	16,067,718
Contig L50	12	12
Contig L90	25	25
QV	64.363	64.5809
Kmer compl.	94.1064	94.0883
BUSCO sing.	97.4%	98.8%
BUSCO dupl.	0.4%	0.3%
BUSCO frag.	1.3%	0.1%
BUSCO miss.	1.0%	0.8%

Warning! BUSCO versions or lineage datasets are not the same across results:

BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: lepidoptera_odb12 (genomes:79, BUSCOs:5760)
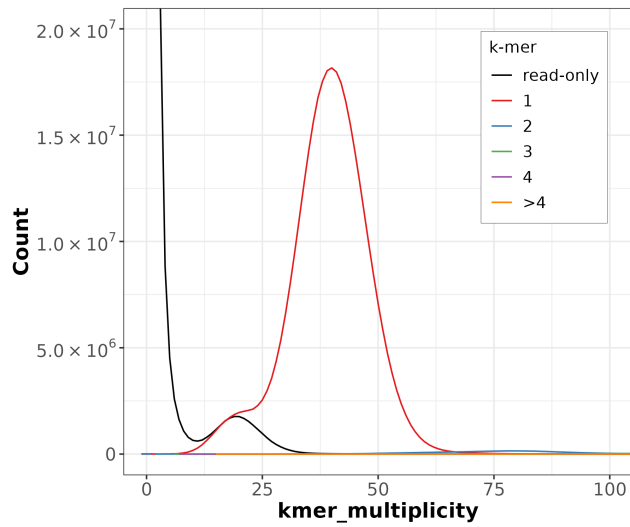
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: lepidoptera_odb12 (genomes:79, BUSCOs:5760)
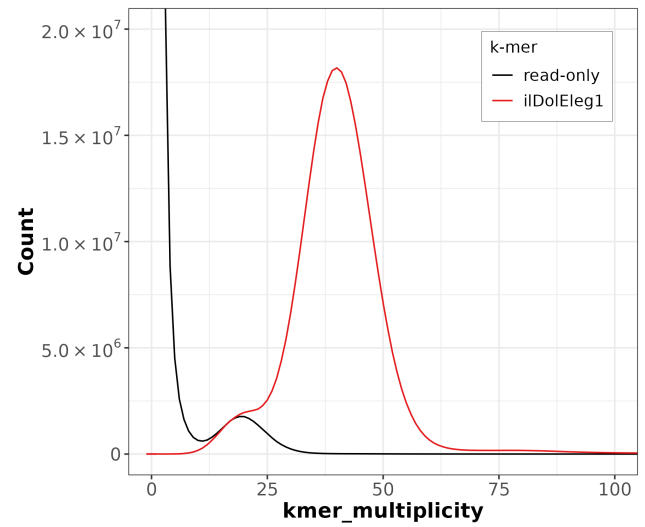
HiC contact map of curated assembly
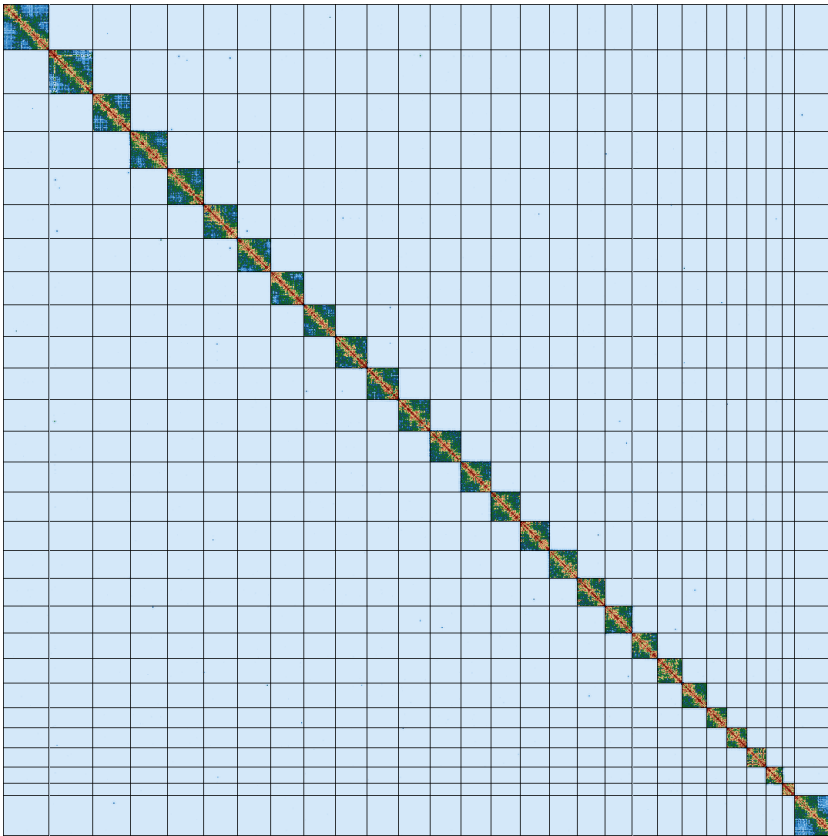
collapsed [LINK]

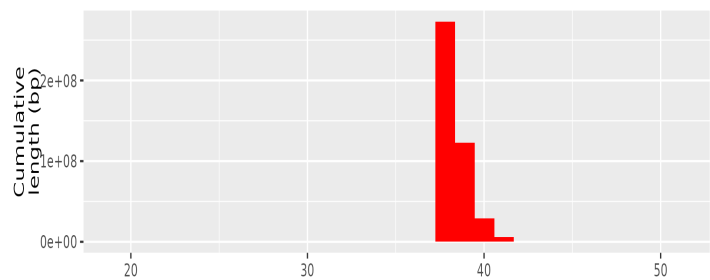# K-mer spectra of curated assembly

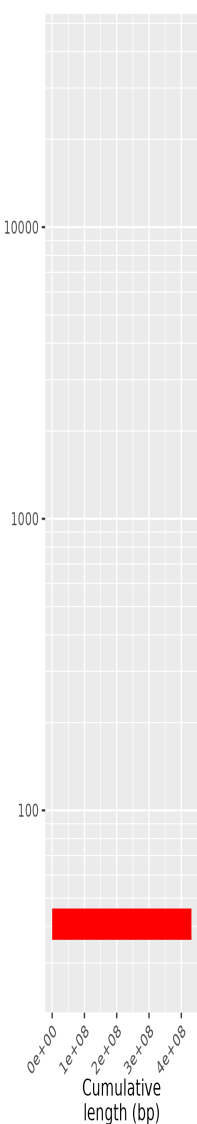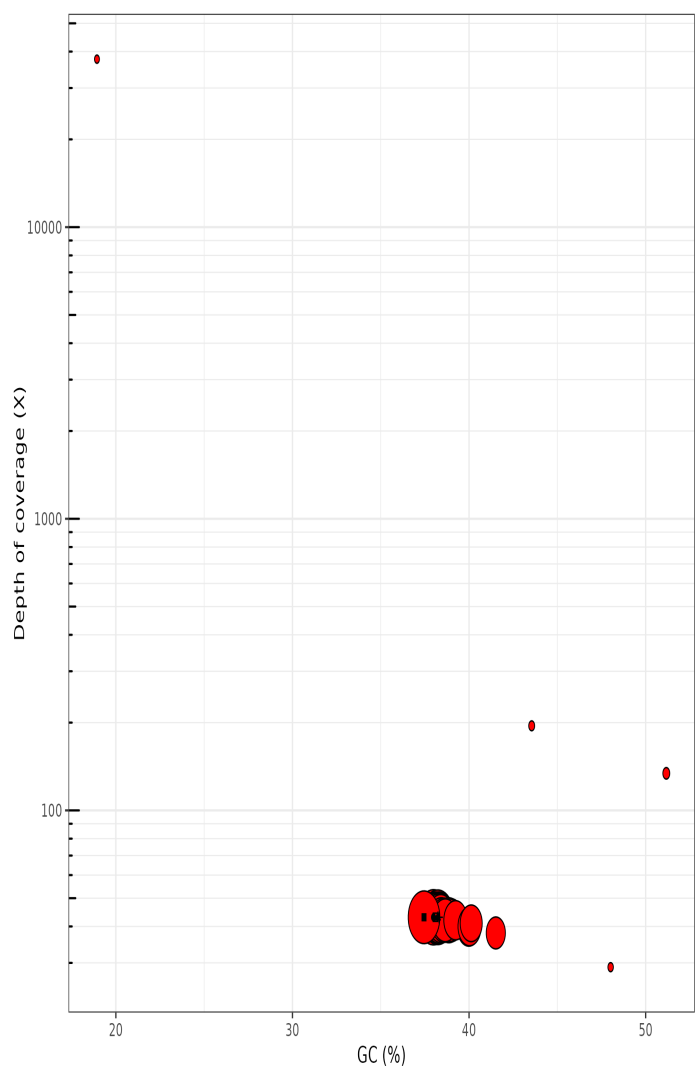Distribution of k-mer counts per copy
numbers found in asm

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

Post-curation contamination screening

TAPAs summary Graph

Longest sequences (bp)

• ilDolEleg1_1 - 23615306 (Eukaryota)

▲ ilDolEleg1_2 - 23032492 (Eukaryota)

■ ilDolEleg1_Z - 21127372 (Eukaryota)

+ ilDolEleg1_3 - 19300041 (Eukaryota)

⊠ ilDolEleg1_4 - 19255577 (Eukaryota)

superkingdom

• Eukaryota

Length (bp)

○ 5.0e+06

○ 1.0e+07

○ 1.5e+07

○ 2.0e+07

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	42	246

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: 2025-10-10 02:47:05 CEST