

# ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	753226
ToLID	<b>i1ScaBole1</b>
Species	Scardia boletella
Class	Insecta
Order	Lepidoptera

Genome Traits	Expected	Observed
Haploid size (bp)	187,880,448	404,657,757
Haploid Number	29 (source: ancestor)	31
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

## EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q58

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

## Curator notes

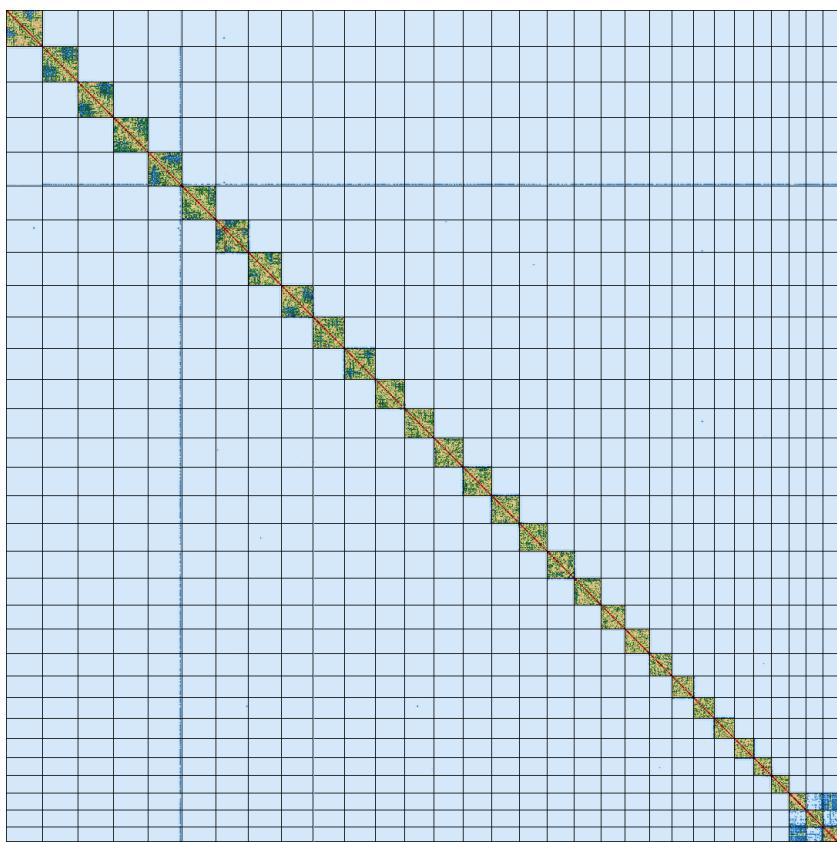
- . Interventions/Gb: 35
- . Contamination notes: ""
- . Other observations: "The assembly of SCARDIA BOLETELLA (i1ScaBole1) is based on 166X PacBio data and 501X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, <https://www.erga-biodiversity.eu/>) via the Biodiversity 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. No contamination was detected. But, 192 regions totaling 6 Mb (with the largest being 996,406 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. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size and the telomere pattern found is TTAGG. "

## Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	404,714,238	404,657,757
GC %	35.15	35.15
Gaps/Gbp	200.14	217.47
Total gap bp	9,200	11,300
Scaffolds	35	33
Scaffold N50	14,210,946	14,210,946
Scaffold L50	13	13
Scaffold L90	27	27
Contigs	115	121
Contig N50	6,207,600	5,201,002
Contig L50	23	25
Contig L90	68	74
QV	58.766	58.7654
Kmer compl.	97.5792	97.579
BUSCO sing.	97.0%	97.0%
BUSCO dupl.	0.9%	0.9%
BUSCO frag.	0.3%	0.3%
BUSCO miss.	1.7%	1.7%

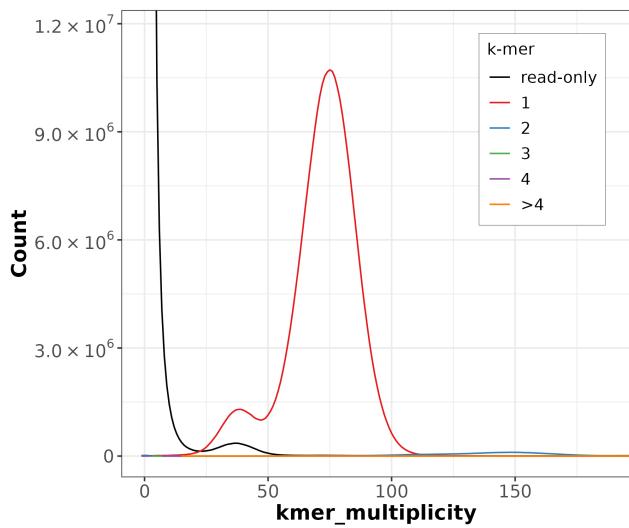
BUSCO: 6.0.0 (euk\_genome\_min, miniprot) / Lineage: lepidoptera\_odb12 (genomes:79, BUSCOs:5760)

# HiC contact map of curated assembly

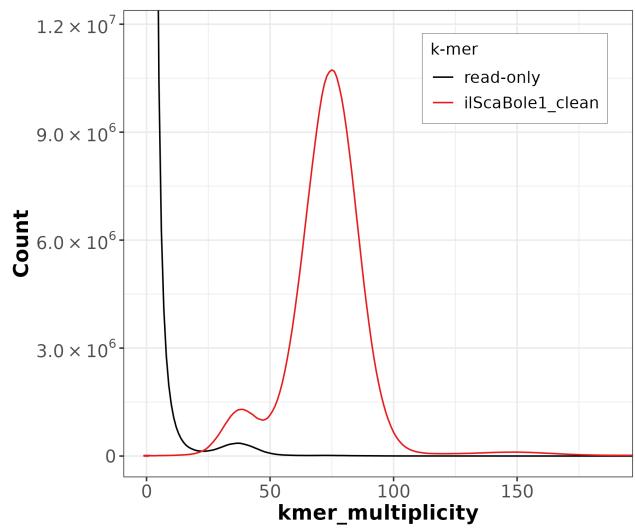


collapsed [\[LINK\]](#)

# K-mer spectra of curated assembly

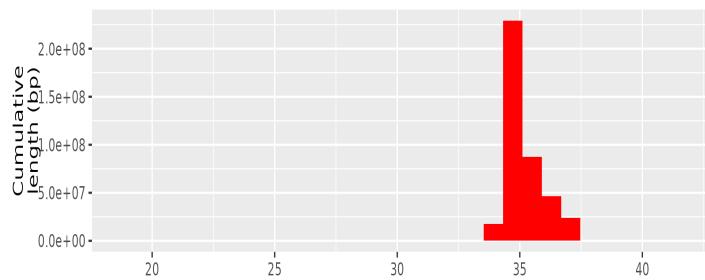


Distribution of k-mer counts per copy numbers found in asm

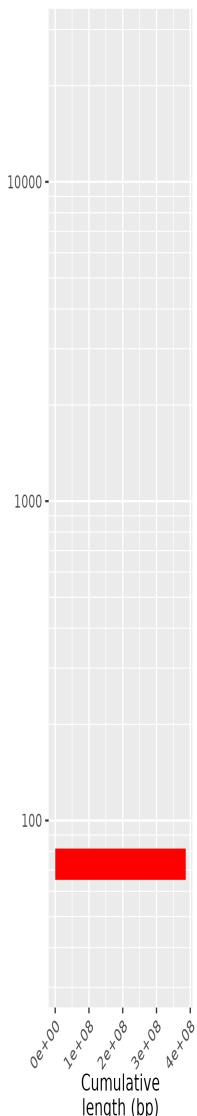
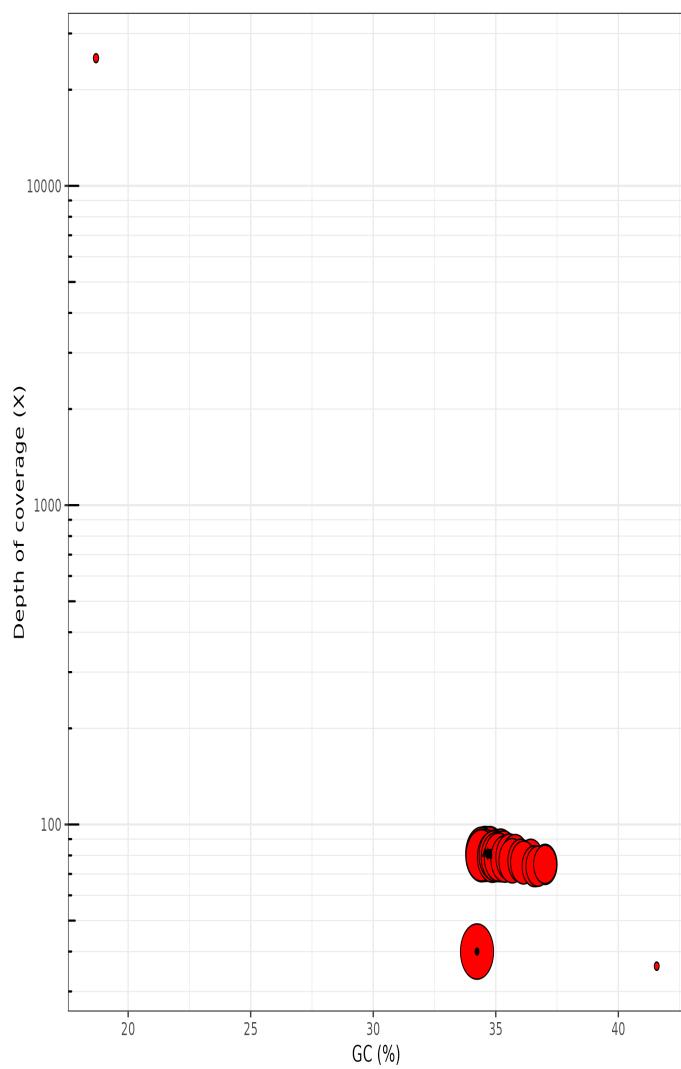


Distribution of k-mer counts coloured by their presence in reads/assemblies

# Post-curation contamination screening



TAPAs summary Graph



Longest sequences (bp)

- llScaBole1\_1 - 17802448 (Eukaryota)
- ▲ llScaBole1\_2 - 17266682 (Eukaryota)
- llScaBole1\_3 - 17138471 (Eukaryota)
- + llScaBole1\_4 - 16963917 (Eukaryota)
- ▣ llScaBole1\_5 - 16640797 (Eukaryota)

superkingdom

- Eukaryota

Length (bp)

- (○) 5.0e+06
- (○) 1.0e+07
- (○) 1.5e+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	166	501

## 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: 2025-11-03 10:58:07 CET