ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	2819898
ToLID	**ihOroJapo4**
Species	Orosanga japonica
Class	Insecta
Order	Hemiptera

Genome Traits	Expected	Observed
Haploid size (bp)	1,516,989,961	1,500,773,387
Haploid Number	13 (source: ancestor)	12
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q65

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

Curator notes

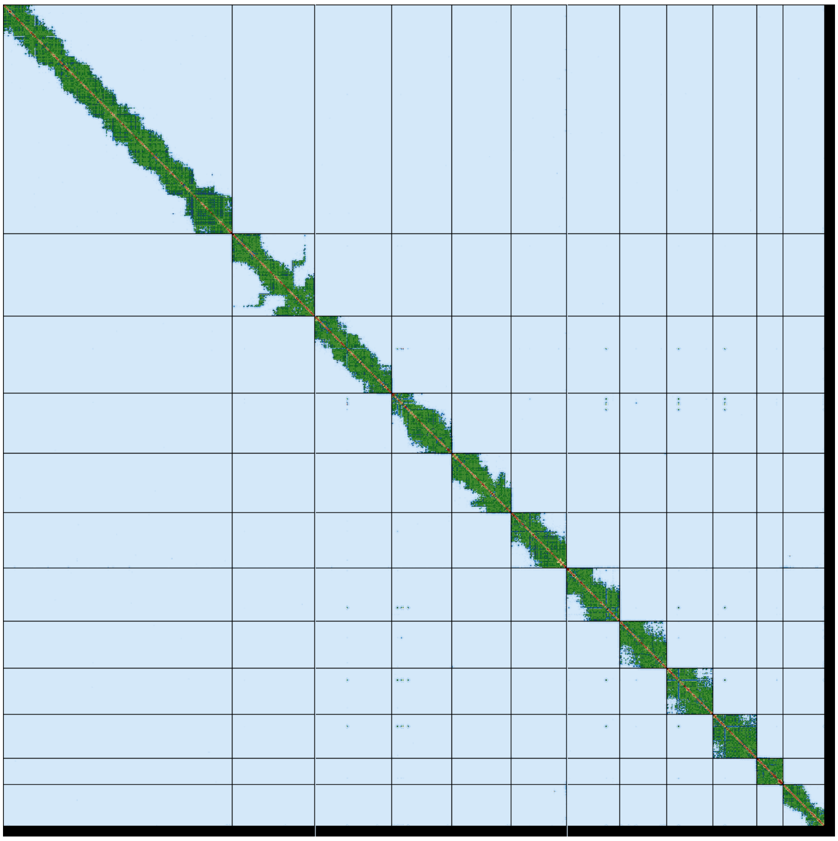
. Interventions/Gb: 36
. Contamination notes: ""
. Other observations: "The assembly of Orosanga japonica (ihOroJapo4) is based on 23,75X PacBio data and 63,98X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics 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, 22 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 4.26 Mb (with the largest being 1.482 Mb). Additionally, 587 regions totaling 58.428 Mb (withthe largest being 2.709 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using OATK. During manual curation, 1 haplotypic regionswere removed, totaling 0.172694Mb (with the largest being 0.172694Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. Based on the long read coverage and HiC signal, we identified the X chromosome and tag it accordingly. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	1,501,003,442	1,500,773,387
GC %	31.93	31.93
Gaps/Gbp	203.86	197.9
Total gap bp	36,700	38,200
Scaffolds	232	197
Scaffold N50	132,285,608	108,566,183
Scaffold L50	4	4
Scaffold L90	10	10
Contigs	518	494
Contig N50	11,227,296	11,191,000
Contig L50	35	36
Contig L90	137	138
QV	65.6269	65.6263
Kmer compl.	88.9303	88.92
BUSCO sing.	90.6%	97.0%
BUSCO dupl.	0.9%	1.0%
BUSCO frag.	6.2%	0.7%
BUSCO miss.	2.2%	1.3%

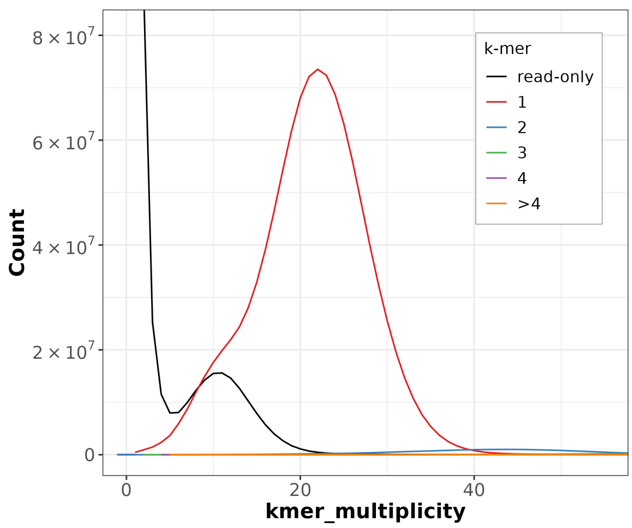
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: hemiptera_odb12 (genomes:32, BUSCOs:3396)
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: hemiptera_odb12 (genomes:32, BUSCOs:3396)

HiC contact map of curated assembly

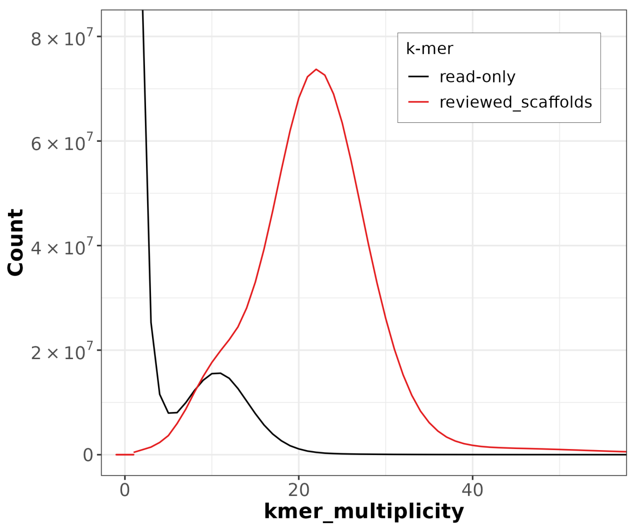


collapsed [LINK]

K-mer spectra of curated assembly

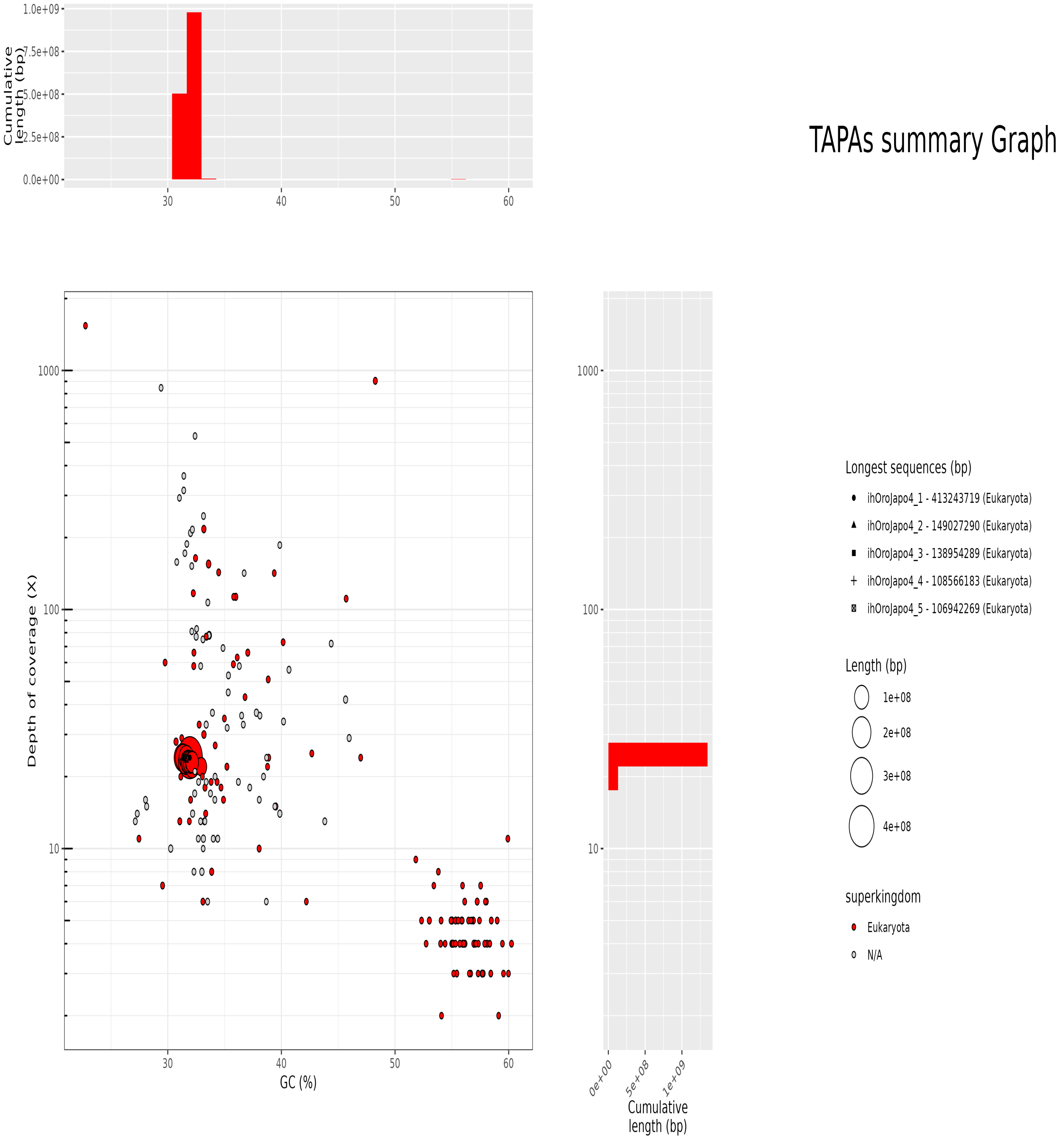


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	23	63

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: Phuong Doan
Affiliation: Genoscope

Date and time: 2025-10-25 14:15:46 CEST