ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1767299
ToLID	**iqAraVest5**
Species	Arachnocephalus vestitus
Class	Insecta
Order	Orthoptera

Genome Traits	Expected	Observed
Haploid size (bp)	3,079,121,200	3,087,915,780
Haploid Number	10 (source: ancestor)	11
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.8.Q59

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

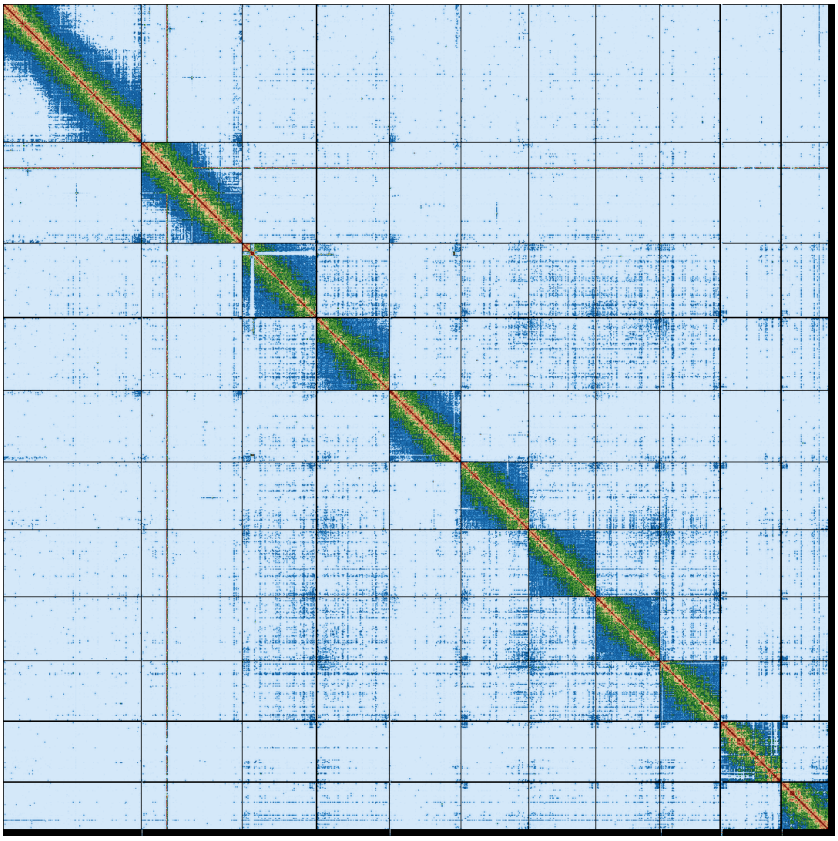
. Interventions/Gb: 79
. Contamination notes: ""
. Other observations: "The assembly of Arachnocephalus vestitus (iqAraVest5) is based on 22X PacBio data and 104X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics 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, 853 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 124 Mb (with the largest being 5.1Mb). Additionally, 2173 regions totaling 194 Mb (with the largest being 1.3 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, 6 haplotypic regions were removed, totaling 4.6 Mb (with the largest being 2.2 Mb). 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	3,092,520,364	3,087,915,780
GC %	43.82	43.82
Gaps/Gbp	695.55	707.27
Total gap bp	261,000	280,700
Scaffolds	482	335
Scaffold N50	265,323,944	265,875,600
Scaffold L50	5	5
Scaffold L90	11	10
Contigs	2,542	2,519
Contig N50	3,133,000	3,128,804
Contig L50	278	279
Contig L90	1,064	1,068
QV	59.2826	59.2822
Kmer compl.	81.656	81.6131
BUSCO sing.	93.0%	93.1%
BUSCO dupl.	4.8%	4.8%
BUSCO frag.	0.6%	0.5%
BUSCO miss.	1.6%	1.6%

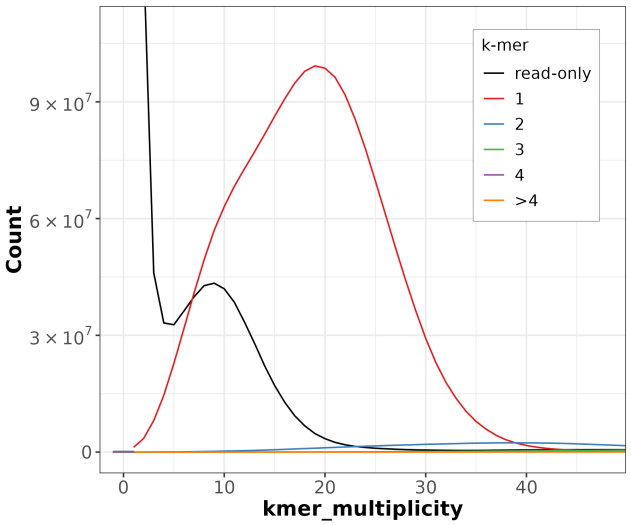
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: insecta_odb12 (genomes:79, BUSCOs:3114)

HiC contact map of curated assembly

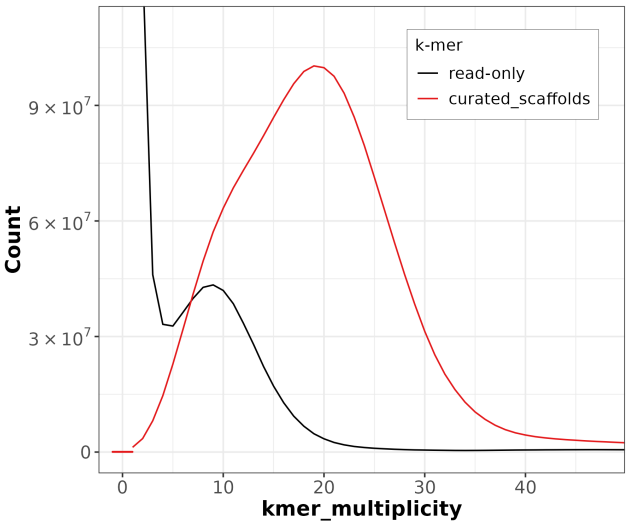


collapsed [LINK]

K-mer spectra of curated assembly

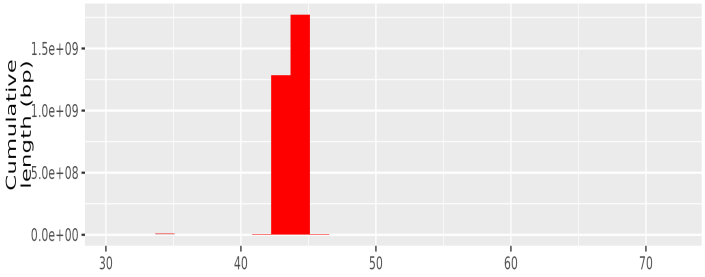


Distribution of k-mer counts per copy numbers found in asm

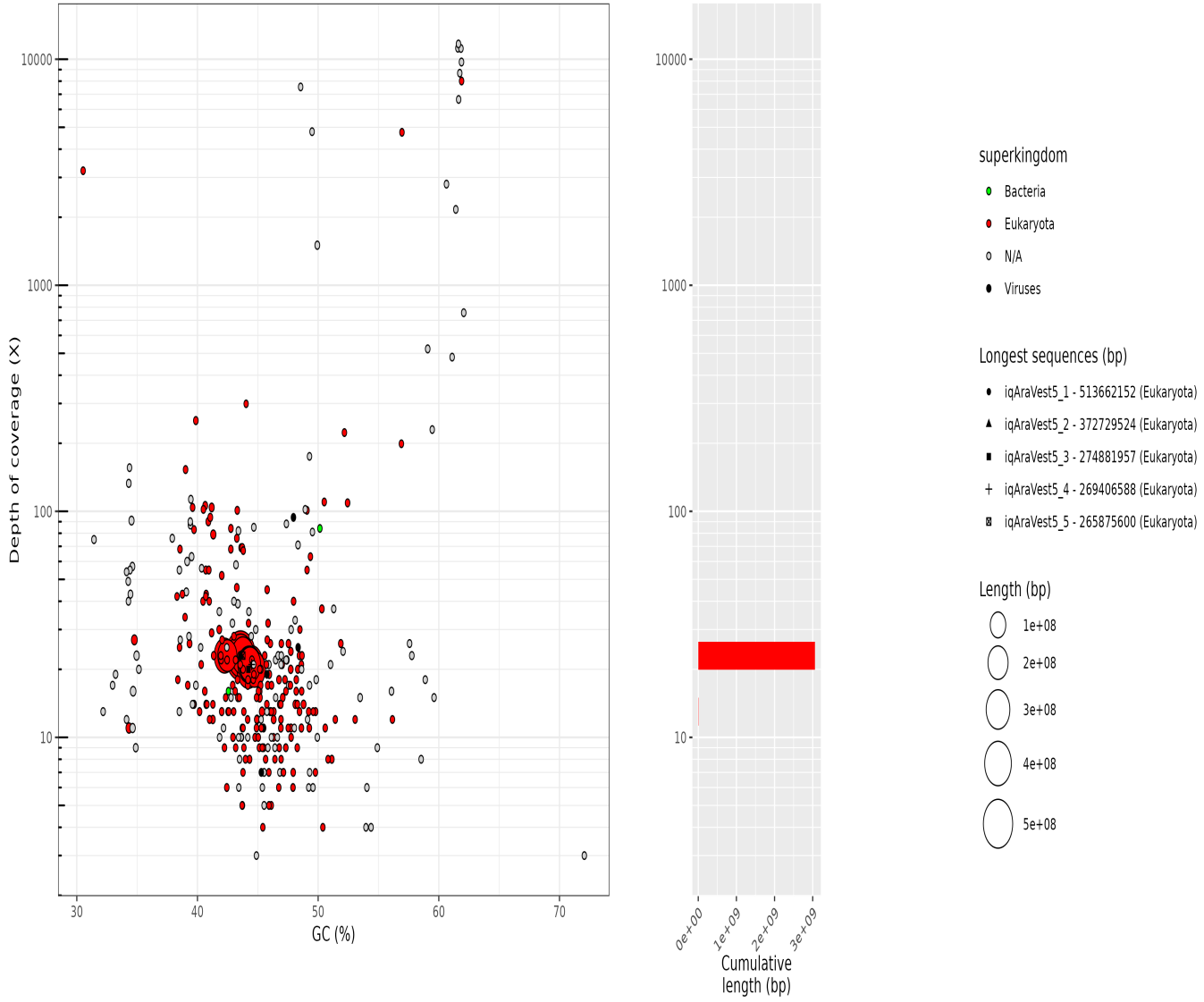


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	22	104

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: 2025-11-22 20:48:02 CET