ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1146903
ToLID	**qcScoCret2**
Species	Scolopendra cretica
Class	Chilopoda
Order	Scolopendromorpha

Genome Traits	Expected	Observed
Haploid size (bp)	1,610,302,235	1,456,208,595
Haploid Number	15 (source: ancestor)	15
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q51

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

. Kmer completeness value is less than 90 for collapsed
. Not 90% of assembly in chromosomes for collapsed

Curator notes

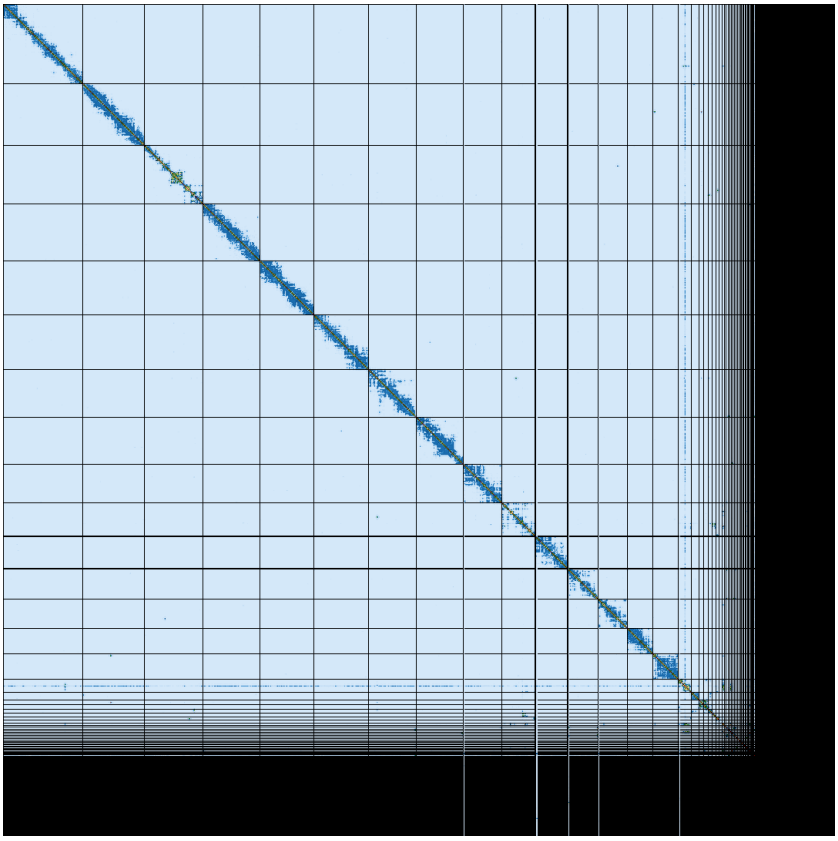
. Interventions/Gb: 70
. Contamination notes: ""
. Other observations: "The assembly of Scolopendra cretica (qcScoCret2.1) is based on 42X PacBio data and 125X of Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics 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 contigs were found to be contaminants (bacterial, archaeal, or viral). Additionally, 720 regions totaling 112.56 Mb (with the largest being 0.98 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, 22 haplotypic regions were removed, totaling 10.73 Mb (with the largest being 2.56 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	1,468,192,980	1,456,208,595
GC %	34.34	34.33
Gaps/Gbp	130.77	135.28
Total gap bp	19,200	23,800
Scaffolds	568	536
Scaffold N50	67,514,300	81,956,192
Scaffold L50	8	8
Scaffold L90	60	46
Contigs	760	733
Contig N50	10,793,000	11,007,542
Contig L50	30	30
Contig L90	189	179
QV	51.0392	51.0366
Kmer compl.	86.5478	86.3077
BUSCO sing.	96.7%	97.1%
BUSCO dupl.	0.9%	0.5%
BUSCO frag.	1.5%	1.5%
BUSCO miss.	0.9%	0.9%

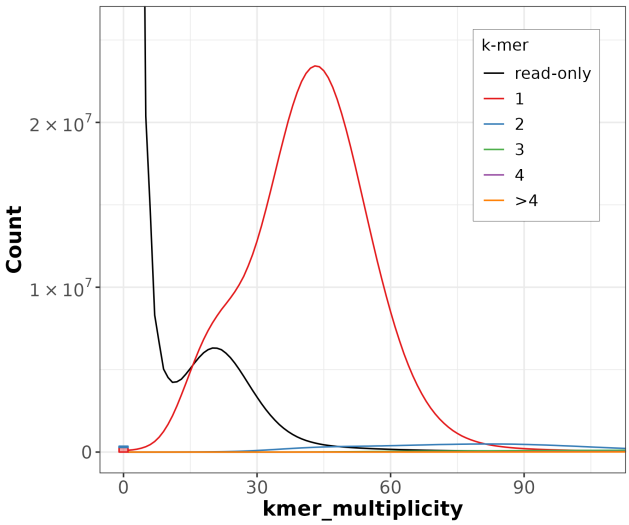
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: arthropoda_odb10 (genomes:90, BUSCOs:1013)

HiC contact map of curated assembly

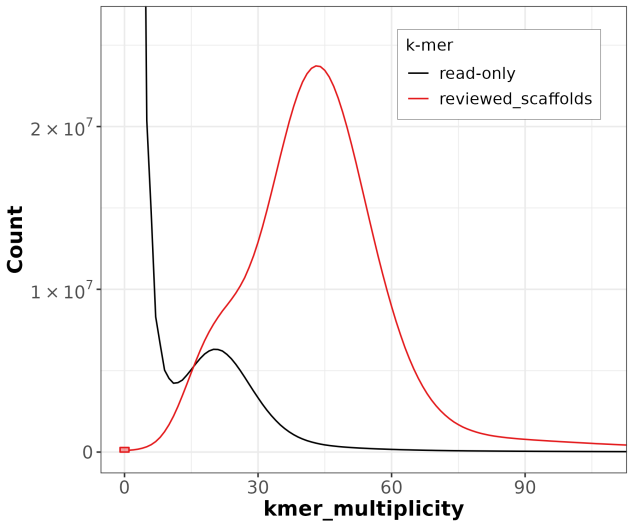


collapsed [LINK]

K-mer spectra of curated assembly

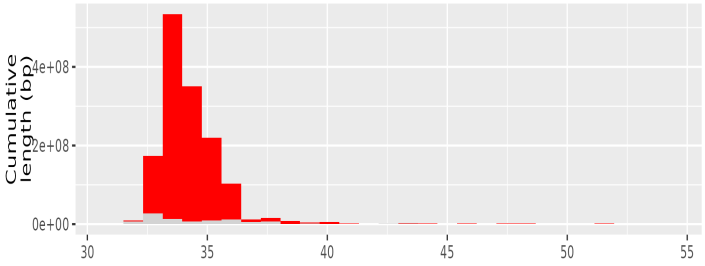


Distribution of k-mer counts per copy numbers found in asm

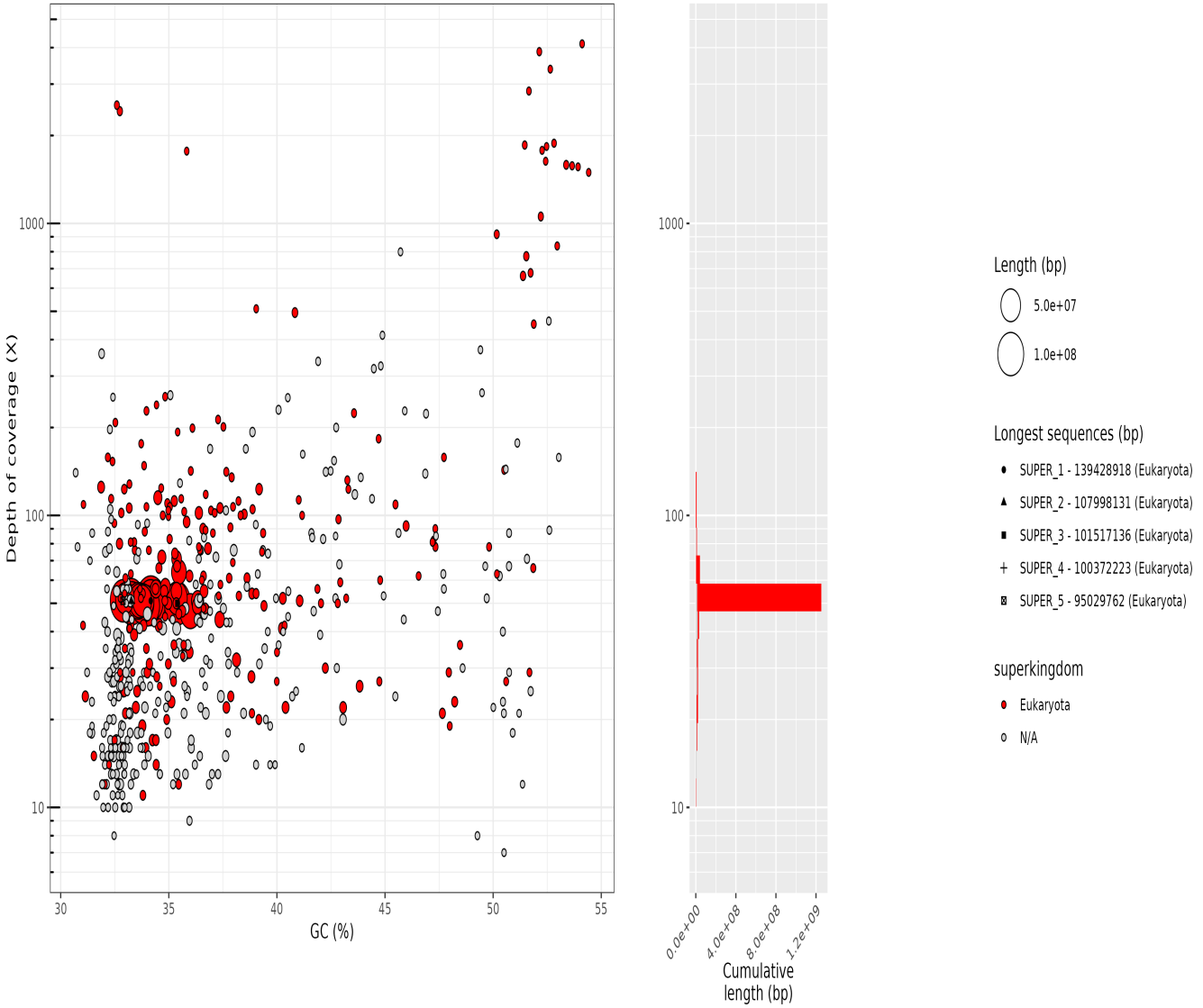


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	PACBIO Hifi	Arima
Coverage	41	122

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: Adama Ndar
Affiliation: Genoscope

Date and time: 2025-08-06 18:45:42 CEST