ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	209733
ToLID	**icCarStae2**
Species	Carabus staehlini
Class	Insecta
Order	Coleoptera

Genome Traits	Expected	Observed
Haploid size (bp)	200,308,448	203,438,952
Haploid Number	14 (source: ancestor)	14
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q61

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

Curator notes

. Interventions/Gb: 335
. Contamination notes: ""
. Other observations: "The assembly of CARABUS STAEHLINI (icCarStae2) is based on 44X PacBio data and 210X 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_dup and Hi-C-based scaffolding with YaHS. In total, 71 contigs were identified as contaminants (bacterial), totaling 3,705,613 pb (with the largest being 1,315,266 pb). Additionally, 215 regions totaling 16,655,681 pb (with the largest being 1,136,177 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 was removed, totaling 1,019,594 pb (with the largest being 597,145 pb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

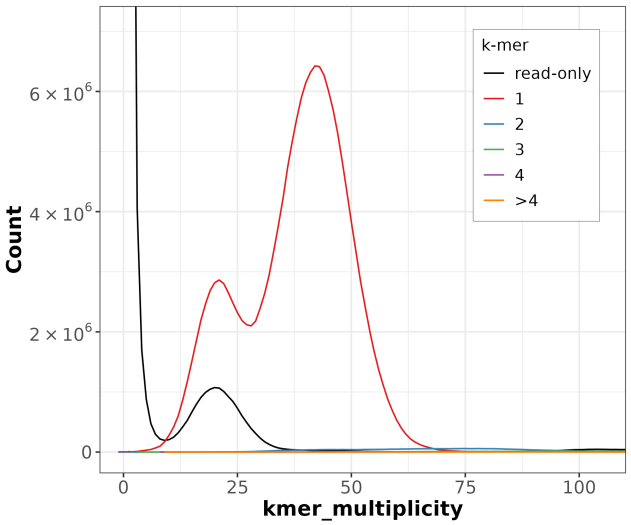
Metrics	Pre-curation collapsed	Curated collapsed
Total bp	205,970,907	203,438,952
GC %	34.58	34.57
Gaps/Gbp	150.51	235.94
Total gap bp	3,200	7,800
Scaffolds	70	47
Scaffold N50	14,277,293	14,277,293
Scaffold L50	6	6
Scaffold L90	13	13
Contigs	99	95
Contig N50	6,325,665	6,325,665
Contig L50	13	13
Contig L90	32	31
QV	61.138	61.197
Kmer compl.	91.1159	90.955
BUSCO sing.	94.9%	94.9%
BUSCO dupl.	0.6%	0.5%
BUSCO frag.	1.4%	1.4%
BUSCO miss.	3.2%	3.2%

BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: coleoptera_odb12 (genomes:64, BUSCOs:3729)

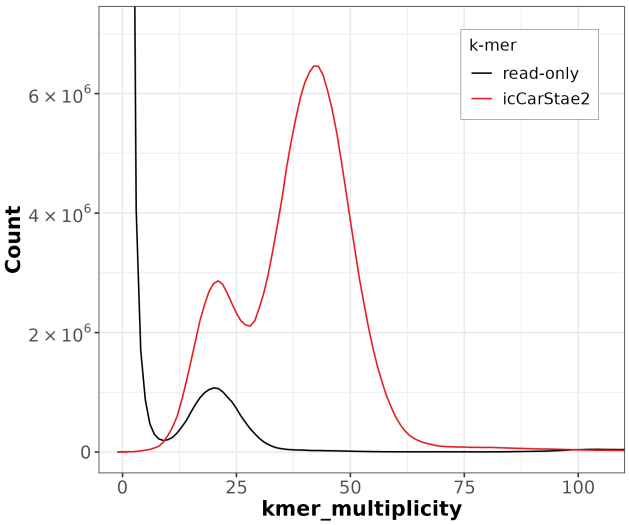
HiC contact map of curated assembly

collapsed [LINK]

K-mer spectra of curated assembly

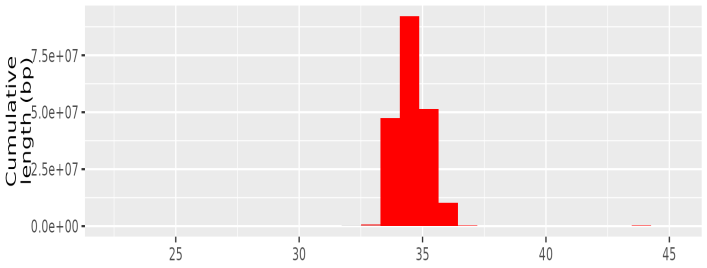


Distribution of k-mer counts per copy numbers found in asm

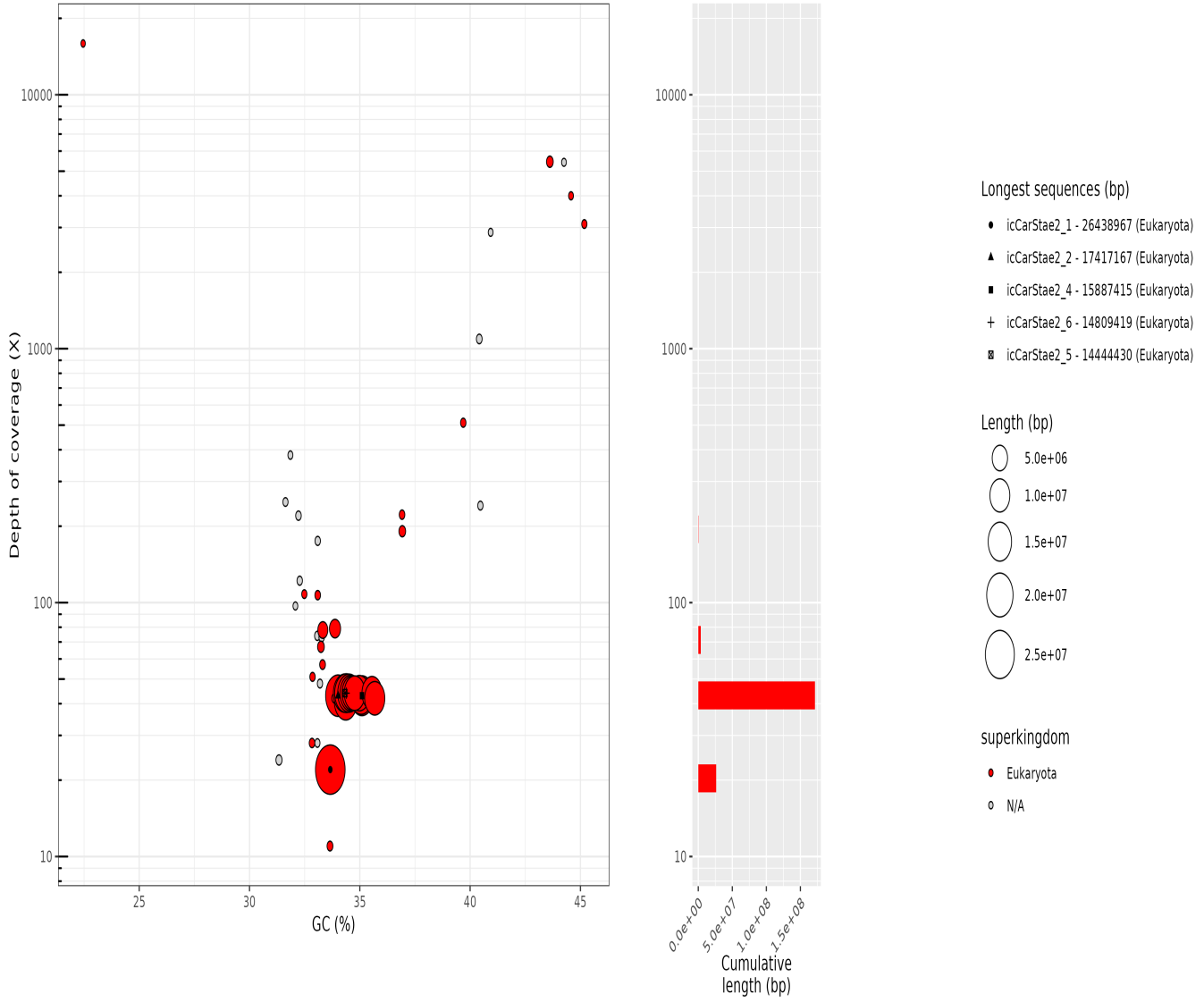


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	44	210

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: 2025-09-05 13:38:55 CEST