ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	3073151
ToLID	**icLeoAngu1**
Species	Leonhardella angulicollis
Class	Insecta
Order	Coleoptera

Genome Traits	Expected	Observed
Haploid size (bp)	485,521,332	493,272,467
Haploid Number	11 (source: ancestor)	11
Ploidy	2 (source: ancestor)	2
Sample Sex	NA	NA

EBP metrics summary and curation notes

Obtained EBP quality metric for pri: 6.7.Q60

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

. Kmer completeness value is less than 90 for pri
. Assembly length loss > 3% for pri

Curator notes

. Interventions/Gb: 364
. Contamination notes: "The PAcBIO HIFI data used in this project was derived from a mix of three individuals for this species. Blobtools allowed us to identify several scaffolds, mostly small, that were made up mostly of bacterial and sometimes fungal sequences. We proceeded to remove 1989 of these scaffolds with no detriment to the busco score (5 Missing BUSCOs pre- and post- elimination of contaminants) even though the contaminated sequences comprised 19% of the total sequence length. The assembly was initially curated prior to filtering out the contaminants. The -scrubbed- assembly was shown to be free of all contaminants. This assembly was subsequently remapped with the HiC data and this is the map we are submitting for review."
. Other observations: "In order to curate the assembly, and given the HiC library was suboptimal, we used both the mq_0 and mq_10 mappings side by side. The save_states in the directory shared for review correspond to the mq_0 mappings. Please refer to save state 3 for review purposes. The shared pretextmap corresponds to a second round of curation, using HiC alignments to the contaminant-free (scrubbed) assembly. It includes a few additional haplotigs we found in the scrubbed assembly. To assemble the genomic sequence of a pooled sample comprising three beetle individuals, we employed a tailored set of parameters in Hifiasm to account for the unique

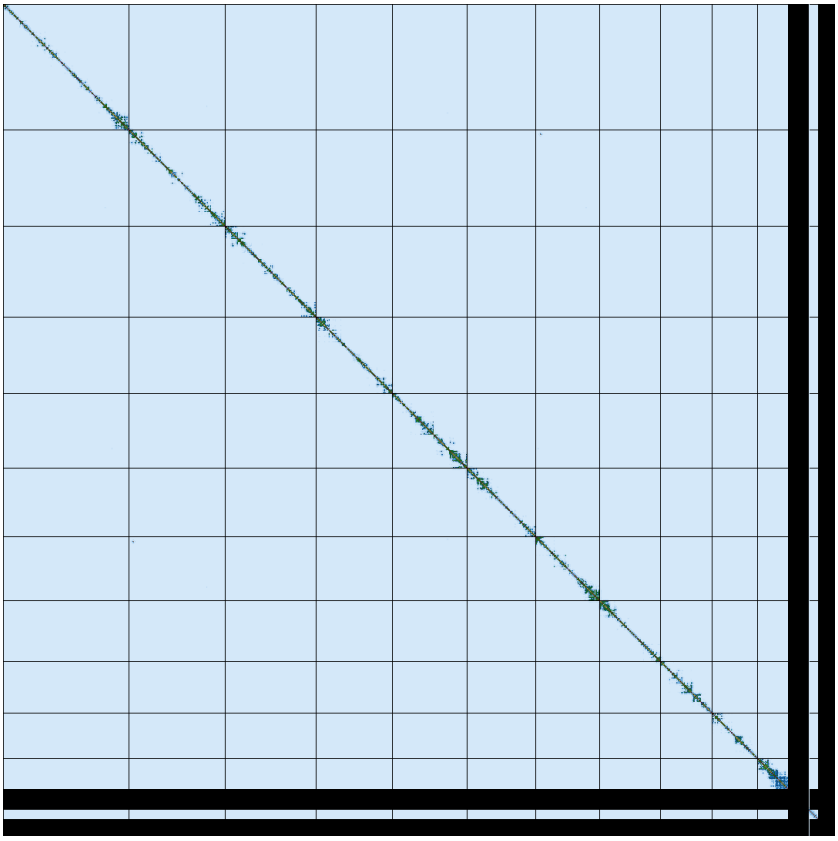
characteristics of the dataset. Given the relatively low heterozygosity observed in the GenomeScope2 plot, we specified --n-hap 3 to reflect the presence of three distinct haplotypes while still accommodating their high similarity. The --hom-cov 44 parameter was manually set to correct for inaccurate peak detection in Hifiasm’s default histogram analysis, ensuring proper separation of homozygous and heterozygous regions. To reduce noise from low-frequency k-mers and improve coverage estimation, we used --min-hist-cnt 10, while --f-perturb 0.1 introduced slight randomness to overlap scoring, helping resolve ambiguities in highly similar regions. Additional parameters such as -N 120 increased the sensitivity of overlap detection, which is crucial when assembling genomes with subtle inter-individual variation. We also included the --primary flag to generate a simplified, non-redundant assembly output, focusing on the most representative contigs. Together, these settings allowed us to balance the need for haplotype resolution with the practical constraints of assembling a genome from a genetically (fairly) homogeneous pool. Based on input from the EAR reviewers, we subsequently determined that most of the SUPER scaffolds were, in fact, chromosome arms. This hypothesis first emerged when we aligned our assembly with those of two closely related species with high-quality genomes—Leptodirus hochenwartii and Catops nigricans. A closer inspection of the mq_10 Pretext map further supported this interpretation, revealing fairly clear evidence of scaffold joins consistent with chromosomal arm structure. These comparative alignments also enabled us to identify the putative X chromosome within our assembly. Additionally, the reviewers helped us detect several haplotigs and guided the reorganization of multiple SUPER scaffolds to improve structural accuracy. This EAR reflects review #2 of the curation, which involved re-orienting some of the chromosome arms and fixing a clear inversion. We have included both the mq_0 and mq_10 pretext maps if needed"

Quality metrics table

Metrics	Pre-curation pri	Curated pri
Total bp	623,527,649	493,272,467
GC %	33.77	32.69
Gaps/Gbp	979.91	488.57
Total gap bp	122,200	48,200
Scaffolds	2,362	342
Scaffold N50	18,918,339	44,176,098
Scaffold L50	13	5
Scaffold L90	555	10
Contigs	2,973	583
Contig N50	2,790,000	3,458,477
Contig L50	67	48
Contig L90	935	160
QV	51.6816	60.2158
Kmer compl.	91.4085	89.9434
BUSCO sing.	96.2%	96.9%
BUSCO dupl.	3.2%	1.7%
BUSCO frag.	0.2%	0.2%
BUSCO miss.	0.4%	1.2%

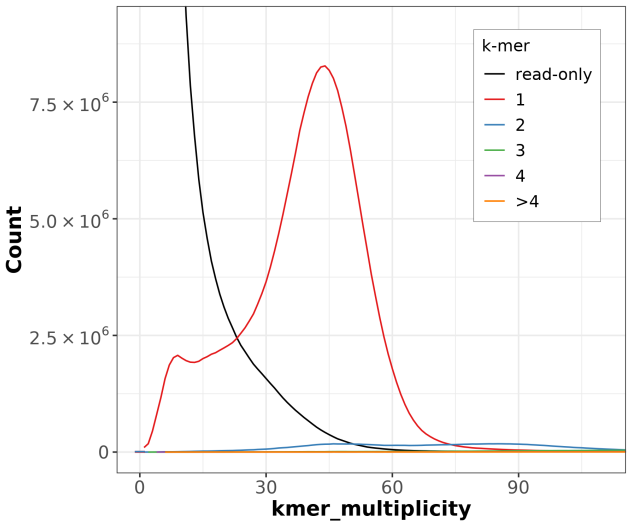
BUSCO: 5.5.0 (euk_genome_met, metaeuk) / Lineage: arthropoda_odb10 (genomes:90, BUSCOs:1013)

HiC contact map of curated assembly

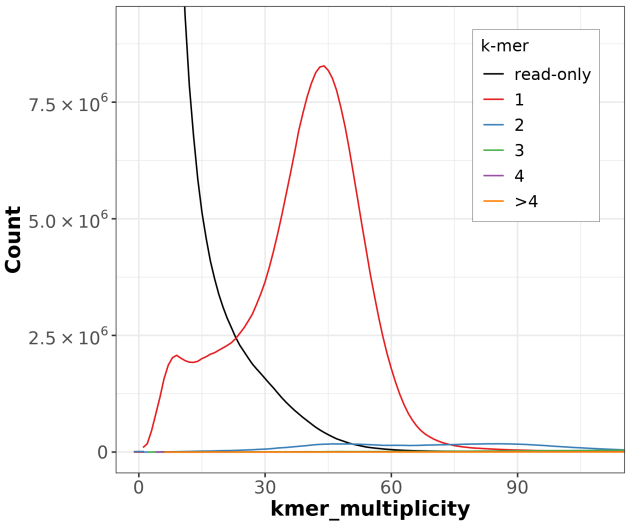


pri [LINK]

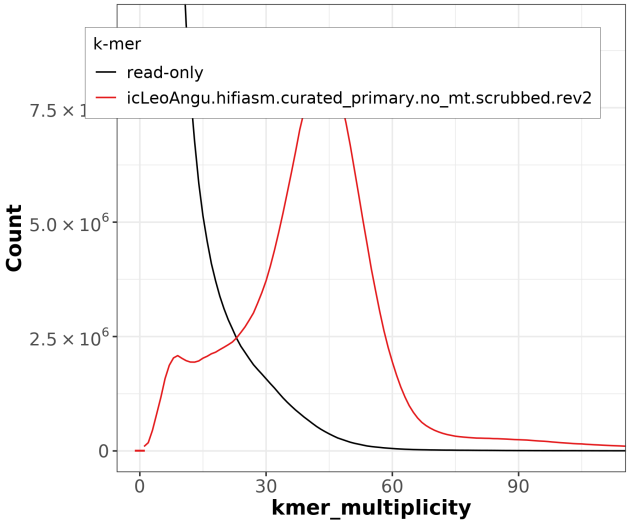
K-mer spectra of curated assembly



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

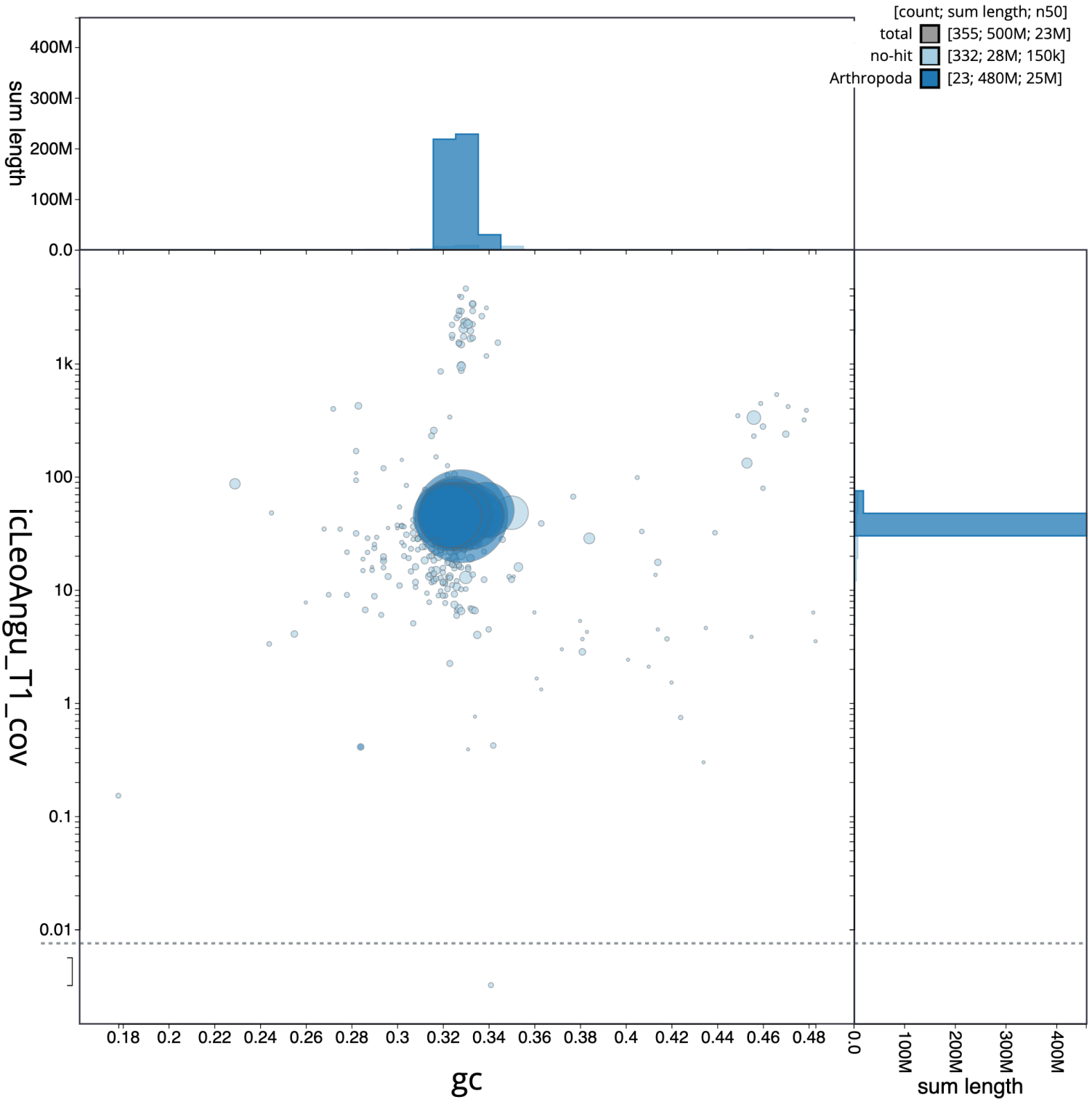


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



pri. 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	OmniC
Coverage	49x	159x

Assembly pipeline

- **hifiasm**
  - |_ *ver:* 0.24.0-r702
  - |_ *key param:* --primary
  - |_ *key param:* --n-hap 3
  - |_ *key param:* --hom-cov 44
  - |_ *key param:* --min-hist-cnt 10
  - |_ *key param:* -N 120
  - |_ *key param:* --f-perturb 0.1
- **purge_dups**
  - |_ *ver:* 1.2.6
  - |_ *key param:* NA
- **YaHS**
  - |_ *ver:* 1.2a
  - |_ *key param:* --no-scaffold-ec
- **CLAWS**
  - |_ *ver:* 2.3
  - |_ *key param:* NA

Curation pipeline

- **PretextViewAI**
  - |_ *ver:* 1.0.3
  - |_ *key param:* NA
- **GRIT_Rapid**
  - |_ *ver:* 2.0
  - |_ *key param:* NA

Submitter: Francisco Camara
Affiliation: CNAG Barcelona

Date and time: 2025-09-26 12:18:26 CEST