

ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1849259
ToLID	icGraMetol
Species	Graciliella metohijensis
Class	Insecta
Order	Coleoptera

Genome Traits	Expected	Observed
Haploid size (bp)	234,579,653	243,216,349
Haploid Number	11 (source: ancestor)	12
Ploidy	2 (source: ancestor)	2
Sample Sex	Mix	Mix

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:

- . Observed Haploid Number is different from Expected
- . Assembly length loss > 3% for pri
- . Not 90% of assembly in chromosomes for pri

Curator notes

- . Interventions/Gb: 176
- . Contamination notes: "The PacBIO HIFI data used in this project was derived from two pooled individuals of this species. Blobtools allowed us to identify several scaffolds, mostly small, that were made up largely of bacterial and fungal sequences. We proceeded to remove 1790 of these scaffolds. In fact 36% of the sequence corresponded to bacterial & fungal contamination. Removing the contaminated scaffolds from the assembly not only did not increase the number missing busco genes but also actually improved not only the busco accuracy but all other assembly quality metrics as well (N50, QV...). The assembly was curated initially prior to filtering out the contaminants. The resulting -scrubbed- assembly was re-analyzed with Blobtools and we were able to find 7 additional short scaffolds made up mostly of bacterial sequence. These will be removed from the assembly submitted for review"
- . Other observations: "The mix of individuals we assembled appears to consist of one male and one female specimen, as suggested by the observation that one chromosome exhibits approximately three-quarters HiFi coverage. To assemble the genome of a beetle derived from two pooled individuals, we applied a customized set of Hifiasm parameters aimed at maximizing contiguity while accommodating subtle haplotype

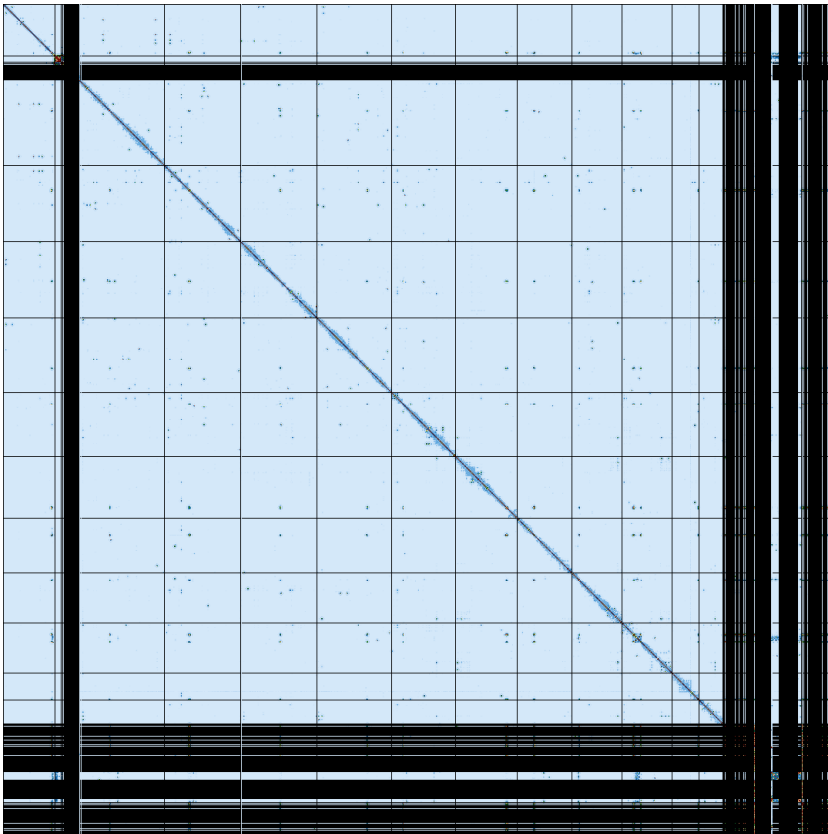
variation. Although the sample contained only two individuals, we specified `--n-hap 4` to allow for the resolution of sub-haplotypes and structural differences that might exist within or between individuals. The manually set `--hom-cov 51` corrected for inaccurate peak detection in coverage histograms, ensuring proper classification of homozygous regions. To reduce noise and improve overlap scoring in this low-diversity context, we used `--min-hist-cnt 5` and increased overlap sensitivity with `-N 150`. To further optimize for a more contiguous primary assembly, we introduced controlled randomness in overlap scoring using `--n-perturb 30000` and `--f-perturb 0.3`, which helped resolve ambiguities in highly similar regions and prevented premature fragmentation. These perturbation settings allowed Hifiasm to better distinguish between nearly identical sequences, improving graph resolution and contig extension. Only the primary assembly achieved a contig N50 greater than 1 Mb, making it the most suitable for downstream analysis. Therefore our focus shifted toward curating this mixed primary assembly to produce a high-quality, representative genomic sequence. This is v2 of the EAR after external reviewing. Several sequence re-arrangements were made and 17 haplotigs were found and removed from the final assembly. The rev2 version of this EAR contains only a couple of minor changes from the previous version to address one of the reviewers concerns"

Quality metrics table

Metrics	Pre-curation pri	Curated pri
Total bp	387,076,371	243,216,349
GC %	35.49	32.05
Gaps/Gbp	382.35	226.14
Total gap bp	29,600	11,000
Scaffolds	2,192	389
Scaffold N50	7,176,294	18,044,559
Scaffold L50	11	6
Scaffold L90	922	46
Contigs	2,340	444
Contig N50	2,026,496	5,769,963
Contig L50	34	13
Contig L90	1,029	96
QV	56.8331	60.2916
Kmer compl.	97.6147	93.8925
BUSCO sing.	97.4%	97.9%
BUSCO dupl.	2.1%	1.6%
BUSCO frag.	0.2%	0.2%
BUSCO miss.	0.3%	0.3%

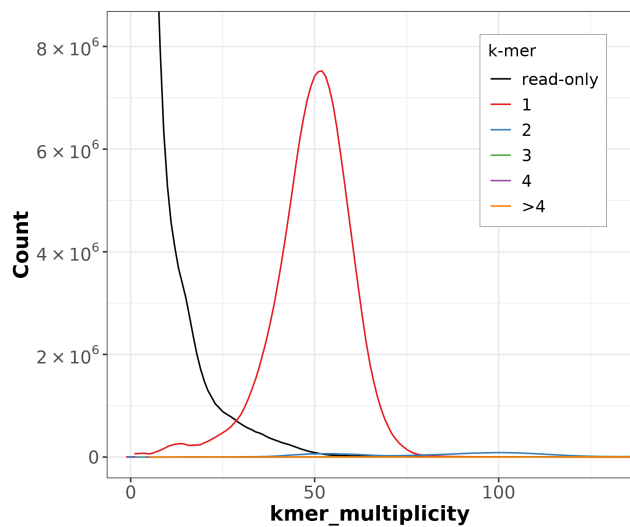
BUSCO: 5.5.0 (euk_genome_met, metaeuk) / Lineage: arthropoda_odb10 (genomes:90, BUSCOs:1013)

HiC contact map of curated assembly

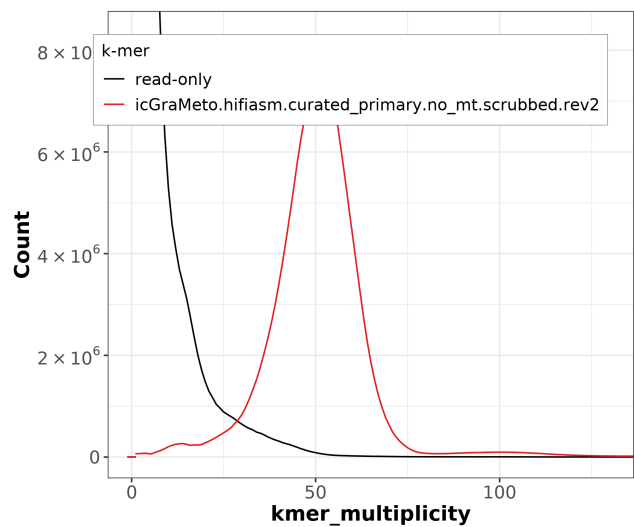


pri [\[LINK\]](#)

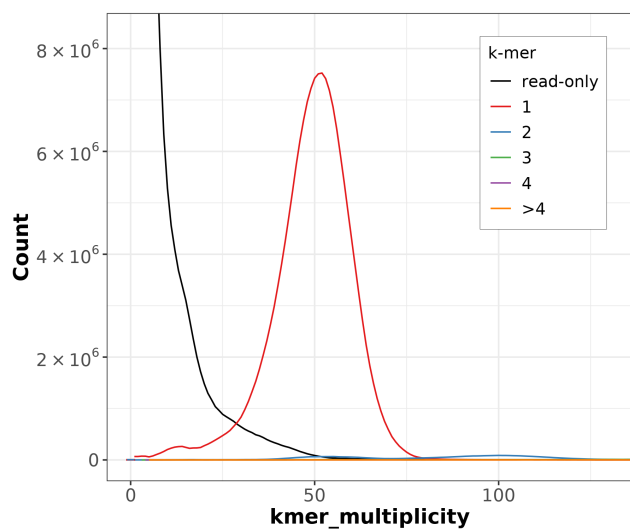
K-mer spectra of curated assembly



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

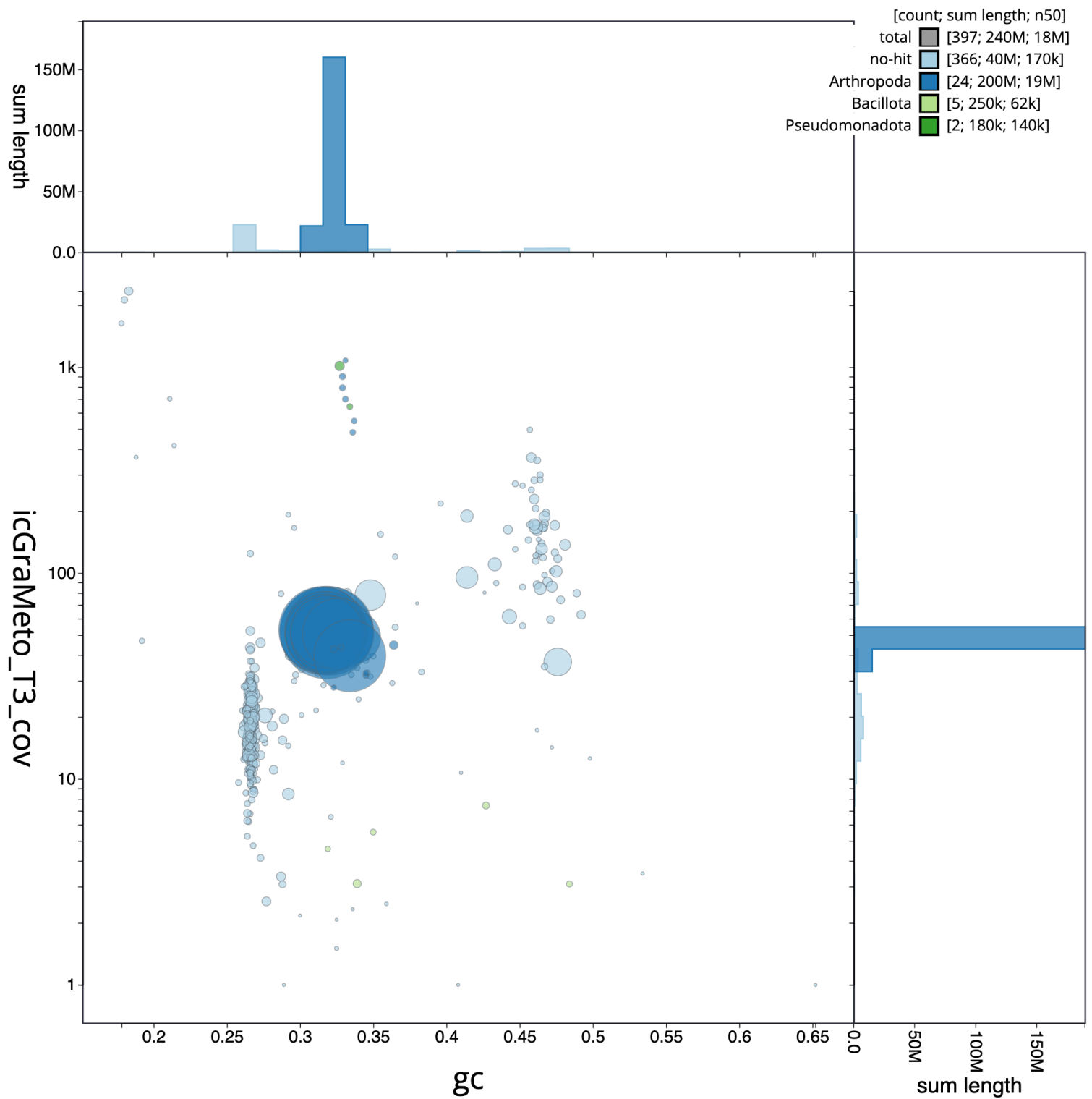


Distribution of k-mer counts coloured by their presence in reads/assemblies



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

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	56x	344x

Assembly pipeline

- **hifiasm**
 - |_ *ver*: 0.24.0-r702
 - |_ *key param*: --n-hap 4
 - |_ *key param*: --hom-cov 51
 - |_ *key param*: --min-hist-cnt 5
 - |_ *key param*: -N 150
 - |_ *key param*: --n-perturb 30000
 - |_ *key param*: --f-perturb 0.3
- **YaHS**
 - |_ *ver*: 1.2a
 - |_ *key param*: --no-scaffold-ec
- **CLAWS**
 - |_ *ver*: 2.3
 - |_ *key param*: NA

Curation pipeline

- **PretextViewAI**
 - |_ *ver*: 1.0.4
 - |_ *key param*: NA
- **GRIT_Rapid**
 - |_ *ver*: 2.0
 - |_ *key param*: NA

Submitter: Francisco Camara

Affiliation: CNAG Barcelona

Date and time: 2025-09-29 15:45:21 CEST