ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	485107
ToLID	**iyForGaga1**
Species	Formica gagates
Class	Insecta
Order	Hymenoptera

Genome Traits	Expected	Observed
Haploid size (bp)	282,970,999	266,142,371
Haploid Number	27 (source: direct)	32
Ploidy	1 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q69

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

. Observed Haploid Number is different from Expected
. Observed Ploidy is different from Expected

Curator notes

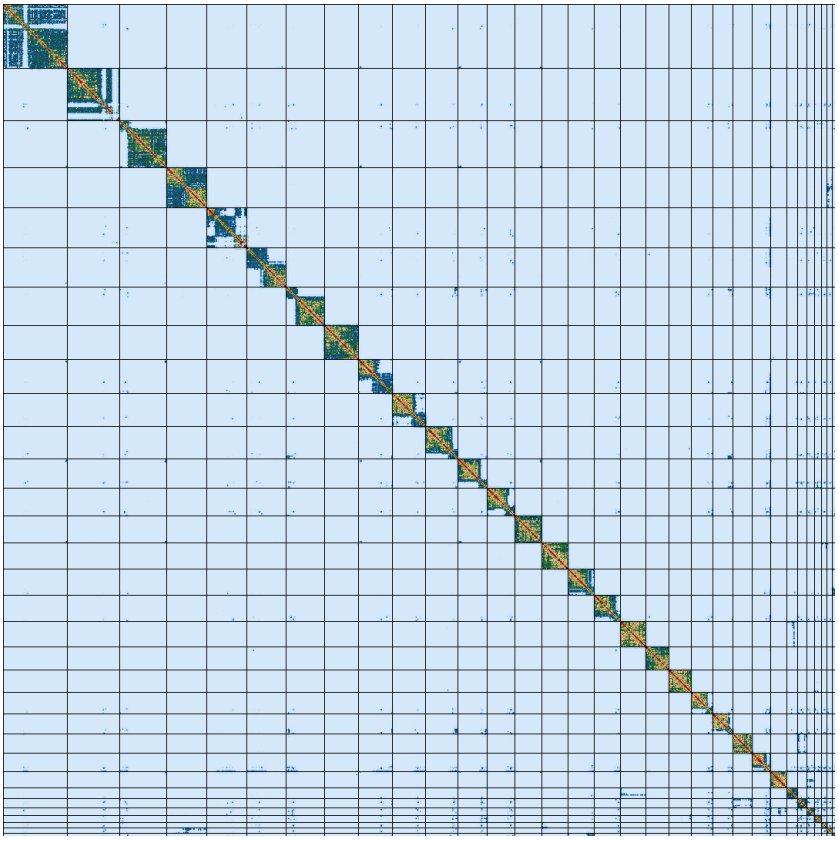
. Interventions/Gb: 42
. Contamination notes: ""
. Other observations: "The assembly of Formica gagates (iyForGaga1) is based on 72X PacBio data and Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe 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, 5 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 2.8 Mb (with the largest being 1.4 Mb). Additionally, 110 regions totaling 24.1 Mb (with the largest being 4.5 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, 2 haplotypic regions were removed,totaling 0.396 Mb (with the largest being 0.214 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	266,537,386	266,142,371
GC %	35.77	35.77
Gaps/Gbp	97.55	120.24
Total gap bp	2,600	3,900
Scaffolds	41	34
Scaffold N50	9,320,565	10,606,726
Scaffold L50	11	10
Scaffold L90	25	23
Contigs	67	66
Contig N50	8,221,297	8,221,297
Contig L50	13	13
Contig L90	30	30
QV	69.9856	69.9791
Kmer compl.	90.6517	90.6473
BUSCO sing.	95.5%	95.6%
BUSCO dupl.	0.4%	0.4%
BUSCO frag.	0.8%	0.8%
BUSCO miss.	3.3%	3.3%

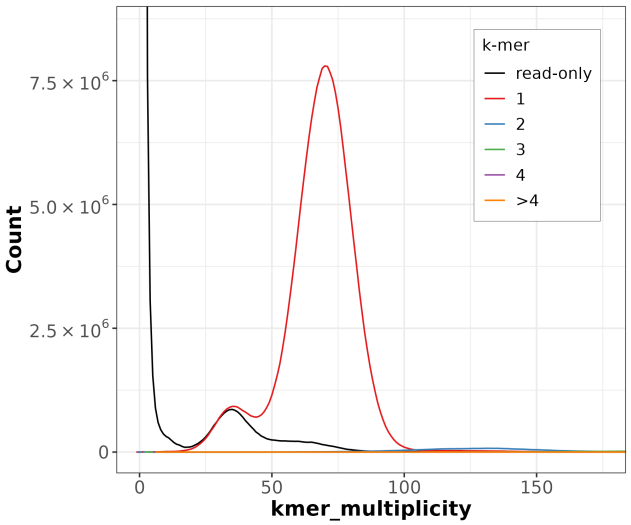
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: formicidae_odb12 (genomes:24, BUSCOs:7266)

HiC contact map of curated assembly

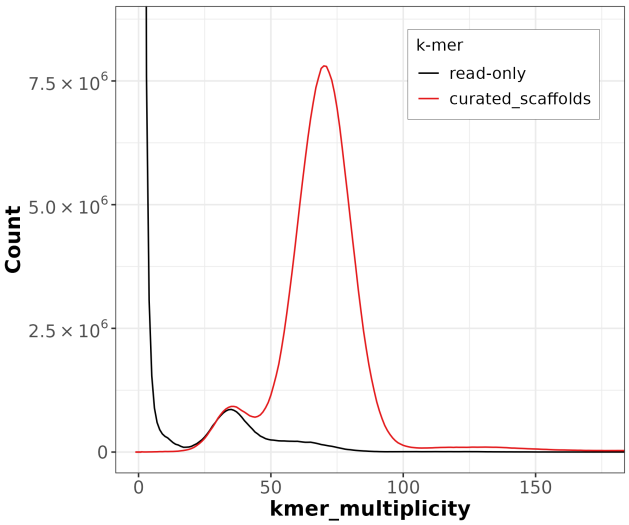


collapsed [LINK]

K-mer spectra of curated assembly

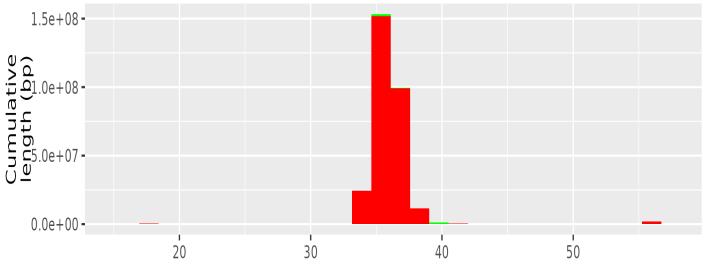


Distribution of k-mer counts per copy numbers found in asm

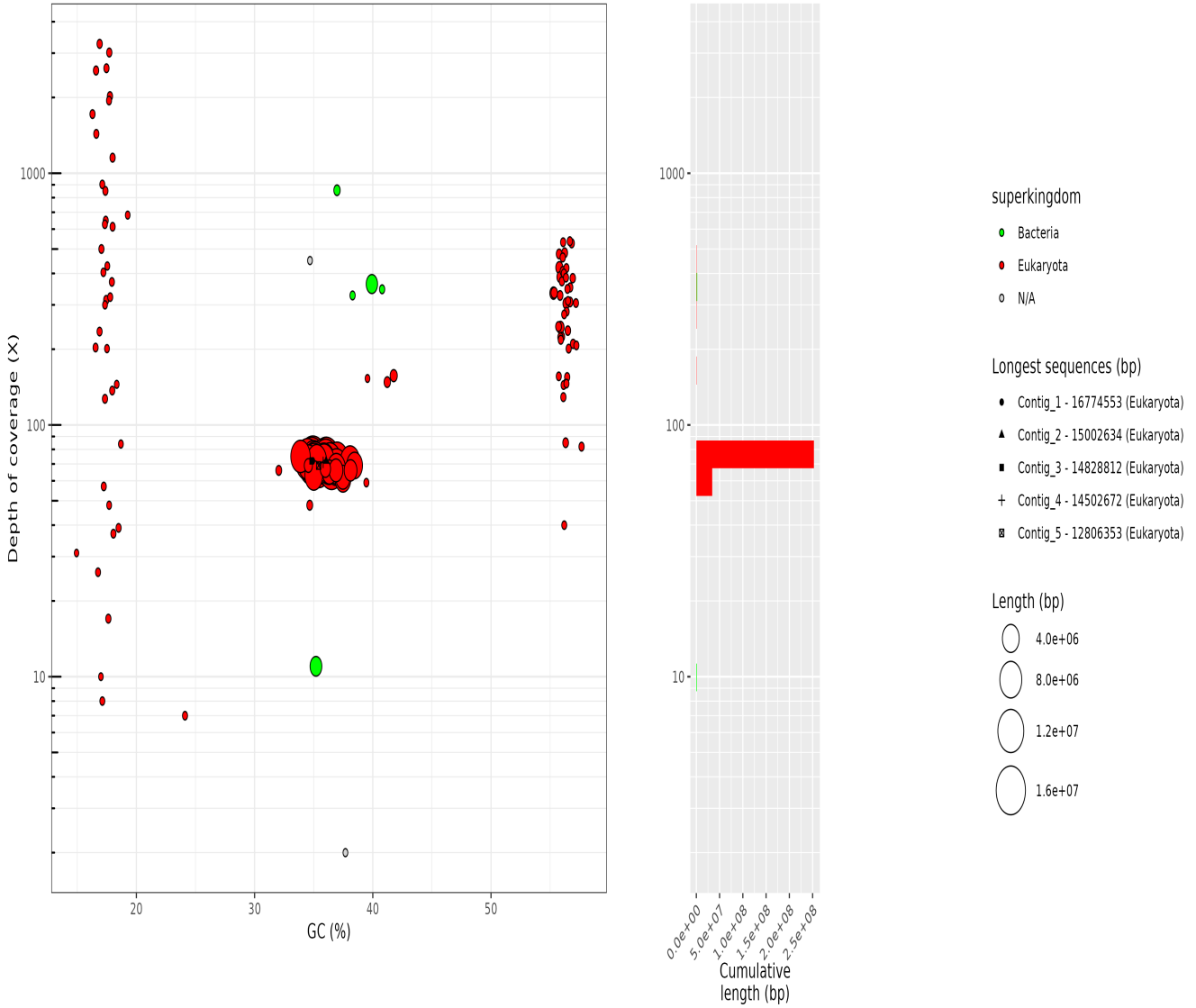


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	72	193

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: 2025-09-04 23:27:38 CEST