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	289,571,368
Haploid Number	27 (source: direct)	25
Ploidy	1 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q66

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

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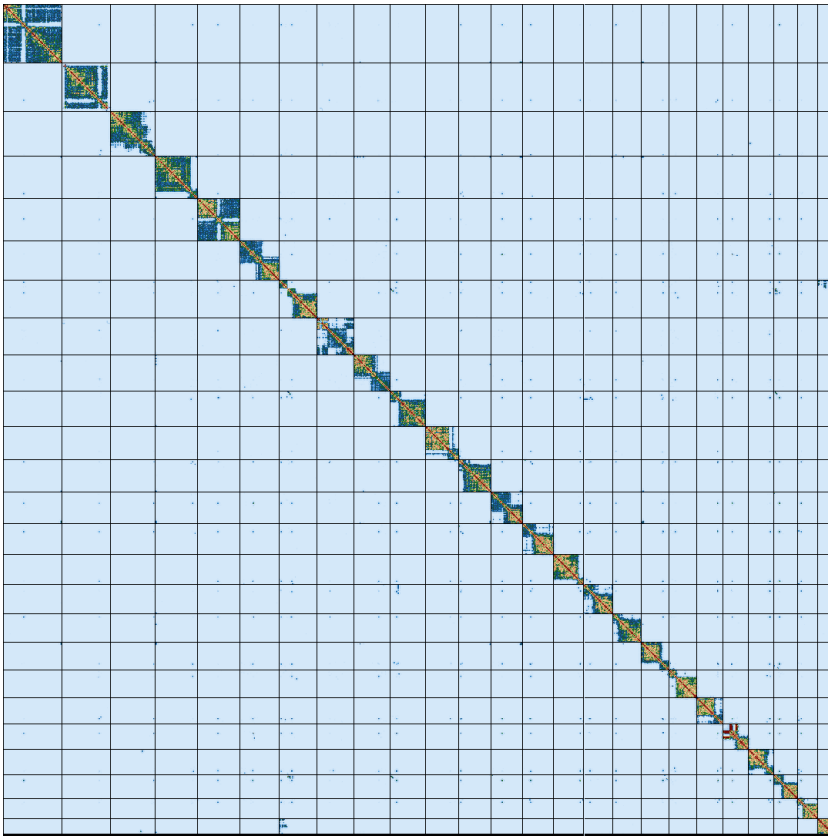. Observed Ploidy is different from Expected

Curator notes

. Interventions/Gb: 460
. 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, 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). The mitochondrial genome was assembled using OATK but the sequence obtained is linear. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 5 haplotypic regions were removed,totaling 0.903 Mb (with the largest being 0.539 Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size "

Quality metrics table

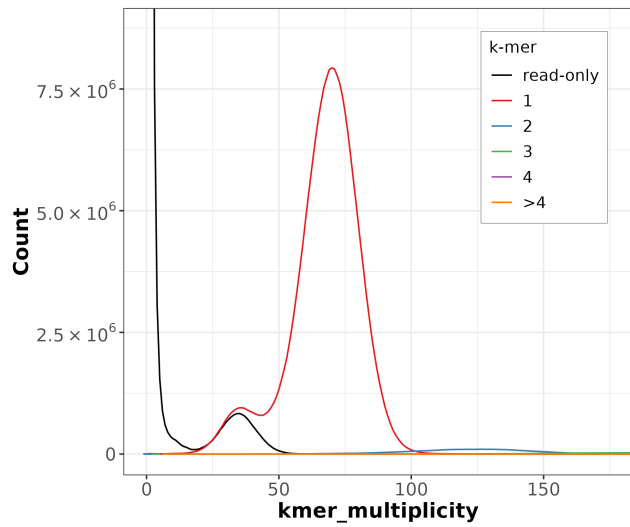| Metrics	Pre-curation collapsed	Curated collapsed
Total bp	266,537,386	289,571,368
GC %	35.77	35.99
Gaps/Gbp	97.55	379.87
Total gap bp	2,600	20,100
Scaffolds	41	56
Scaffold N50	9,320,565	12,250,416
Scaffold L50	11	10
Scaffold L90	25	22
Contigs	67	166
Contig N50	8,221,297	7,685,475
Contig L50	13	15
Contig L90	30	40
QV	69.9856	66.9781
Kmer compl.	90.6517	93.6067
BUSCO sing.	95.5%	96.2%
BUSCO dupl.	0.4%	0.4%
BUSCO frag.	0.8%	0.7%
BUSCO miss.	3.3%	2.7%

BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: formicidae_odb12 (genomes:24, BUSCOs:7266)
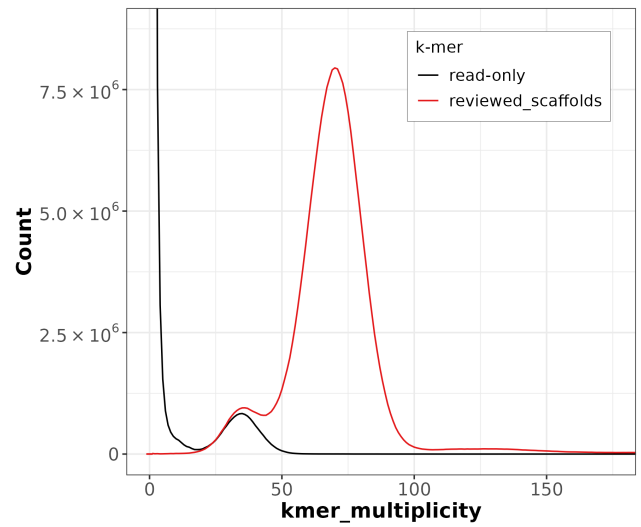
HiC contact map of curated assembly

collapsed [LINK]

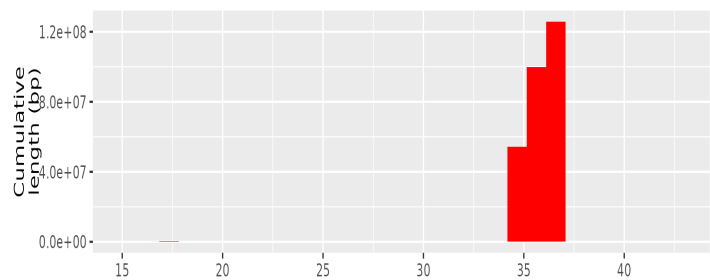# K-mer spectra of curated assembly

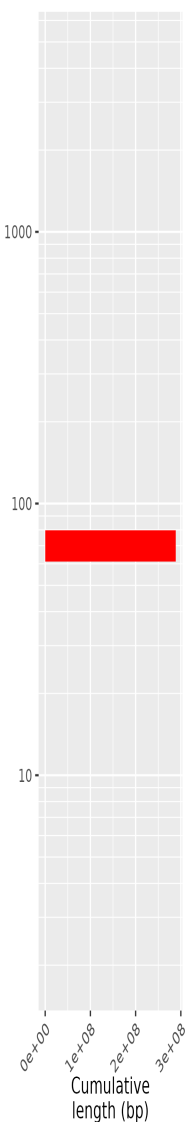
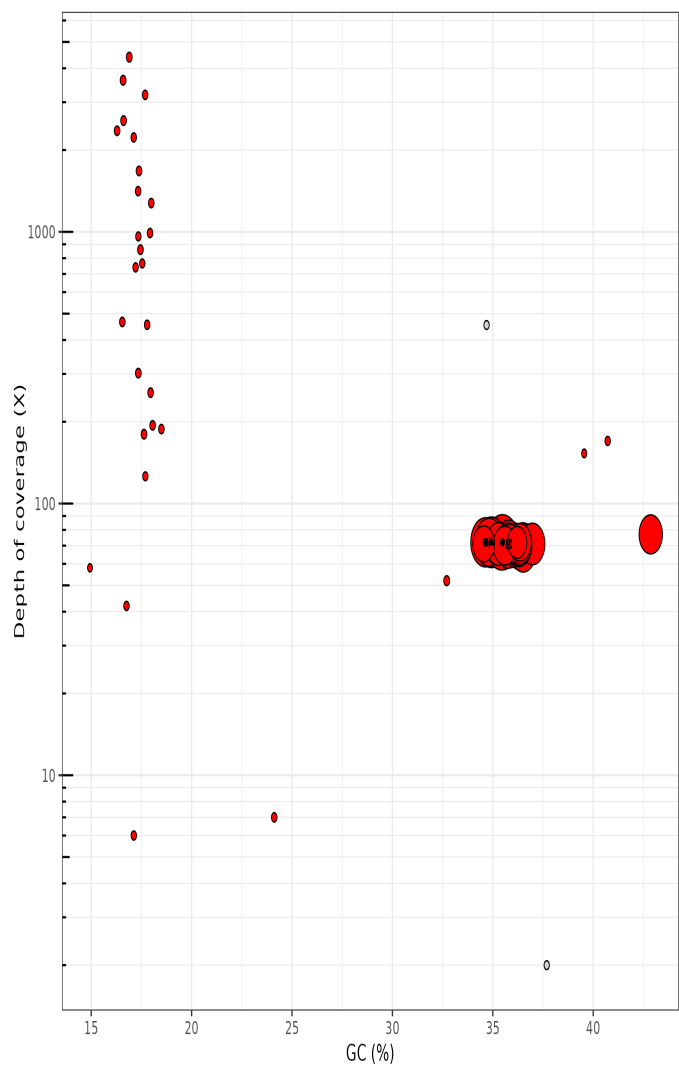Distribution of k-mer counts per copy
numbers found in asm

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

Post-curation contamination screening

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	Long reads	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: 2025-09-28 02:47:13 CEST