ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1980904
ToLID	**xgMetNove1**
Species	Metafruticicola noverca
Class	Gastropoda
Order	Stylommatophora

Genome Traits	Expected	Observed
Haploid size (bp)	1,632,108,443	1,752,039,481
Haploid Number	21 (source: ancestor)	26
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q51

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

. Observed Haploid Number is different from Expected

. Kmer completeness value is less than 90 for collapsed
. BUSCO single copy value is less than 90% for collapsed
. Assembly length loss > 3% for collapsed
. More than 1000 gaps/Gbp for collapsed

Curator notes

. Interventions/Gb: 587
. Contamination notes: ""
. Other observations: "The assembly of Metafruticicola noverca (xgMetNove1.1) is based on 63X PacBio data and 29X 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, 35 contigs were identified as contaminants (bacterial), totaling 1.8 Mb (with the largest being 647 Kb). Additionally, 11,422 regions totaling 2.14 Gb (with the largest being 5.1 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, 659 haplotypic regions were removed, totaling 355.2 Mb (with the largest being 38.3 Mb). Chromosome-scale scaffolds

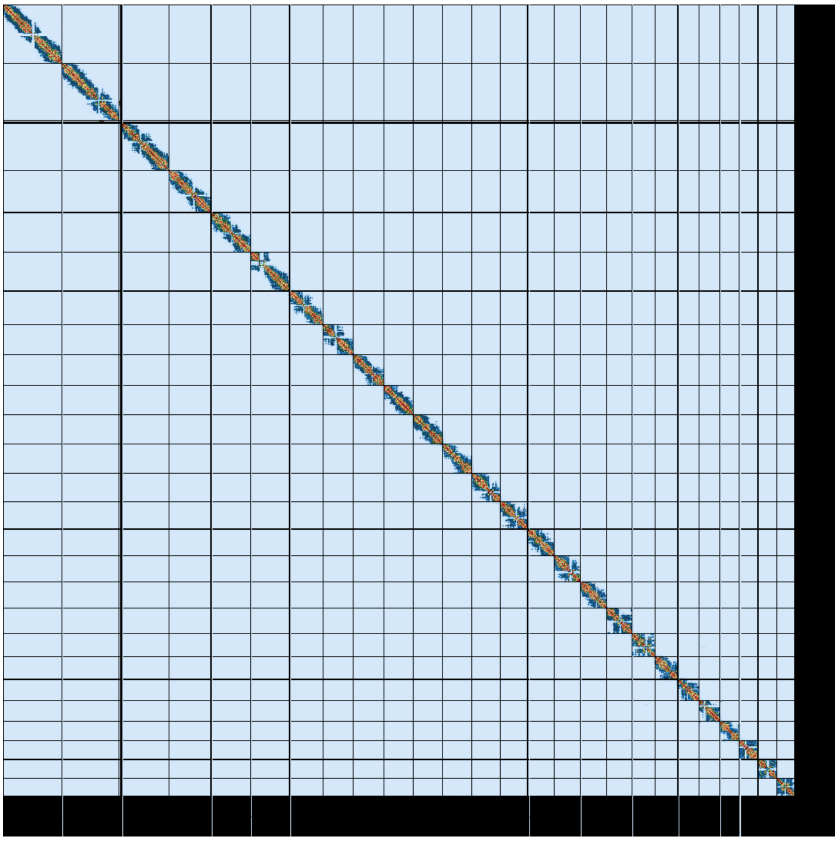
confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	2,109,473,454	1,752,039,481
GC %	41.44	41.34
Gaps/Gbp	1,336.83	1,270.52
Total gap bp	282,000	229,200
Scaffolds	1,630	1,198
Scaffold N50	65,344,123	62,101,932
Scaffold L50	12	11
Scaffold L90	52	24
Contigs	4,450	3,424
Contig N50	1,029,000	1,270,493
Contig L50	620	406
Contig L90	2,068	1,480
QV	52.1751	51.857
Kmer compl.	91.1952	85.5639
BUSCO sing.	84.5%	85.3%
BUSCO dupl.	8.5%	3.1%
BUSCO frag.	5.4%	5.4%
BUSCO miss.	1.6%	6.2%

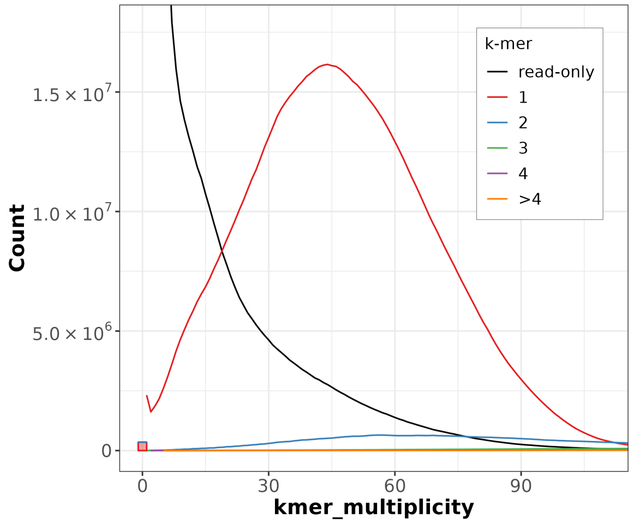
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: eukaryota_odb12 (genomes:456, BUSCOs:129)

HiC contact map of curated assembly

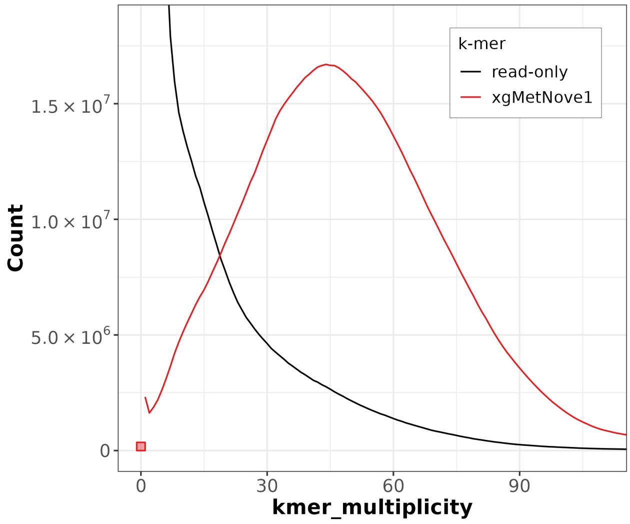


collapsed [LINK]

K-mer spectra of curated assembly

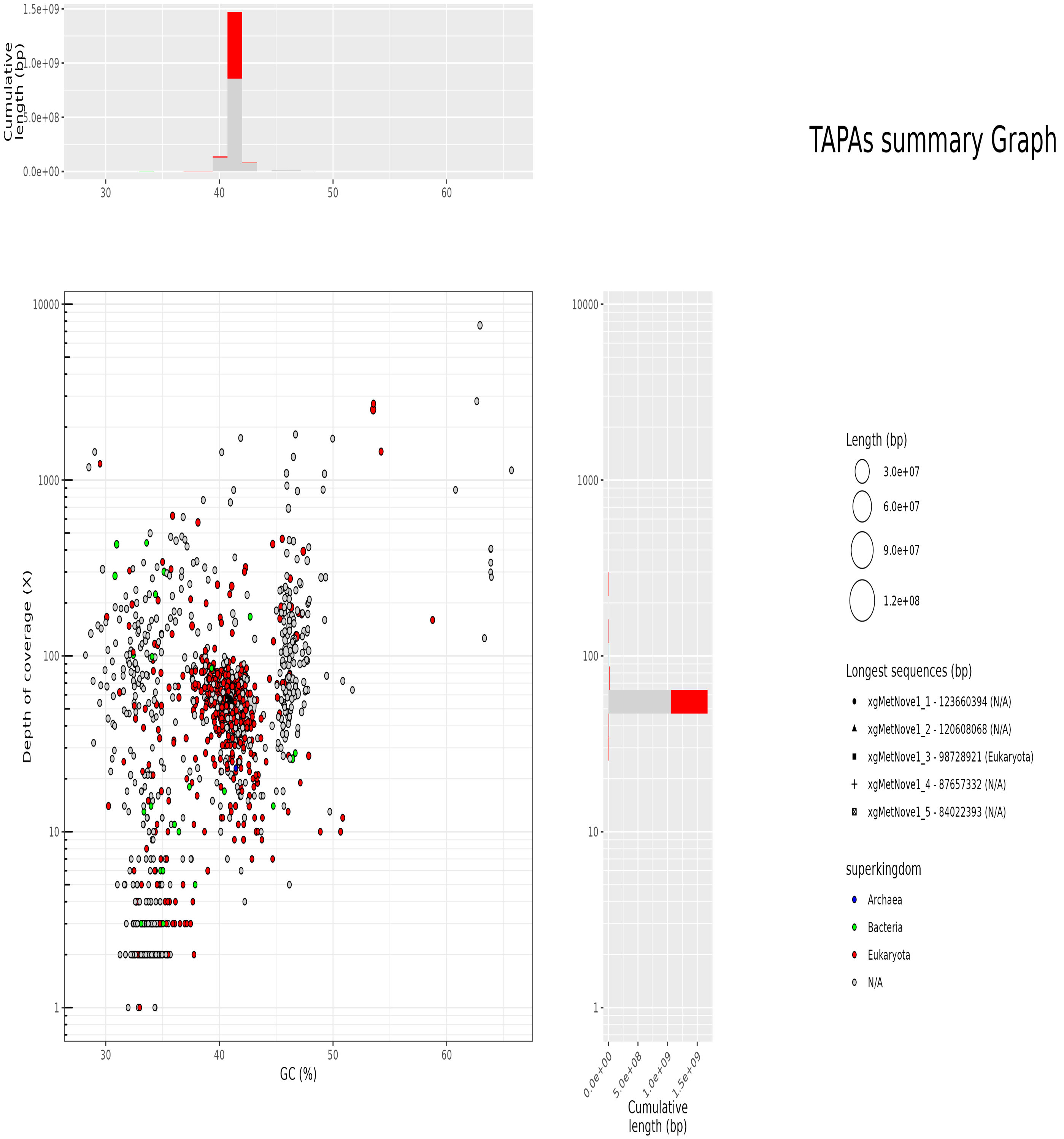


Distribution of k-mer counts per copy numbers found in asm



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	PACBIO Hifi	Arima
Coverage	63	29

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: S.Duprat
Affiliation: Genoscope

Date and time: 2025-04-25 13:52:46 CEST