ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	395376
ToLID	**kaPhaFumi1.1**
Species	Phallusia fumigata
Class	Ascidiacea
Order	Phlebobranchia

Genome Traits	Expected	Observed
Haploid size (bp)	190,593,277	189,978,158
Haploid Number	9 (source: ancestor)	8
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q61

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

Curator notes

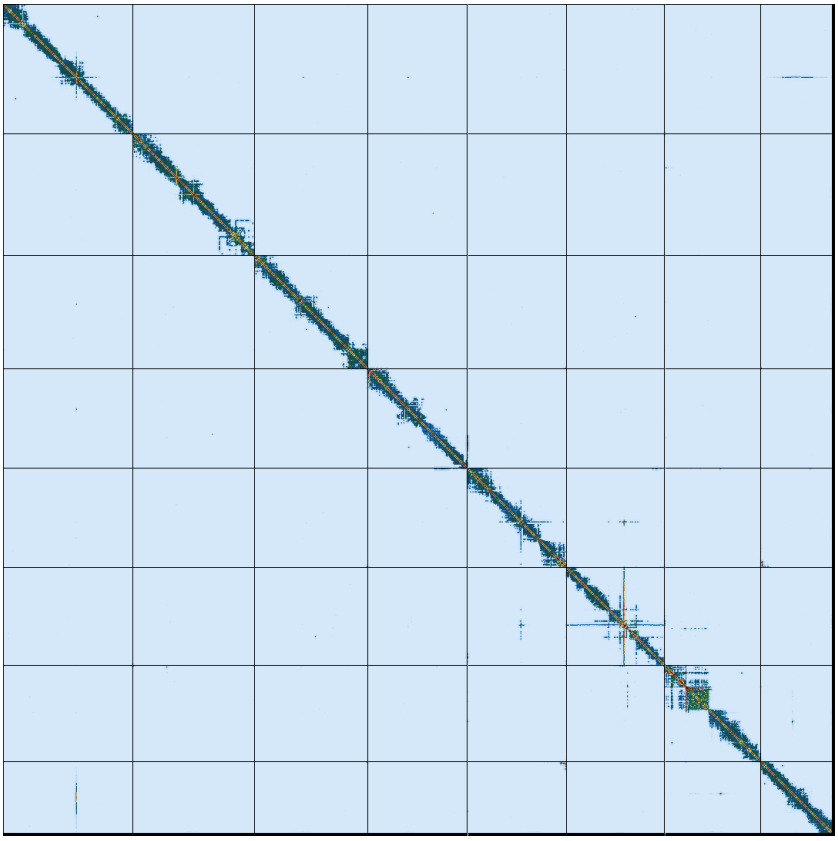
. Interventions/Gb: 292
. Contamination notes: ""
. Other observations: "The assembly of Phallusia fumigata (kaPhaFumi1) is based on 129X PacBio data and Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio 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, 1 contig was identified as contaminant (bacterial), totaling 57 kb. Additionally, 173 regions totaling 19 Mb (with the largest being 3.6 Mb) were identified as haplotypic duplications and removed using purge_dups. The mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 15 haplotypic regions and 3 contaminant sequences were removed, totaling 2 Mb and 61 kb respectively (with the largest being 980 kb and 27 kb). 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	192,368,061	189,978,158
GC %	38.85	38.83
Gaps/Gbp	135.16	178.97
Total gap bp	2,600	5,200
Scaffolds	39	29
Scaffold N50	24,043,024	22,805,738
Scaffold L50	4	4
Scaffold L90	7	7
Contigs	65	63
Contig N50	9,985,990	9,971,324
Contig L50	8	8
Contig L90	21	21
QV	47.7126	61.1674
Kmer compl.	68.2718	68.3525
BUSCO sing.	92.3%	92.5%
BUSCO dupl.	0.6%	0.5%
BUSCO frag.	2.6%	2.6%
BUSCO miss.	4.5%	4.4%

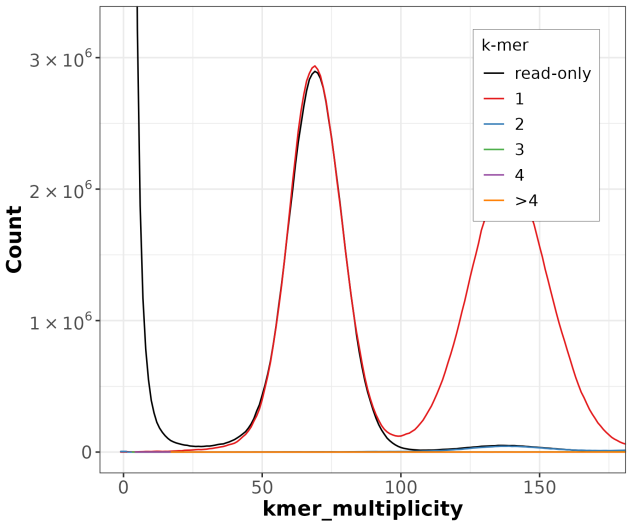
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: metazoa_odb10 (genomes:65, BUSCOs:954)

HiC contact map of curated assembly

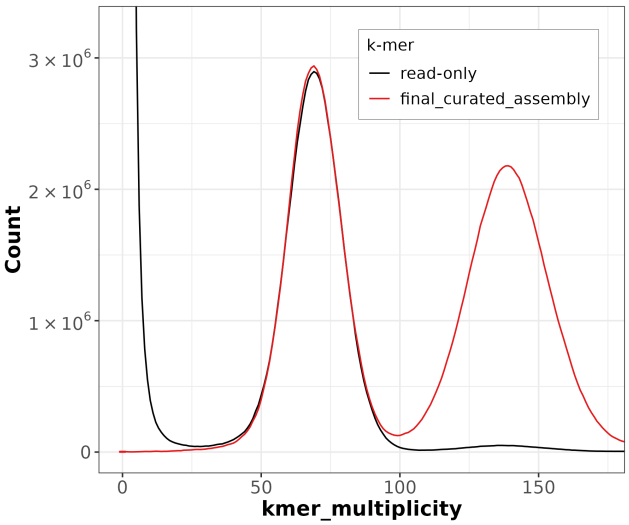


collapsed [LINK]

K-mer spectra of curated assembly

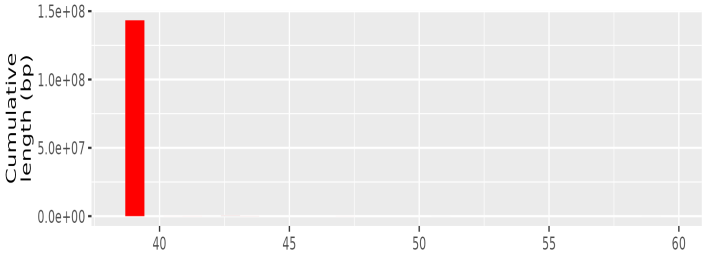


Distribution of k-mer counts per copy numbers found in asm

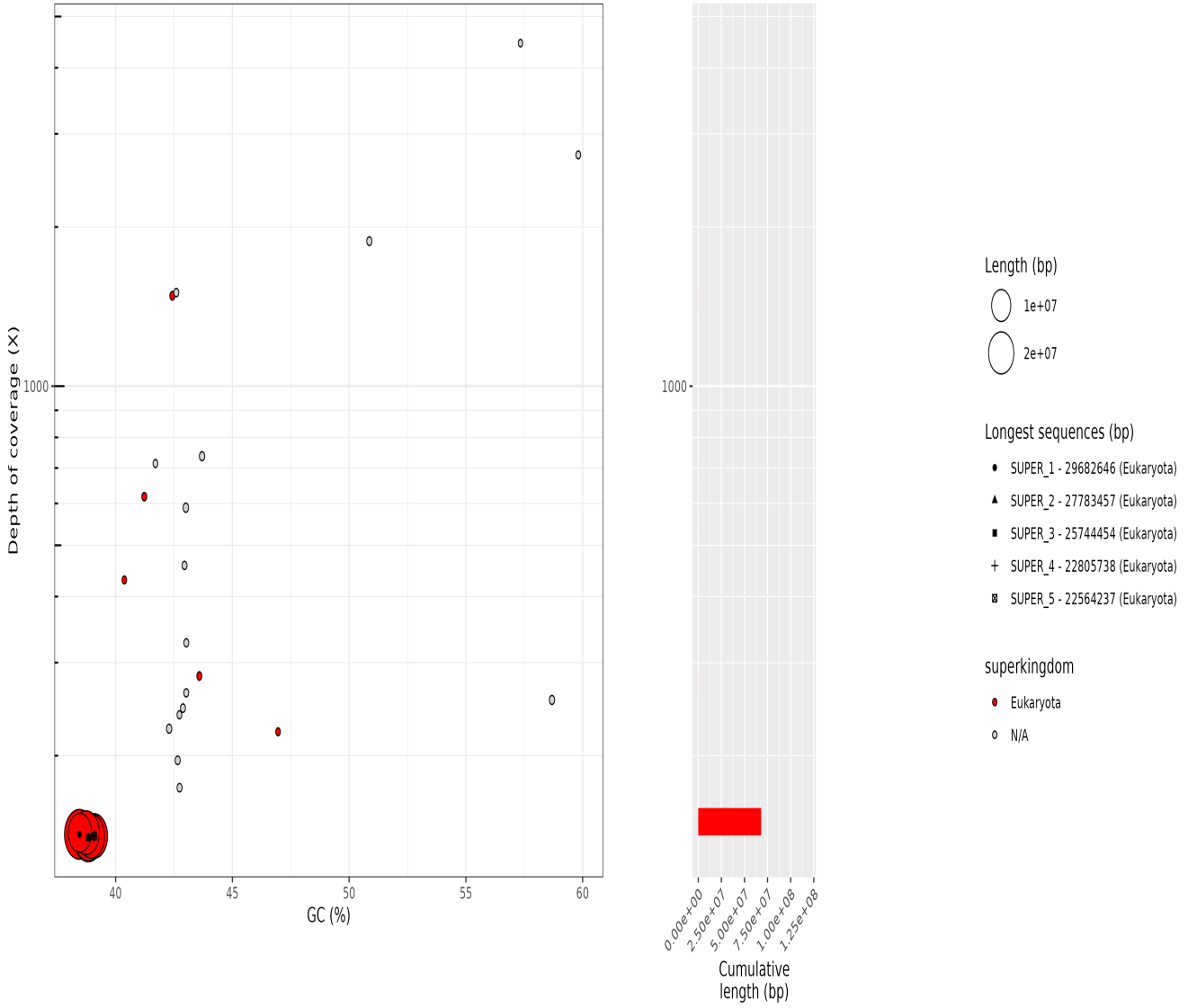


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	129	263

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: Emilie Teodori
Affiliation: Genoscope

Date and time: 2024-12-16 09:01:45 CET