

ERGA Assembly Report

v24.10.15

Tags: ATLASEa[INVALID TAG]

TxID	201963
ToLID	kaHalPapi5
Species	Halocynthia papillosa
Class	Ascidacea
Order	Stolidobranchia

Genome Traits	Expected	Observed
Haploid size (bp)	150,213,787	152,245,280
Haploid Number	16 (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.Q60

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
- . Not 90% of assembly in chromosomes for collapsed

Curator notes

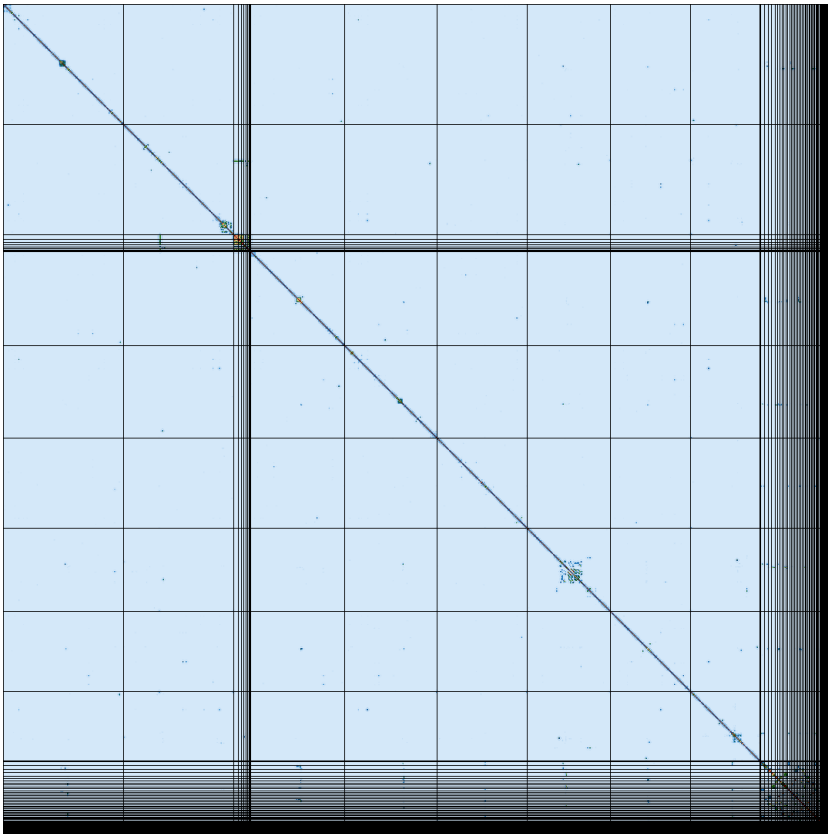
- . Interventions/Gb: 129
- . Contamination notes: ""
- . Other observations: "The assembly of Halocynthia papillosa (kaHalPapi5.1) is based on 34X of PacBio data and Arima Hi-C data generated as part of the ATLASEa programme (<https://www.atlasea.fr>). The assembly process included the following steps: initial PacBio assembly generation with Hifiiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 77 contigs were identified as contaminants (bacterial), totaling 2.7 Mb (with the largest being 0.45 Mb). Additionally, 291 regions totaling 24.9 Mb (with the largest being 0.84 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, 7 contaminant sequences and one haplotig region totaling respectively 0.19 Mb and 0.042 Mb (with the largest being 0.096 Mb and 0.042) were removed. 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	152,894,805	152,245,280
GC %	35.1	35.13
Gaps/Gbp	130.81	197.05
Total gap bp	2,000	4,400
Scaffolds	117	89
Scaffold N50	12,496,639	16,897,662
Scaffold L50	5	4
Scaffold L90	19	11
Contigs	137	119
Contig N50	6,987,731	6,987,731
Contig L50	9	9
Contig L90	36	35
QV	49.2663	60.4723
Kmer compl.	80.8387	83.3539
BUSCO sing.	91.9%	91.9%
BUSCO dupl.	0.2%	0.2%
BUSCO frag.	3.4%	3.4%
BUSCO miss.	4.5%	4.5%

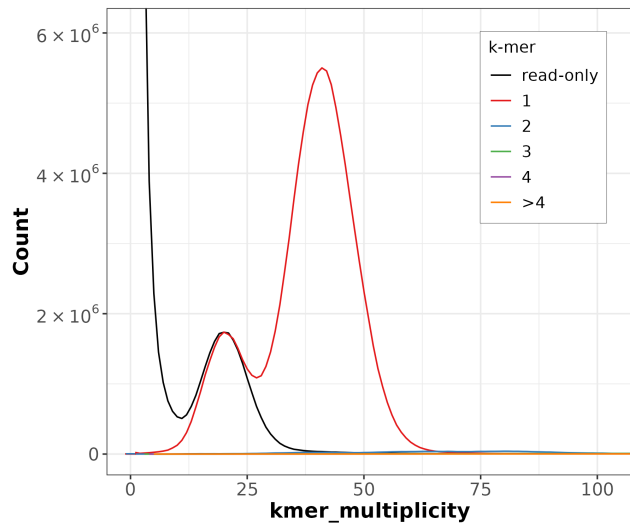
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: metazoa_odb10 (genomes:65, BUSCOs:954)

HiC contact map of curated assembly

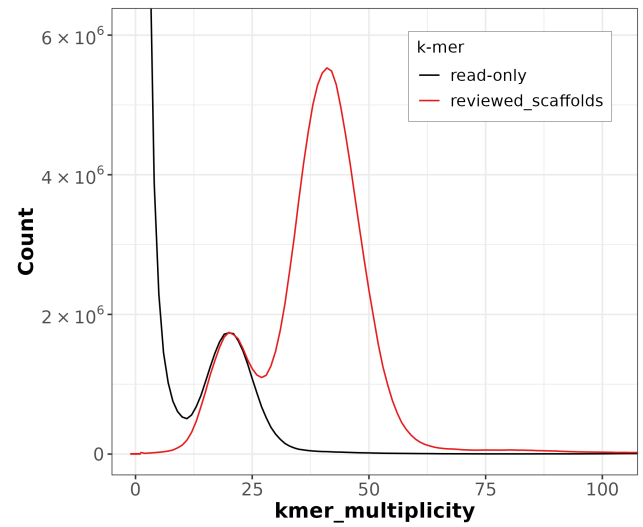


collapsed [\[LINK\]](#)

K-mer spectra of curated assembly

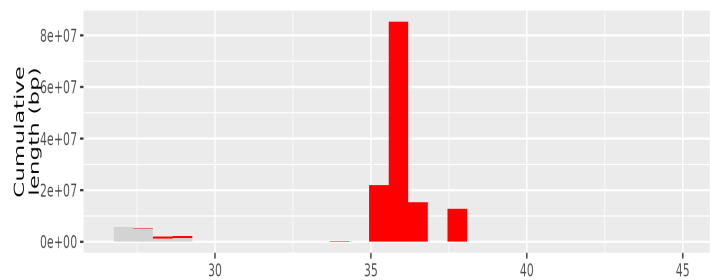


Distribution of k-mer counts per copy numbers found in asm

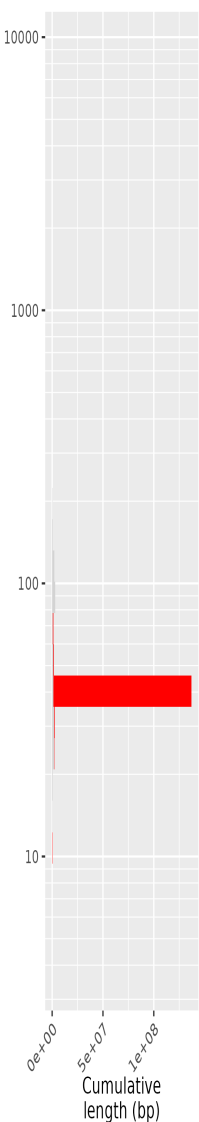
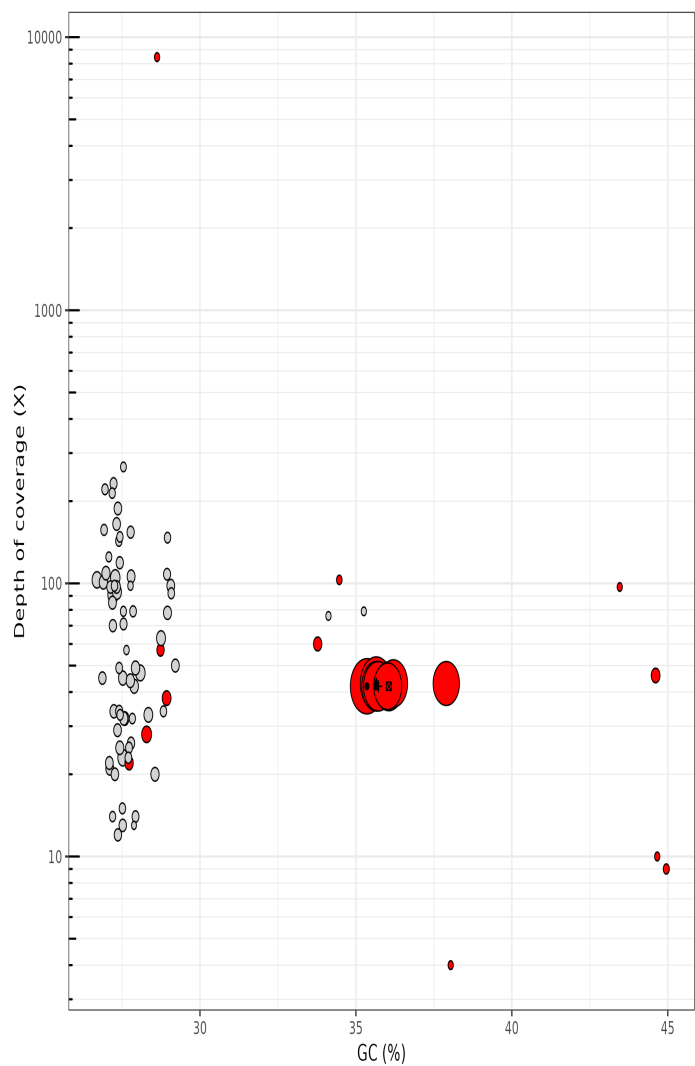


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



Longest sequences (bp)

- SUPER_1 - 22013227 (Eukaryota)
- ▲ SUPER_2 - 20136828 (Eukaryota)
- SUPER_3 - 17119572 (Eukaryota)
- + SUPER_4 - 16897662 (Eukaryota)
- ▣ SUPER_5 - 16508509 (Eukaryota)

superkingdom

- Eukaryota
- N/A

Length (bp)

- 5.0e+06
- 1.0e+07
- 1.5e+07
- 2.0e+07

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	33	1104

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: Adama Ndar

Affiliation: Genoscope

Date and time: 2025-04-03 11:53:36 CEST