ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3095174
ToLID	**wpHarAreo1**
Species	Harmothoe areolata
Class	Polychaeta
Order	Phyllodocida

Genome Traits	Expected	Observed
Haploid size (bp)	1,676,708,486	1,663,694,339
Haploid Number	10 (source: ancestor)	18
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q65

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
. Assembly length loss > 3% for collapsed

Curator notes

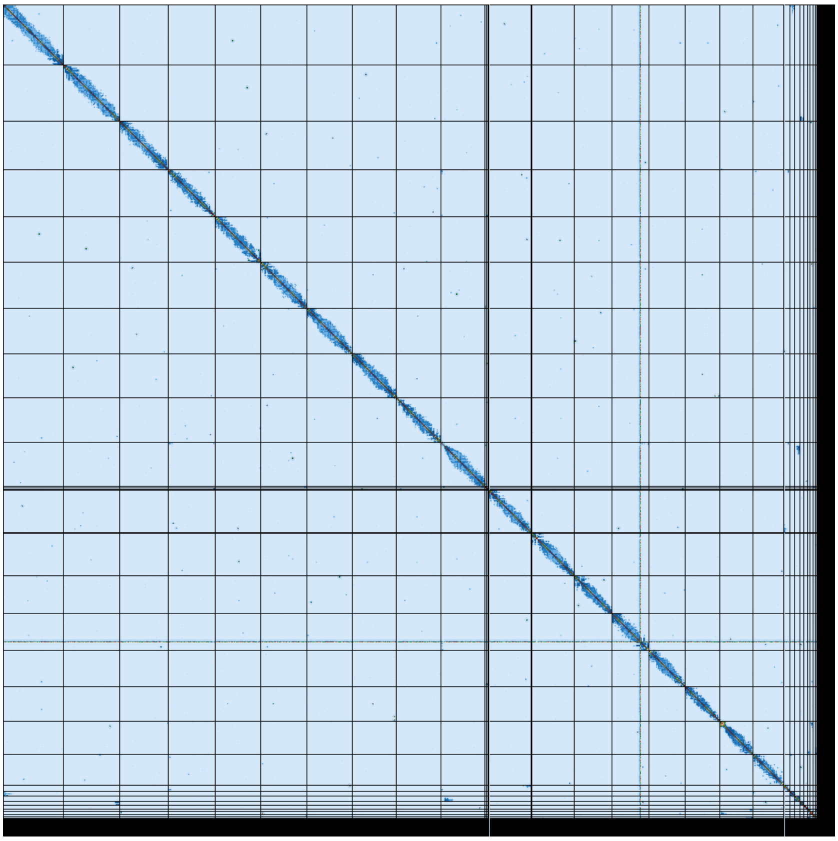
. Interventions/Gb: 51
. Contamination notes: ""
. Other observations: "The assembly of Harmothoe areolata (wpHarAreo1.1) is based on 41X PacBio data and Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio 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, 853 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 40.89 Mb (with the largest being 1.79 Mb). Additionally, 591 regions totaling 125.49 Mb (with the largest being 6.41 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, 42 haplotypic regions were removed, totaling 72.56 Mb (with the largest being 6.43 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	1,736,240,422	1,663,694,339
GC %	41.41	41.41
Gaps/Gbp	159.54	142.45
Total gap bp	27,700	26,300
Scaffolds	304	279
Scaffold N50	91,980,063	88,728,759
Scaffold L50	8	9
Scaffold L90	17	18
Contigs	581	516
Contig N50	19,258,107	20,708,742
Contig L50	24	22
Contig L90	116	97
QV	50.0958	65.013
Kmer compl.	68.7894	67.2441
BUSCO sing.	91.2%	93.7%
BUSCO dupl.	3.7%	1.0%
BUSCO frag.	2.8%	2.8%
BUSCO miss.	2.3%	2.5%

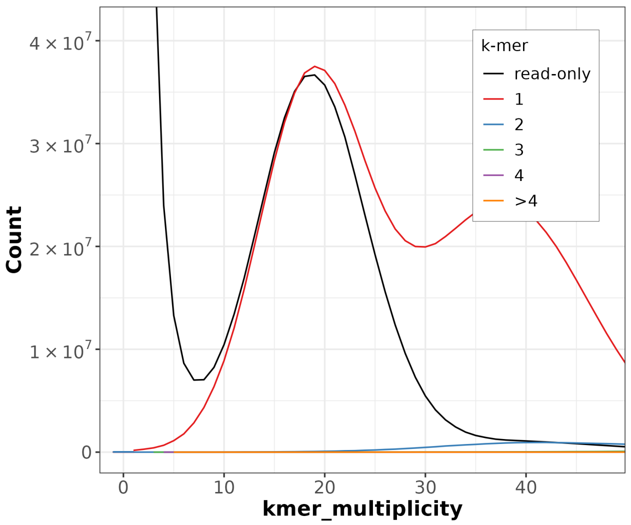
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: metazoa_odb10 (genomes:65, BUSCOs:954)

HiC contact map of curated assembly

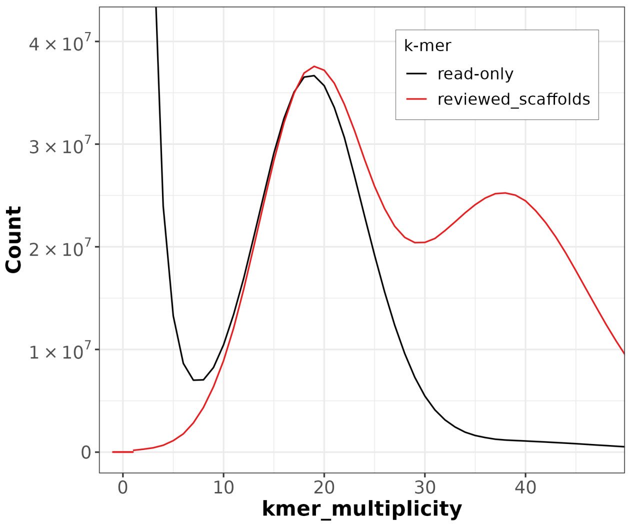


collapsed [LINK]

K-mer spectra of curated assembly

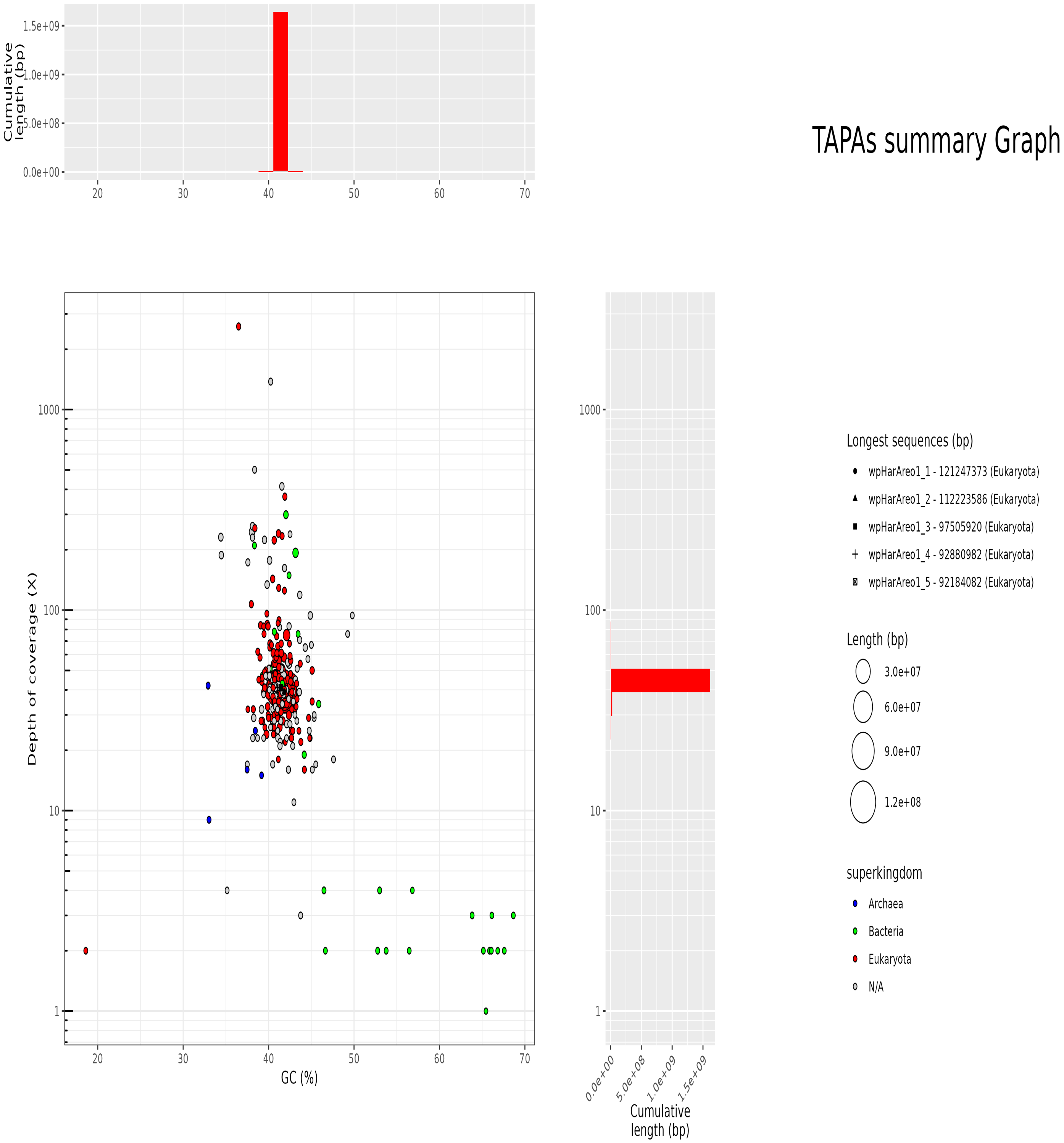


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	41	186

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: Jean-Marc Aury
Affiliation: Genoscope

Date and time: 2025-04-27 15:39:07 CEST