ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1928866
ToLID	**wjLeoHara2**
Species	Leodice harassii
Class	Polychaeta
Order	Eunicida

Genome Traits	Expected	Observed
Haploid size (bp)	2,285,927,748	2,313,196,816
Haploid Number	3 (source: ancestor)	12
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q48

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

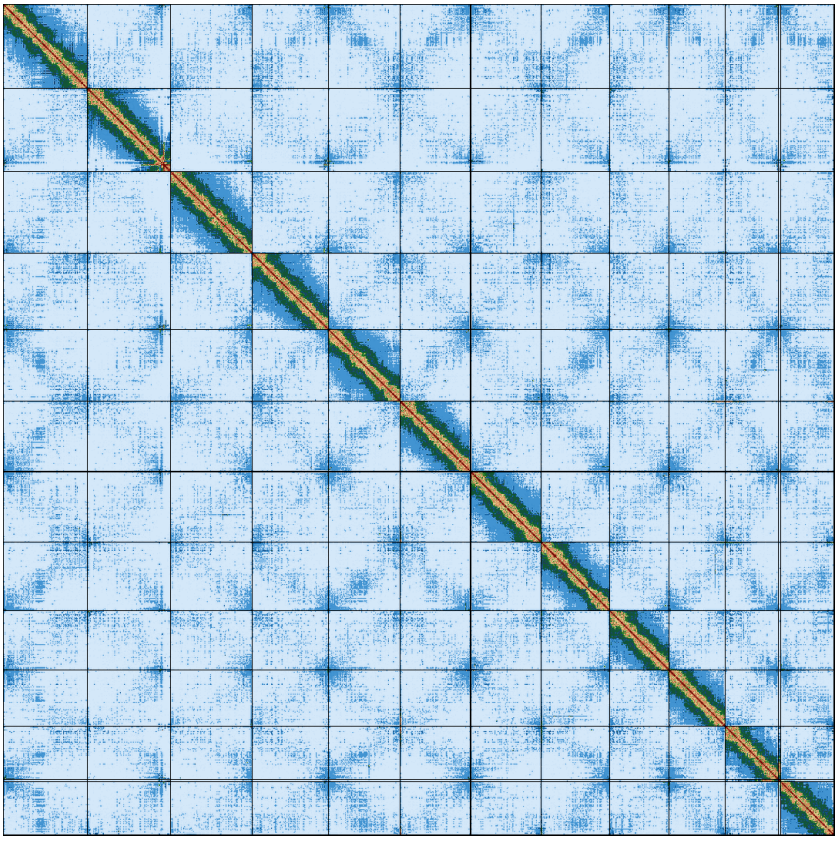
. Interventions/Gb: 32
. Contamination notes: ""
. Other observations: "The assembly of Leodice harassii (wjLeoHara2) is based on 44,7X PacBio data and 185,6X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial 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, 8 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 2.413 Mb (with the largest being 2.016 Mb). Additionally, 439 regions totaling 281.739 Mb (with the largest being 10.238 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, 26 haplotypic regions were removed, totaling 206.5Mb, (with the largest being 31.8Mb). 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	2,520,279,737	2,313,196,816
GC %	39.54	39.54
Gaps/Gbp	40.87	43.23
Total gap bp	10,300	13,200
Scaffolds	79	59
Scaffold N50	218,242,820	195,450,055
Scaffold L50	6	6
Scaffold L90	11	11
Contigs	182	159
Contig N50	44,928,000	50,651,150
Contig L50	17	15
Contig L90	67	55
QV	48.871	48.8593
Kmer compl.	66.6931	63.0761
BUSCO sing.	85.1%	91.1%
BUSCO dupl.	6.8%	0.6%
BUSCO frag.	6.6%	6.7%
BUSCO miss.	1.5%	1.5%

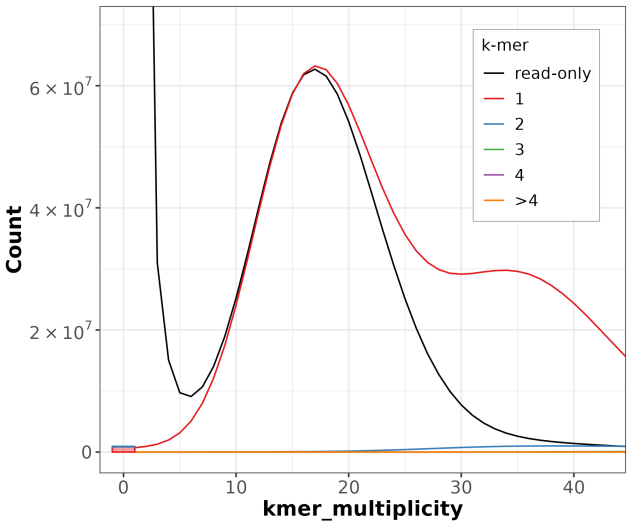
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: lophotrochozoa_odb12 (genomes:75, BUSCOs:1252)

HiC contact map of curated assembly

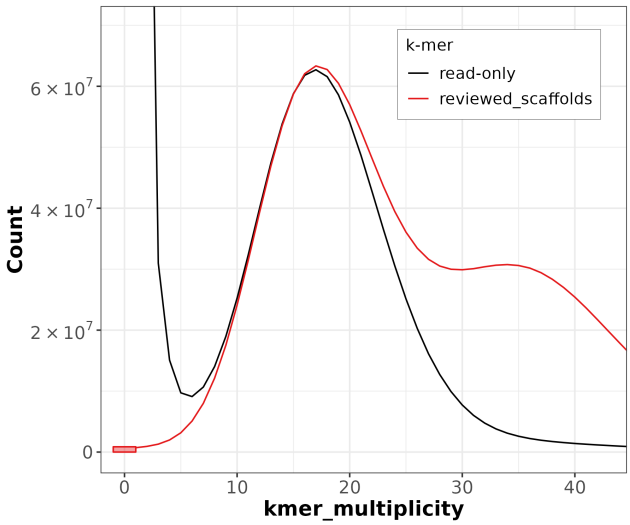


collapsed [LINK]

K-mer spectra of curated assembly

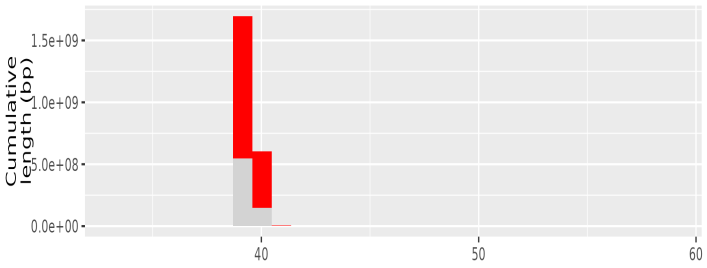


Distribution of k-mer counts per copy numbers found in asm

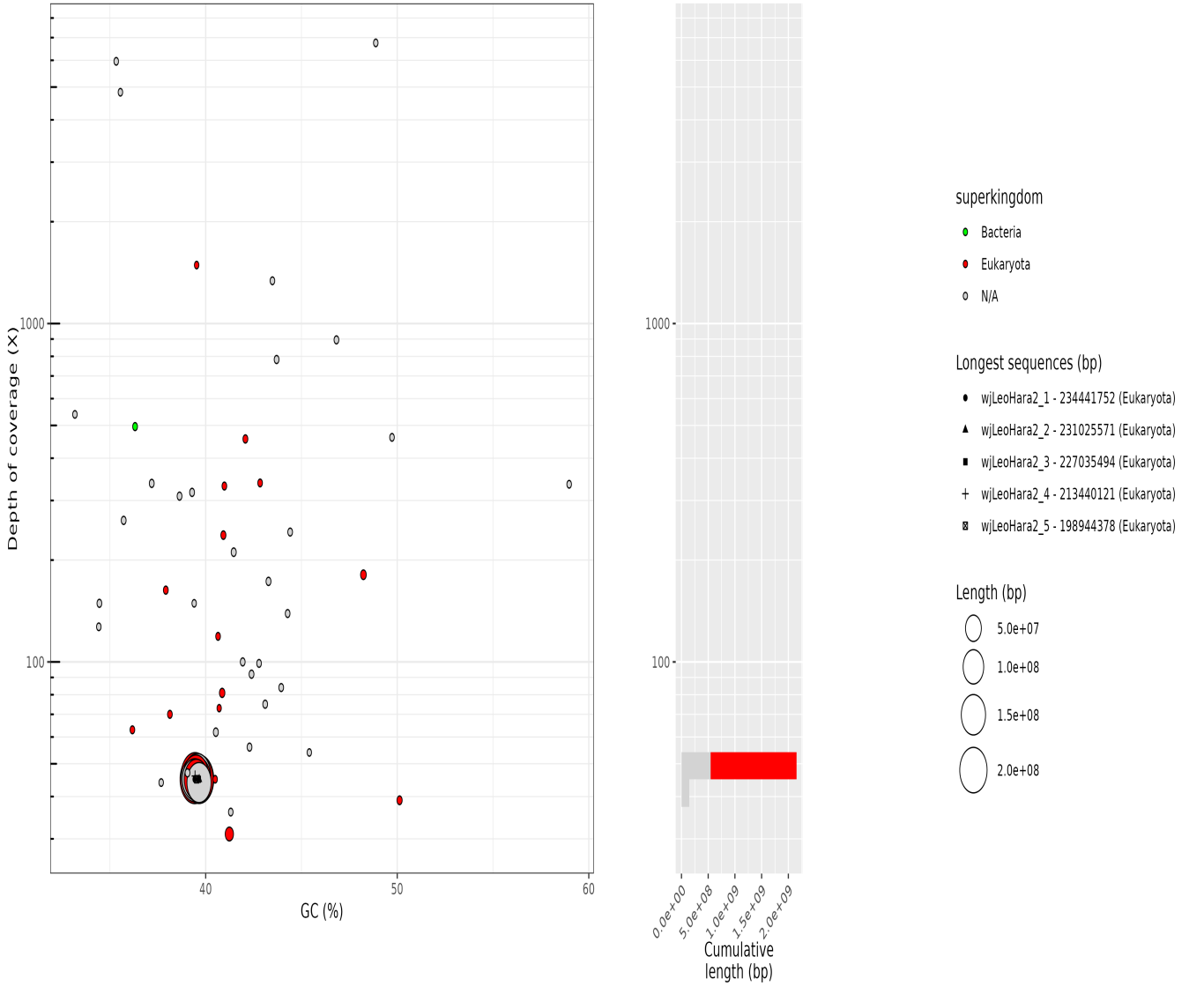


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	44	185

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: 2025-07-26 23:03:16 CEST