ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3269394
ToLID	**wsPerRubr1**
Species	Perkinsiana rubra
Class	Polychaeta
Order	Sabellida

Genome Traits	Expected	Observed
Haploid size (bp)	1,512,736,028	1,488,416,639
Haploid Number	4 (source: ancestor)	13
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q49

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

Curator notes

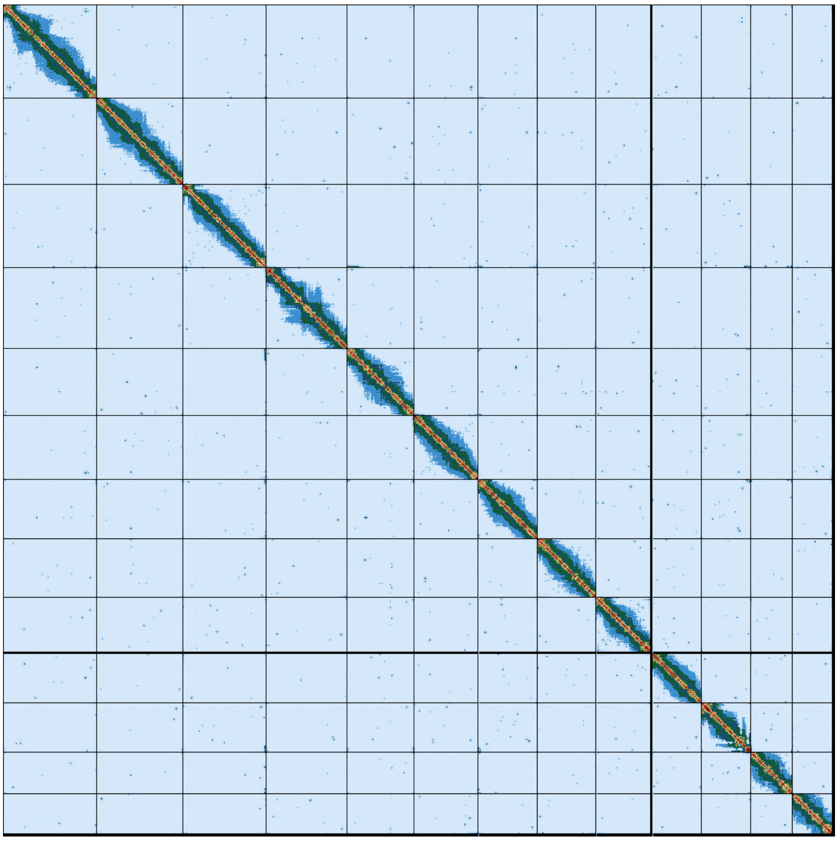
. Interventions/Gb: 56
. Contamination notes: ""
. Other observations: "The assembly of Perkinsiana rubra (wsPerRubr1) is based on 97X ONT data and 194X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial ONT 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, 155 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 1.908 Mb (with the largest being 0.328 Mb). Additionally, 5541 regions totaling 110.222 Mb (with the largest being 6.629 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, 44 haplotypic regions were removed, totaling 137.478Mb, (with the largest being 60.15Mb). 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	1,627,983,106	1,488,416,639
GC %	31.94	31.92
Gaps/Gbp	69.41	81.97
Total gap bp	11,300	16,100
Scaffolds	2,207	86
Scaffold N50	117,859,084	114,571,942
Scaffold L50	6	6
Scaffold L90	13	12
Contigs	2,320	208
Contig N50	17,924,912	18,485,000
Contig L50	27	23
Contig L90	90	77
QV	48.6145	49.7691
Kmer compl.	78.993	75.0511
BUSCO sing.	85.5%	85.5%
BUSCO dupl.	2.9%	1.4%
BUSCO frag.	9.2%	9.1%
BUSCO miss.	2.5%	4.0%

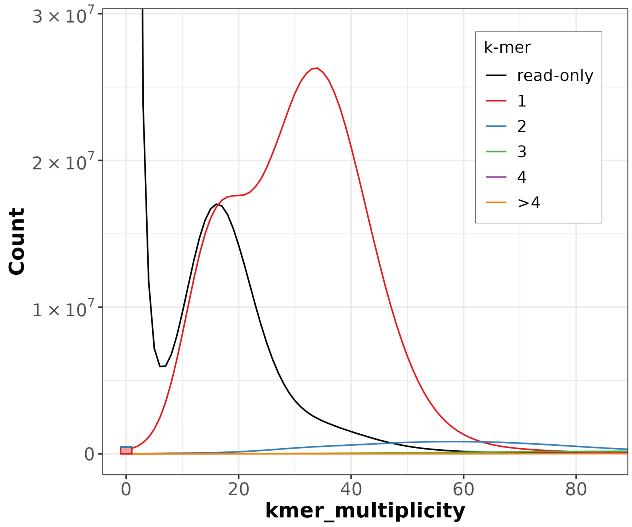
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: lophotrochozoa_odb12 (genomes:75, BUSCOs:1252)

HiC contact map of curated assembly

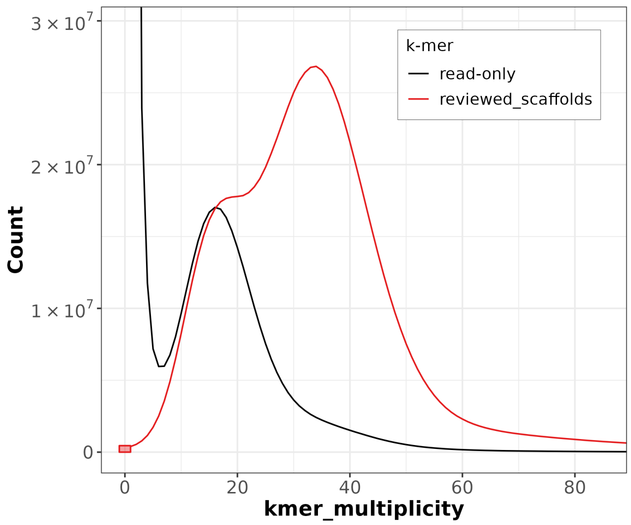


collapsed [LINK]

K-mer spectra of curated assembly

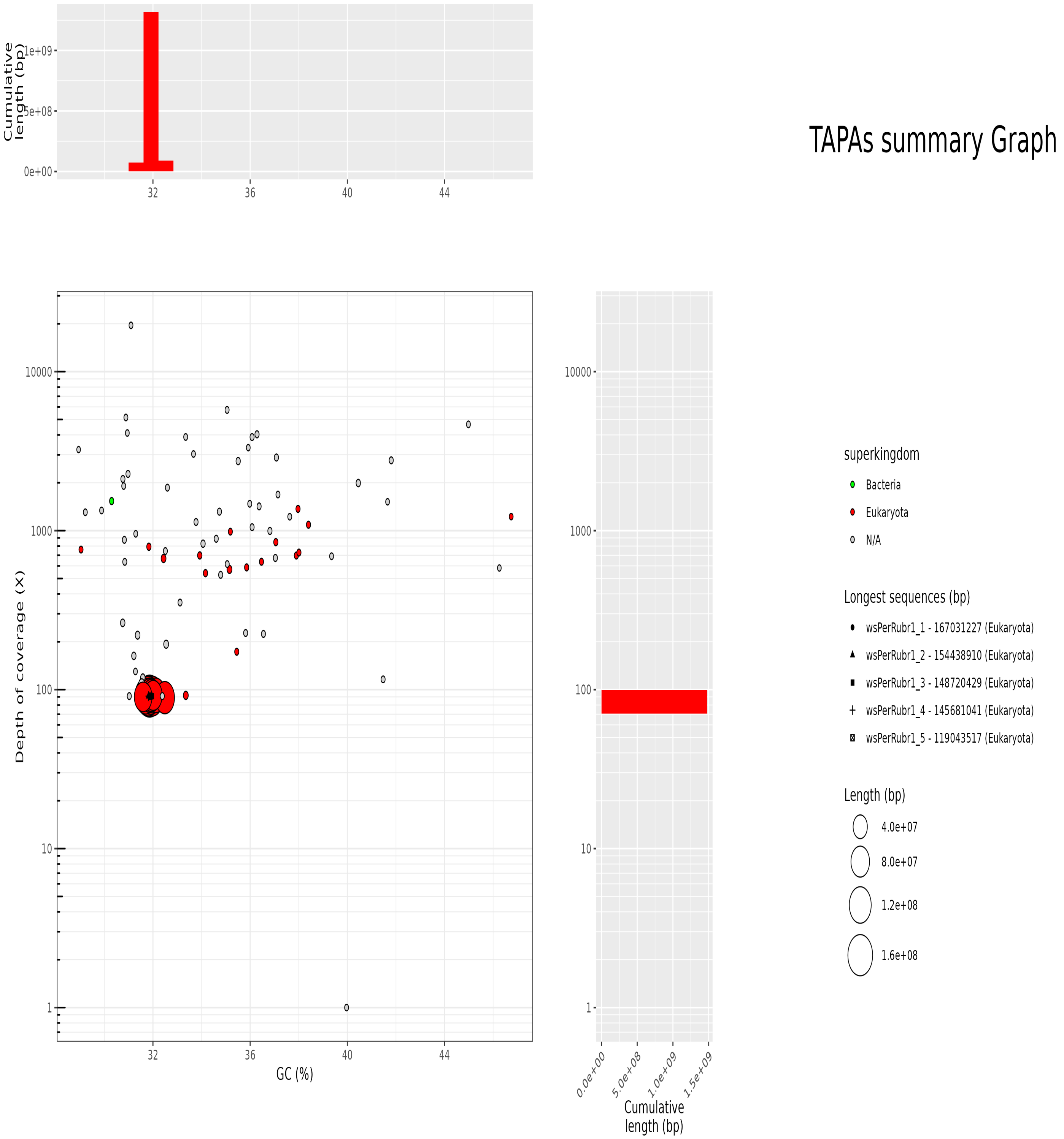


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	96	194

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: 2025-08-02 16:29:04 CEST