ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	154647
ToLID	**xgMarBlai1**
Species	Marionia blainvillea
Class	Gastropoda
Order	Nudibranchia

Genome Traits	Expected	Observed
Haploid size (bp)	1,163,329,747	1,257,981,301
Haploid Number	11 (source: ancestor)	21
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q43

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
. More than 1000 gaps/Gbp for collapsed
. Not 90% of assembly in chromosomes for collapsed

Curator notes

. Interventions/Gb: 74
. Contamination notes: ""
. Other observations: "The assembly of Marionia blainvillea (xgMarBlai1) is based on 41X PacBio data and 238X 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, 71 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 12.571 Mb (with the largest being 2.518 Mb). Additionally, 2211 regions totaling 293.632 Mb (with the largest being 3.838 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 haplotypic regions and 218 contaminant sequences were removed, totaling 2.484 Mb and 26.2 Mb, respectively (with the largest being 0.474 Mb and 1.999 Mb). Chromosome-scale scaffolds confirmed by

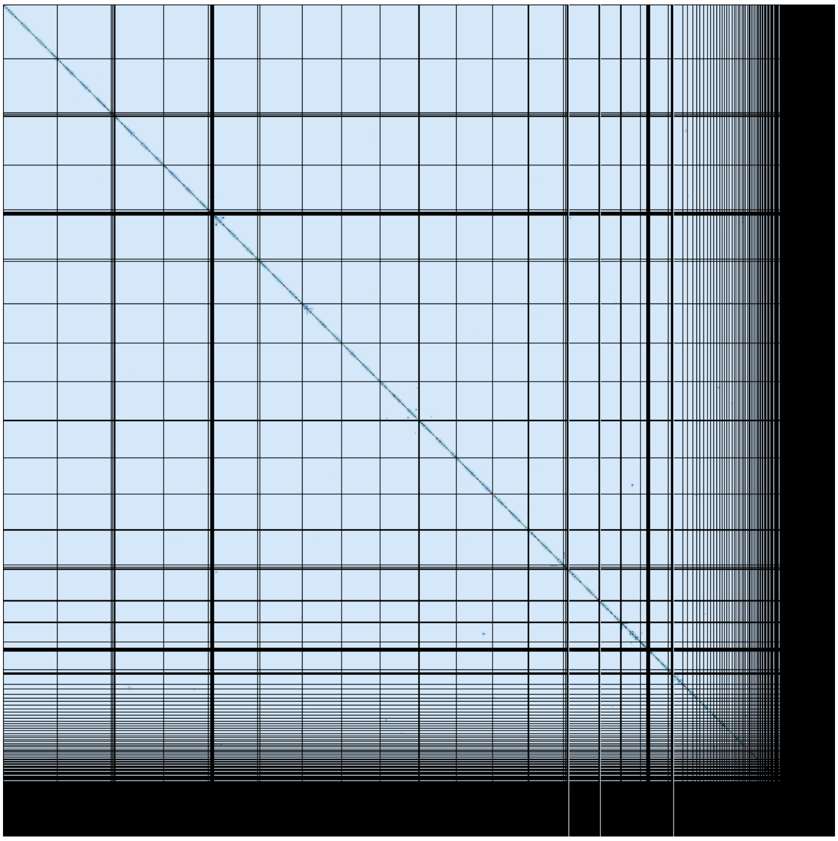
Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	1,295,160,762	1,257,981,301
GC %	36.34	36.63
Gaps/Gbp	1,393.65	1,332.29
Total gap bp	180,500	171,000
Scaffolds	1,482	802
Scaffold N50	34,031,641	56,625,305
Scaffold L50	11	10
Scaffold L90	83	62
Contigs	3,287	2,478
Contig N50	1,312,002	1,347,335
Contig L50	276	262
Contig L90	1,106	1,002
QV	42.1643	43.5652
Kmer compl.	72.2657	71.5231
BUSCO sing.	85.3%	85.5%
BUSCO dupl.	1.7%	1.5%
BUSCO frag.	2.8%	2.8%
BUSCO miss.	10.2%	10.2%

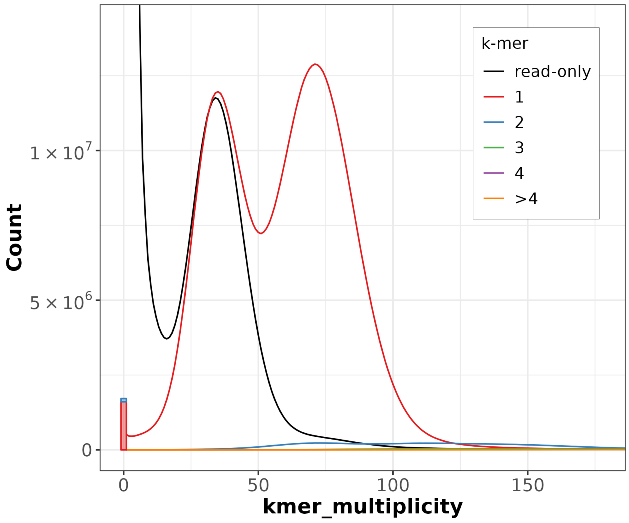
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: mollusca_odb10 (genomes:7, BUSCOs:5295)

HiC contact map of curated assembly

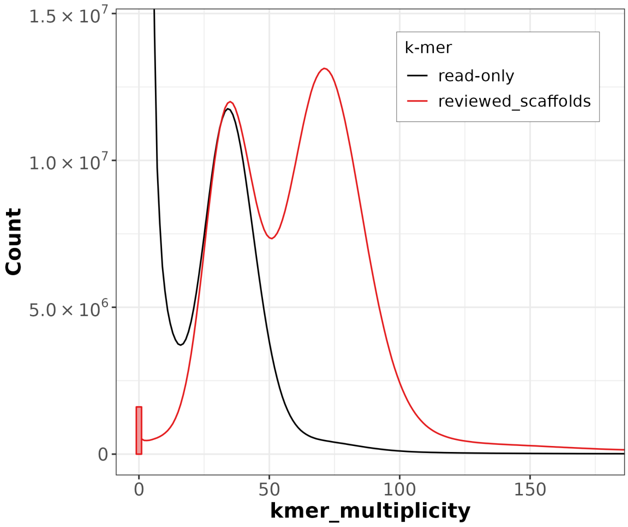


collapsed [LINK]

K-mer spectra of curated assembly

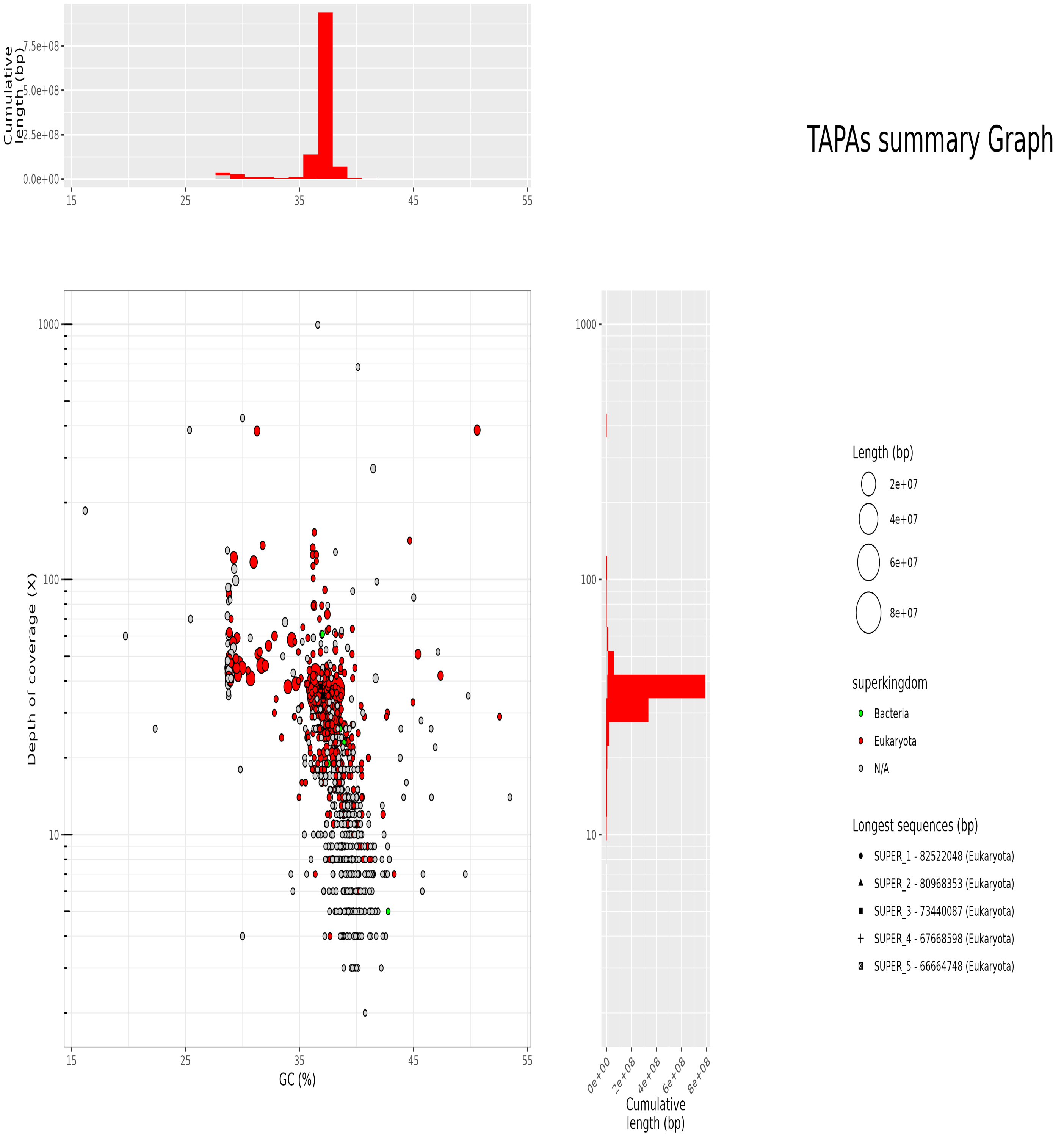


Distribution of k-mer counts per copy numbers found in asm



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	172

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: Benjamin Istace
Affiliation: Genoscope

Date and time: 2025-06-04 09:48:29 CEST