ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	516032
ToLID	**kaDidVexi2**
Species	Didemnum vexillum
Class	Ascidiacea
Order	Aplousobranchia

Genome Traits	Expected	Observed
Haploid size (bp)	844,609,776	829,469,921
Haploid Number	9 (source: ancestor)	19
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.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

Curator notes

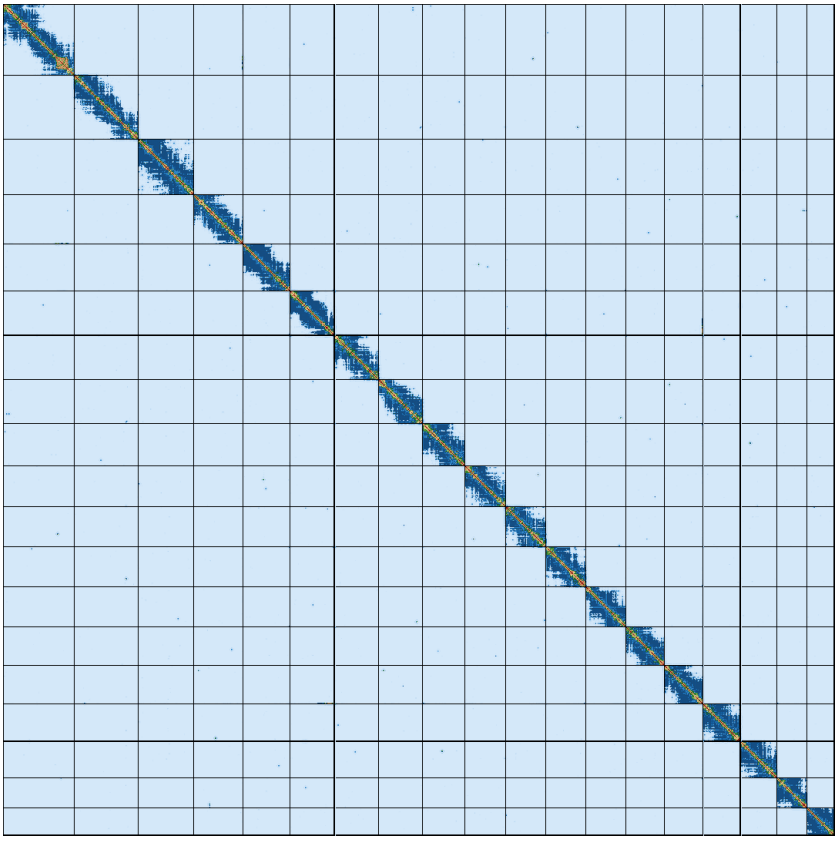
. Interventions/Gb: 31
. Contamination notes: ""
. Other observations: "The assembly of Didemnum vexillum (kaDidVexi2) is based on 104X ONT data and 154X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial ONT 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, 164 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 23.101 Mb (with the largest being 8.618 Mb). Additionally, 1611 regions totaling 55.117 Mb (with the largest being 12.259 Mb) were identified as haplotypic duplications and removed. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 9 haplotypic regions and 22 contaminant sequences were removed, totaling 8.859 Mb and 1.138 Mb, respectively (with the largest being 3.657 Mb and 0.154 Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. 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	842,886,083	829,469,921
GC %	36.09	36.13
Gaps/Gbp	28.47	40.99
Total gap bp	2,400	4,900
Scaffolds	2,711	91
Scaffold N50	42,855,594	43,759,100
Scaffold L50	9	8
Scaffold L90	17	17
Contigs	2,735	125
Contig N50	29,972,850	31,732,028
Contig L50	11	10
Contig L90	28	25
QV	39.8016	43.2696
Kmer compl.	75.4282	75.1122
BUSCO sing.	76.8%	77.7%
BUSCO dupl.	2.7%	1.6%
BUSCO frag.	12.9%	12.9%
BUSCO miss.	7.6%	7.7%

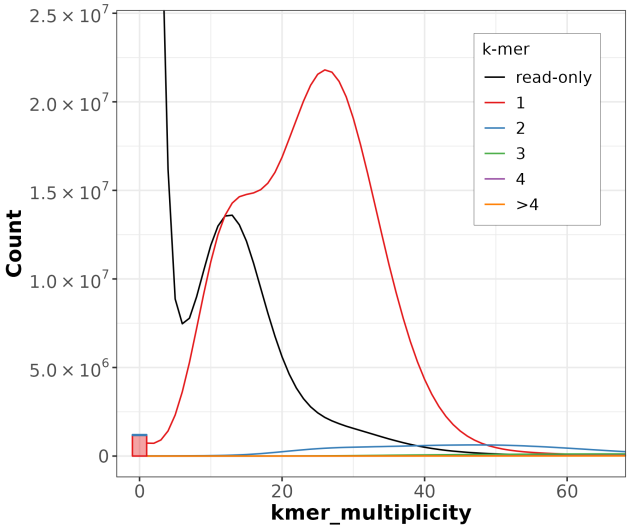
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: metazoa_odb12 (genomes:206, BUSCOs:672)

HiC contact map of curated assembly

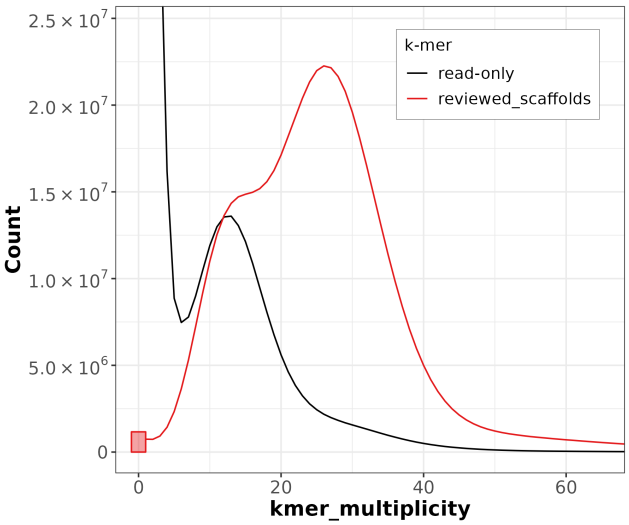


collapsed [LINK]

K-mer spectra of curated assembly

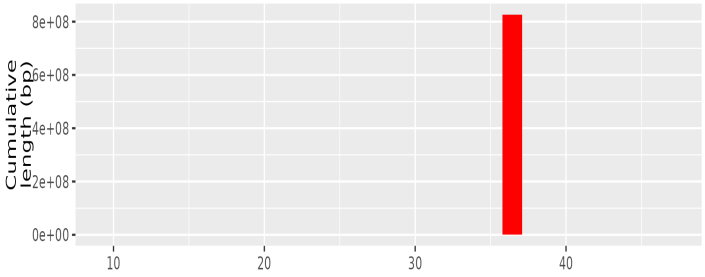


Distribution of k-mer counts per copy numbers found in asm

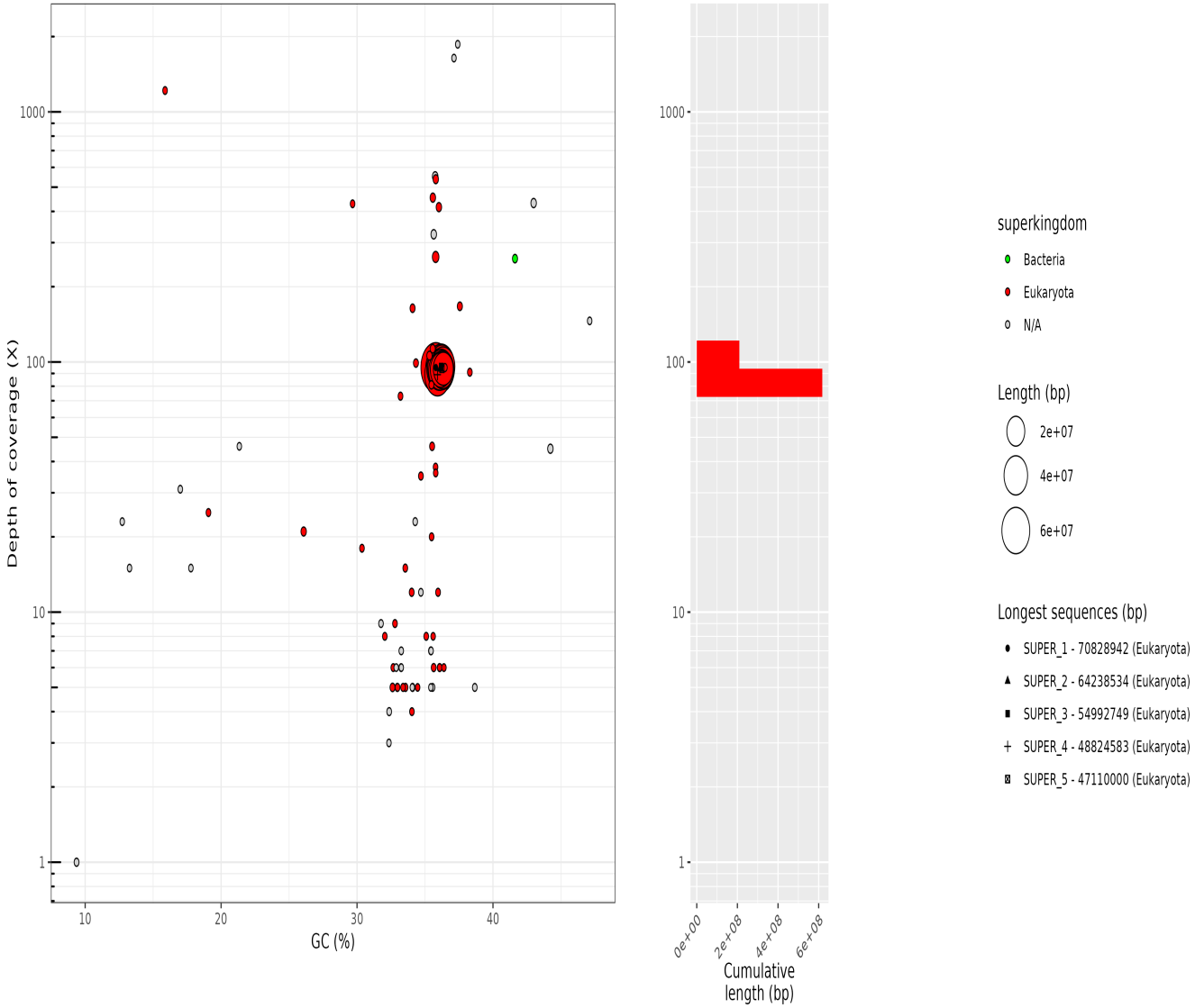


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	104	154

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: 2025-07-14 08:32:38 CEST