ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3074273
ToLID	**xbDonSemi1**
Species	Donax semistriatus
Class	Bivalvia
Order	Cardiida

Genome Traits	Expected	Observed
Haploid size (bp)	1,325,838,155	1,383,602,231
Haploid Number	26 (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.Q67

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

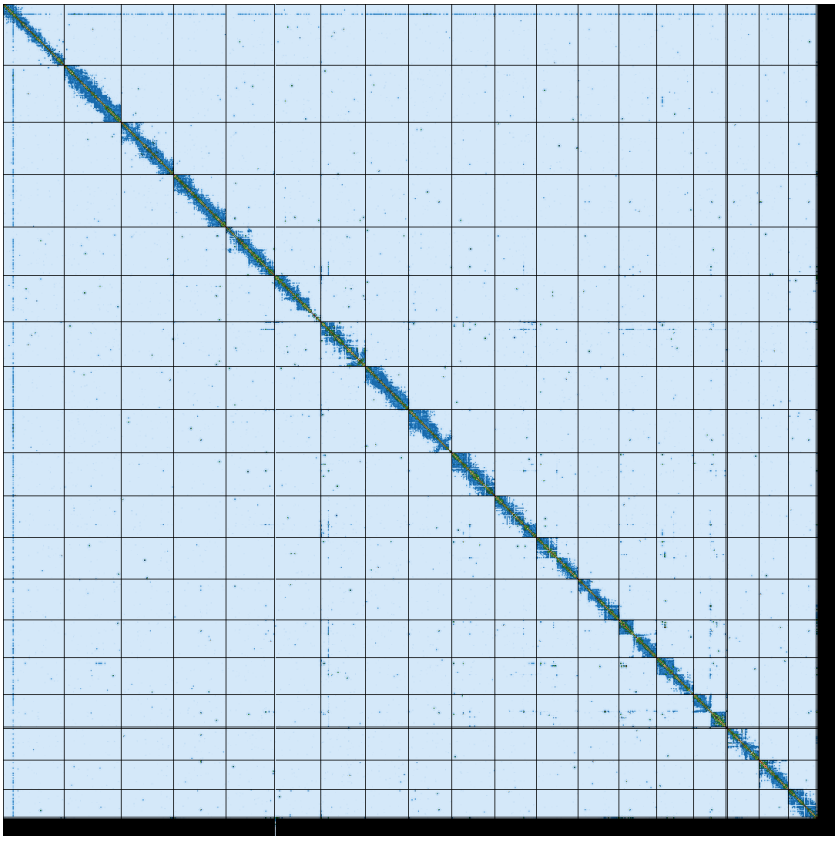
. Interventions/Gb: 66
. Contamination notes: ""
. Other observations: "The assembly of Donax semistriatus (xbDonSemi1) is based on 60X PacBio data and Arima HighCoverage Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr).The assembly process included the following steps: initial 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, 2 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 679 Kb (with the largest being 571 Kb). Additionally, 405 regions totaling 1.1 Gb (with the largest being 28.8 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, 31 haplotypic regions were removed,totaling 89 Mb (with the largest being 14.7 Mb) "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	1,474,078,632	1,383,602,231
GC %	34.46	34.48
Gaps/Gbp	126.18	125.04
Total gap bp	18,600	19,400
Scaffolds	190	176
Scaffold N50	78,631,728	71,661,531
Scaffold L50	8	9
Scaffold L90	17	17
Contigs	376	349
Contig N50	15,236,000	17,138,325
Contig L50	30	24
Contig L90	93	77
QV	49.4658	67.811
Kmer compl.	60.6009	57.6476
BUSCO sing.	74.1%	78.5%
BUSCO dupl.	5.7%	1.0%
BUSCO frag.	4.3%	4.3%
BUSCO miss.	15.9%	16.2%

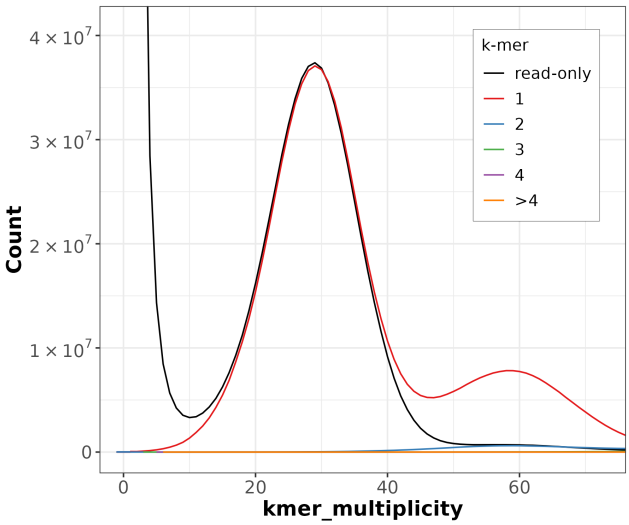
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: mollusca_odb10 (genomes:7, BUSCOs:5295)

HiC contact map of curated assembly

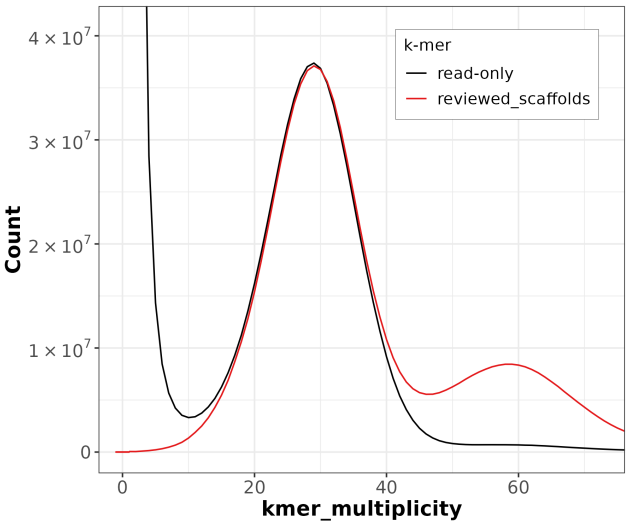


collapsed [LINK]

K-mer spectra of curated assembly

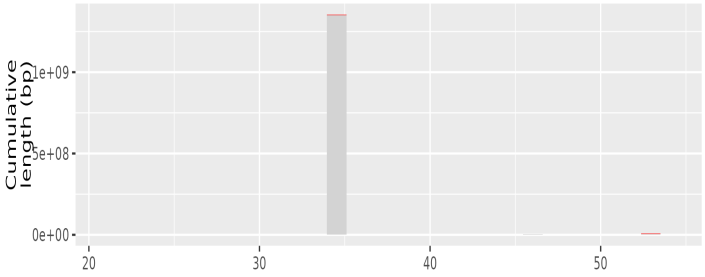


Distribution of k-mer counts per copy numbers found in asm

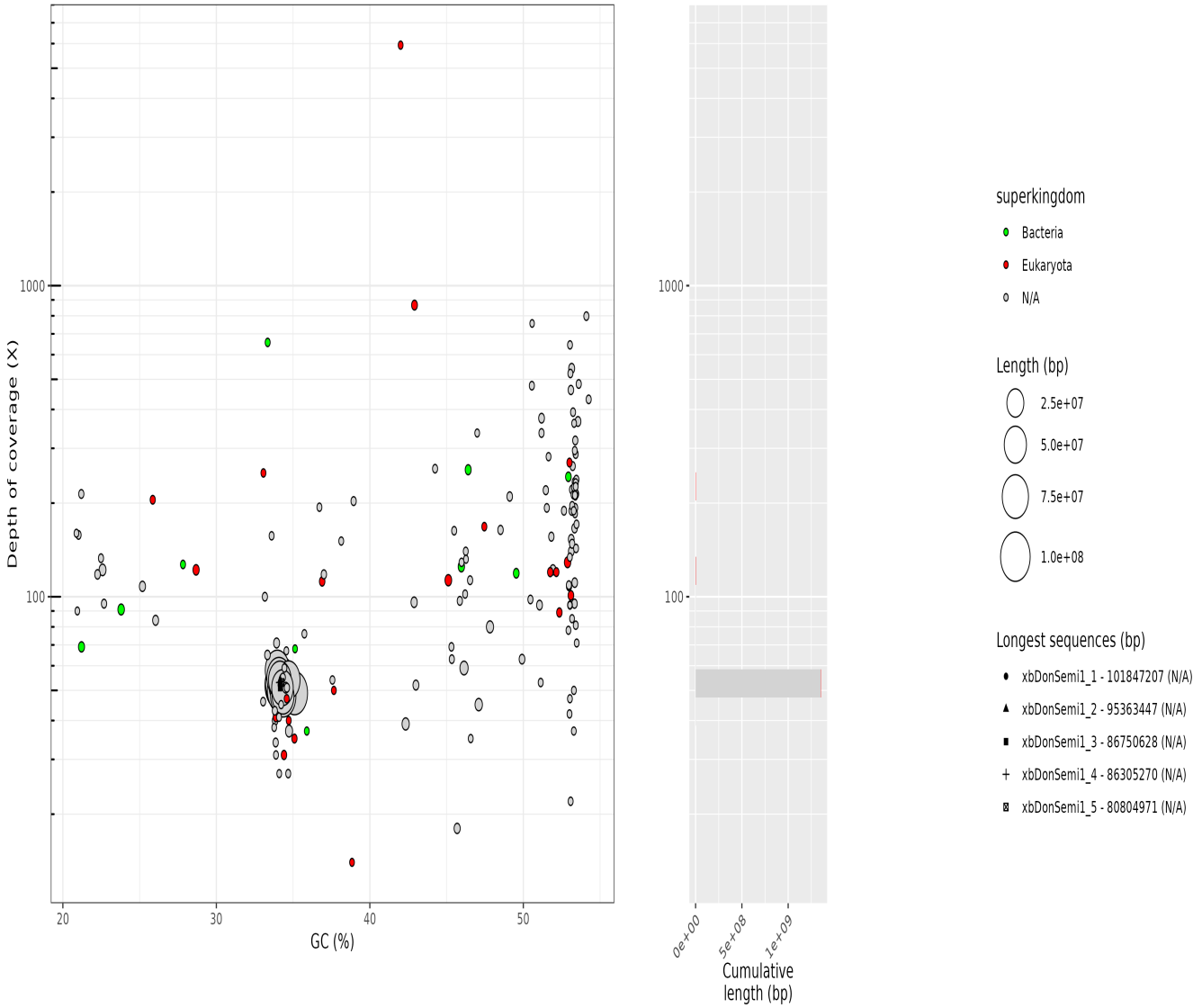


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



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	59	200

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: Caroline Menguy
Affiliation: Genoscope

Date and time: 2025-04-05 18:06:33 CEST