ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	30281
ToLID	**kaMolManh1**
Species	Molgula manhattensis
Class	Ascidiacea
Order	Stolidobranchia

Genome Traits	Expected	Observed
Haploid size (bp)	185,741,328	198,142,259
Haploid Number	16 (source: ancestor)	18
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.6.Q30

The following metrics were automatically flagged as below EBP recommended standards or different from expected:

. Observed Haploid Number is different from Expected

. QV value is less than 40 for collapsed
. Kmer completeness value is less than 90 for collapsed
. Not 90% of assembly in chromosomes for collapsed

Curator notes

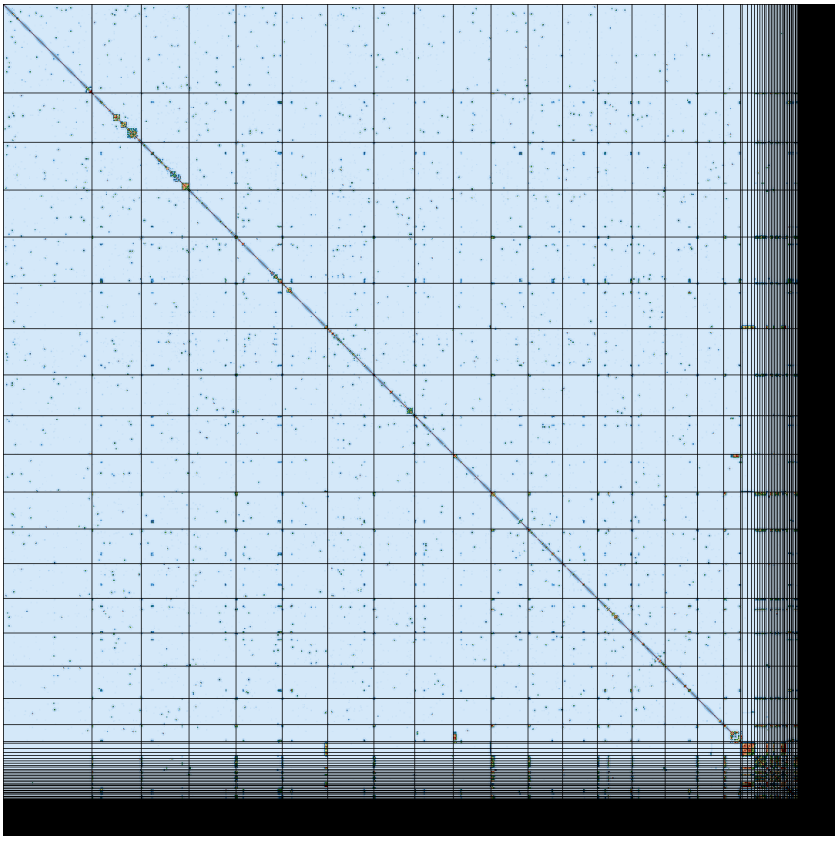
. Interventions/Gb: 199
. Contamination notes: ""
. Other observations: "The assembly of Molgula manhattensis (kaMolManh1) is based on 39.11X PacBio data and 205.28X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: 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, 82 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 3.774 Mb (with the largest being 0.882 Mb). Additionally, 175 regions totaling 13.281 Mb (with the largest being 0.29 Mb) were identified as haplotypic duplications and removed. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 7 haplotypic regions and 17 contaminant sequences were removed, totaling 2.361Mb and 0.784Mb, respectively (with the largest being 0.829Mb and 0.115Mb). 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	201,282,337	198,142,259
GC %	34.82	34.86
Gaps/Gbp	144.08	247.3
Total gap bp	2,900	7,300
Scaffolds	267	235
Scaffold N50	9,178,282	9,132,242
Scaffold L50	9	9
Scaffold L90	30	21
Contigs	296	284
Contig N50	6,303,038	5,564,133
Contig L50	12	12
Contig L90	51	51
QV	30.0325	30.5025
Kmer compl.	70.805	70.3745
BUSCO sing.	91.4%	92.4%
BUSCO dupl.	1.6%	0.6%
BUSCO frag.	2.8%	2.8%
BUSCO miss.	4.2%	4.2%

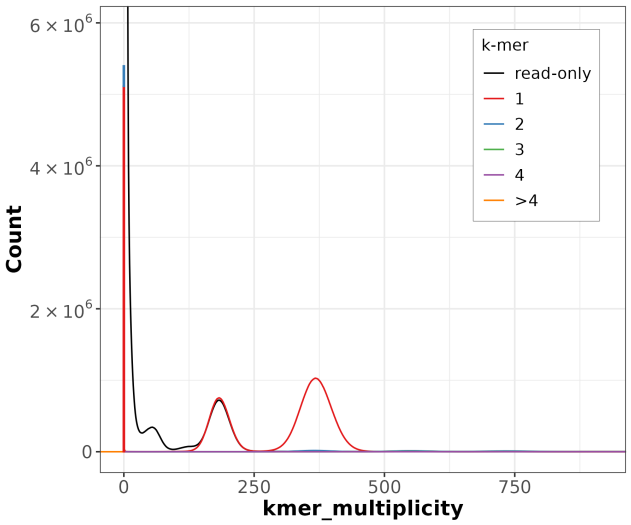
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: metazoa_odb12 (genomes:206, BUSCOs:672)
BUSCO: 6.1.0 (euk_genome_min, miniprot) / Lineage: metazoa_odb12 (genomes:206, BUSCOs:672)

HiC contact map of curated assembly

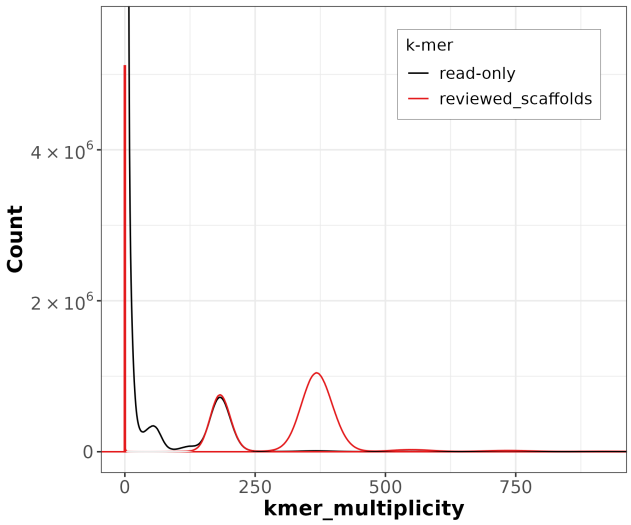


collapsed [LINK]

K-mer spectra of curated assembly

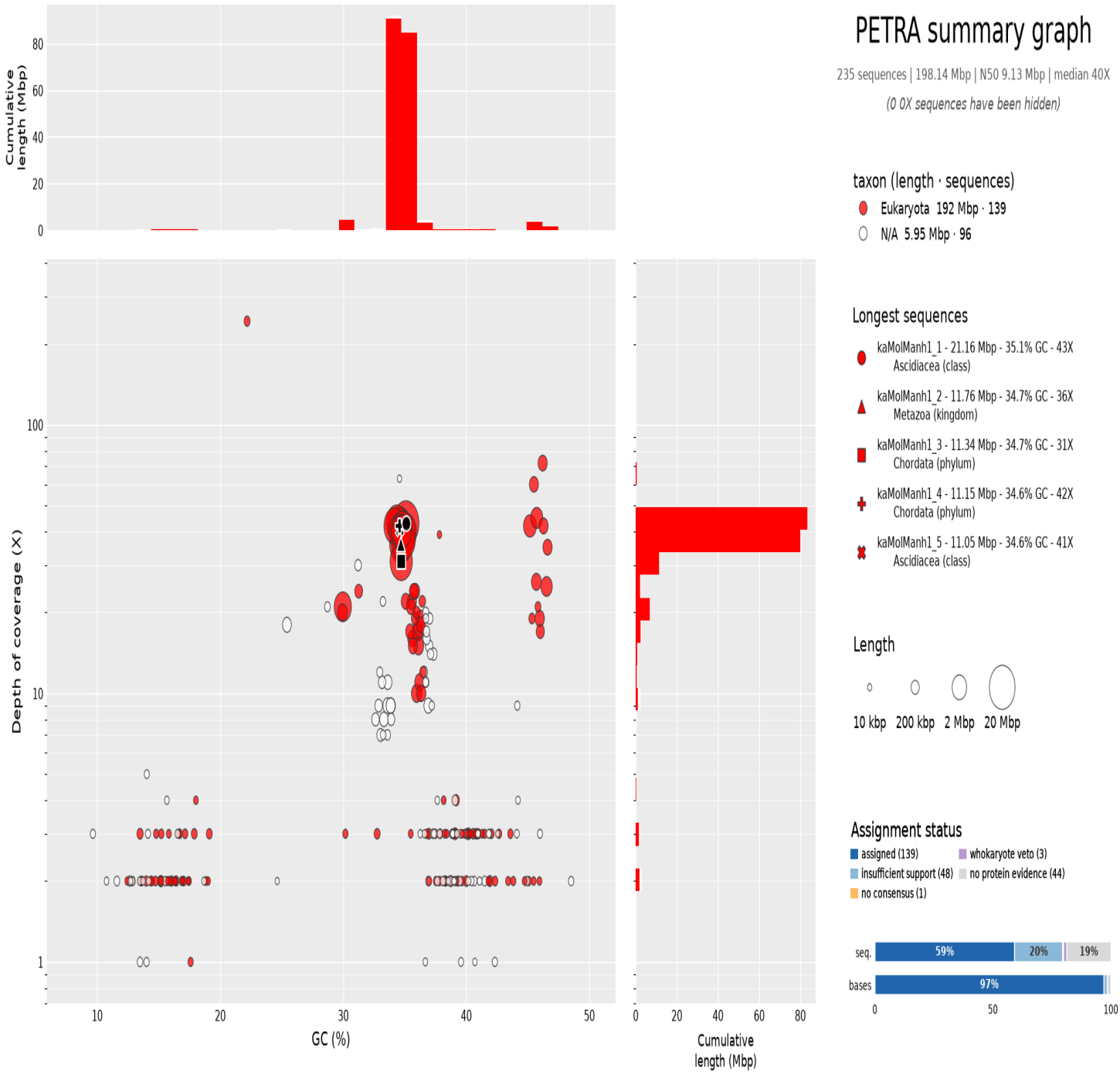


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	Long reads	Arima
Coverage	39	208

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:* 0.2.4
  |_ *key param:* NA
- **PretextView**
  |_ *ver:* 1.0.7
  |_ *key param:* NA

Submitter: Caroline Menguy
Affiliation: Genoscope

Date and time: 2026-10-06 09:26:49 CEST