ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	154642
ToLID	**xgCraPere1**
Species	Cratena peregrina
Class	Gastropoda
Order	Nudibranchia

Genome Traits	Expected	Observed
Haploid size (bp)	782,644,359	795,988,301
Haploid Number	11 (source: ancestor)	15
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q46

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

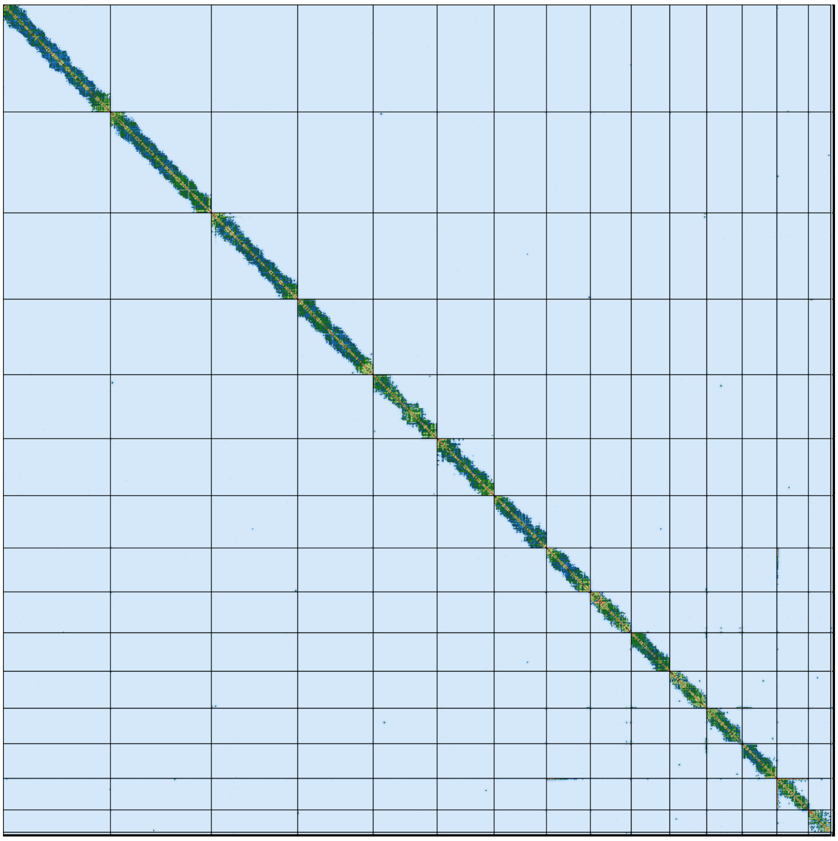
. Interventions/Gb: 62
. Contamination notes: ""
. Other observations: "The assembly of Cratena peregrina (xgCraPere1) is based on 63X PacBio data and 186X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio 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, 89 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 4.249 Mb (with the largest being 0.504 Mb). Additionally, 196 regions totaling 51.353 Mb (with the largest being 2.942 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, 12 haplotypic regions and 6 contaminant sequences were removed, totaling 9.554 Mb and 1.866 Mb, respectively (with the largest being 2.132 Mb and 1.605 Mb). During manual curation, 12 haplotypic regions were removed, totaling 9.554Mb, (with the largest being 2.132Mb). 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	807,537,413	795,988,301
GC %	38.82	38.83
Gaps/Gbp	147.36	165.83
Total gap bp	11,900	15,800
Scaffolds	87	53
Scaffold N50	61,050,429	61,050,429
Scaffold L50	5	5
Scaffold L90	13	13
Contigs	206	185
Contig N50	11,905,730	12,254,148
Contig L50	22	21
Contig L90	76	75
QV	46.1062	46.4837
Kmer compl.	75.0867	74.6671
BUSCO sing.	86.9%	87.7%
BUSCO dupl.	1.5%	0.7%
BUSCO frag.	8.4%	8.4%
BUSCO miss.	3.2%	3.2%

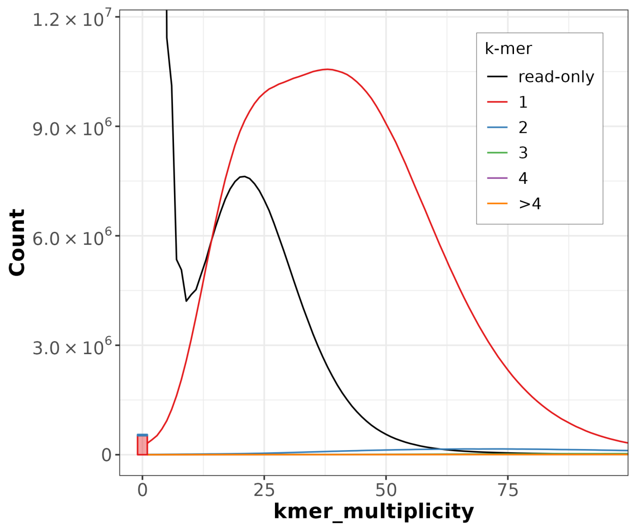
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: mollusca_odb12 (genomes:36, BUSCOs:4421)

HiC contact map of curated assembly

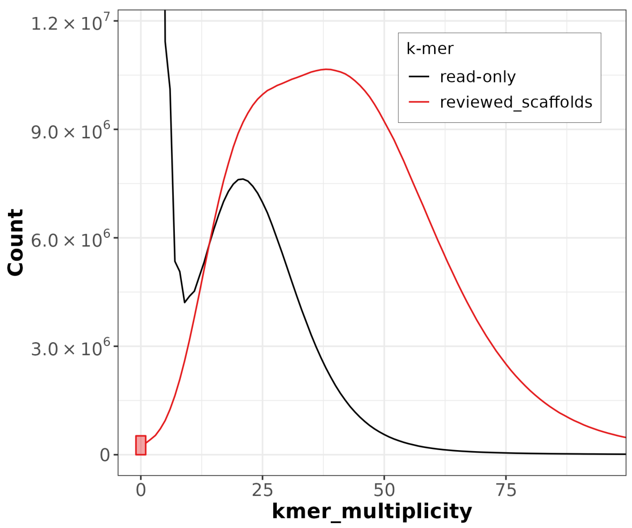


collapsed [LINK]

K-mer spectra of curated assembly

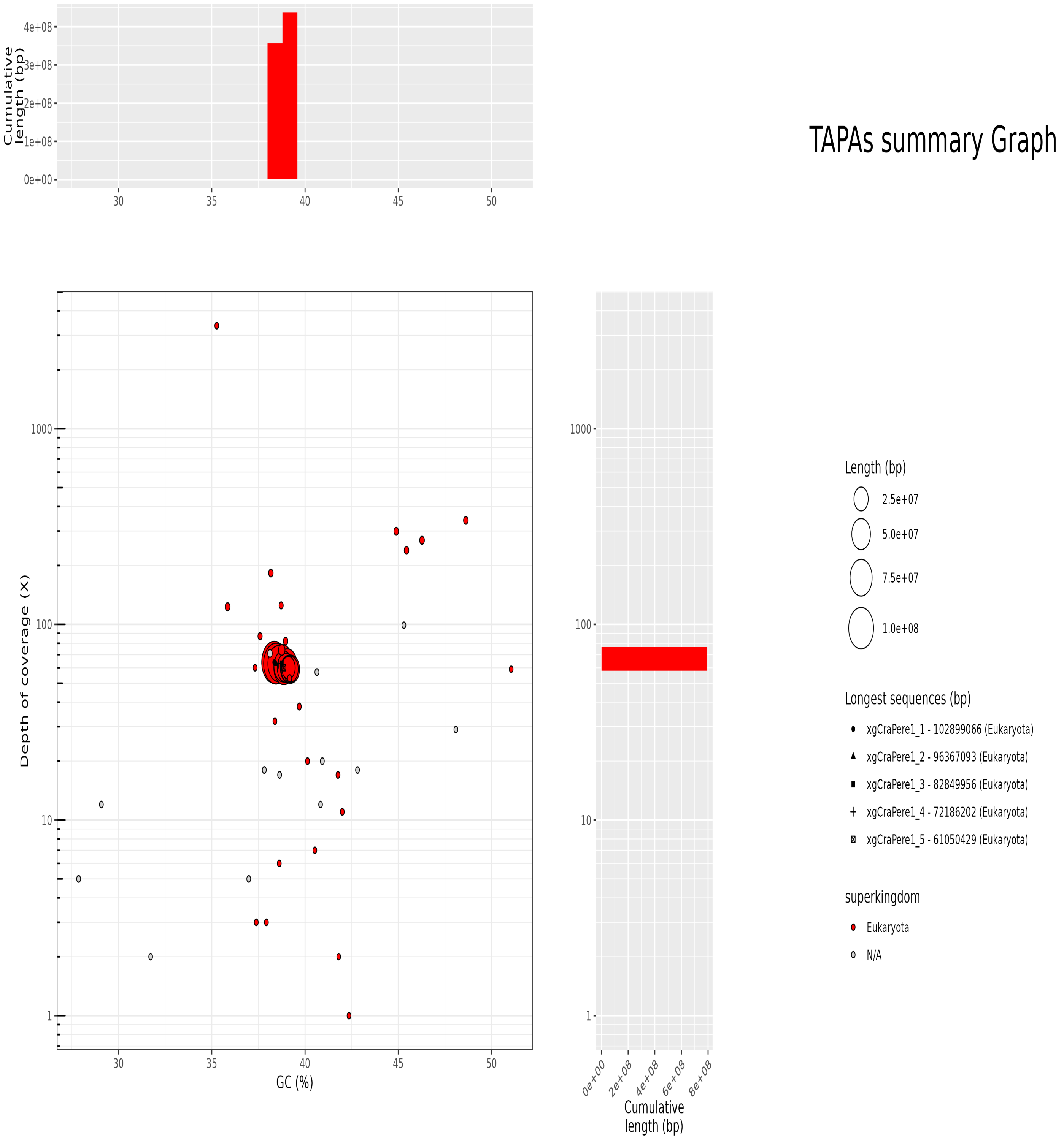


Distribution of k-mer counts per copy numbers found in asm



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	63	186

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: 2025-08-10 02:43:29 CEST