ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	470445
ToLID	**qmPorSegn1**
Species	Portunus segnis
Class	Malacostraca
Order	Decapoda

Genome Traits	Expected	Observed
Haploid size (bp)	1,321,034,932	984,116,969
Haploid Number	52 (source: ancestor)	52
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.7.Q39

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

. Observed Haploid size (bp) has >20% difference with Expected

. QV value is less than 40 for collapsed
. Kmer completeness value is less than 90 for collapsed
. More than 1000 gaps/Gbp for collapsed

Curator notes

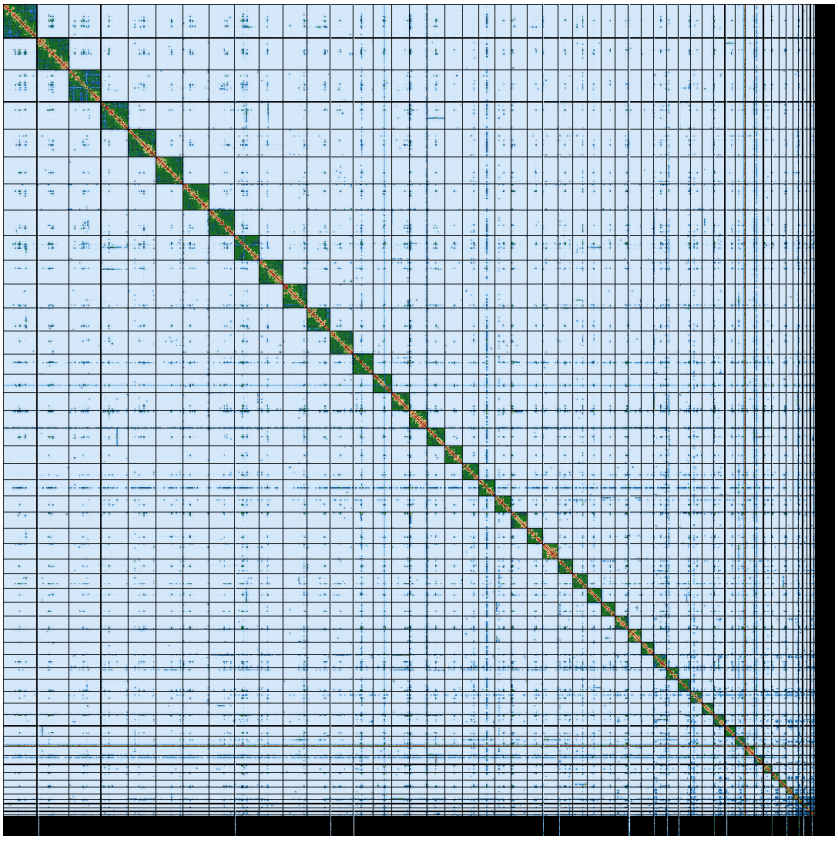
. Interventions/Gb: 300
. Contamination notes: ""
. Other observations: "The assembly of PORTUNUS SEGNIS (qmPorSegn1) is based on 35X ONT data and 133X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe project (BGE, https://biodiversitygenomics.eu/). The assembly process included the following steps: initial PacBio assembly generation with Flye, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 30 contigs were identified as viral, totaling 605,419 pb (with the largest being 245,493 pb) and 108 contigs were identified as bacterial, totaling 2,163,401 pb (with the largest being 537,362 pb). Additionally, 8 622 regions totaling 44 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, 79 contaminant sequences were removed, totaling 5.3 Mb (with the largest being 178,570 pb). 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	1,001,363,491	984,116,969
GC %	41.21	41.23
Gaps/Gbp	2,345.8	2,511.9
Total gap bp	234,900	267,300
Scaffolds	6,558	1,182
Scaffold N50	19,830,070	21,026,853
Scaffold L50	18	17
Scaffold L90	48	41
Contigs	8,907	3,654
Contig N50	925,711	938,765
Contig L50	276	267
Contig L90	1,450	1,323
QV	33.304	39.6733
Kmer compl.	89.3201	88.9701
BUSCO sing.	93.3%	93.3%
BUSCO dupl.	0.7%	0.7%
BUSCO frag.	3.6%	3.4%
BUSCO miss.	2.4%	2.6%

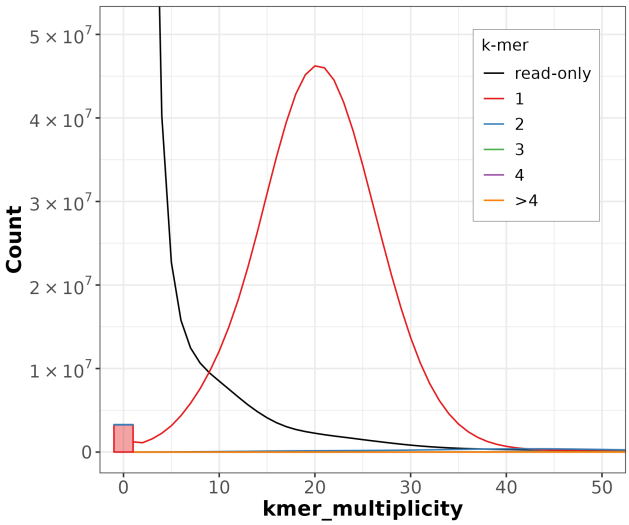
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: arthropoda_odb10 (genomes:90, BUSCOs:1013)

HiC contact map of curated assembly

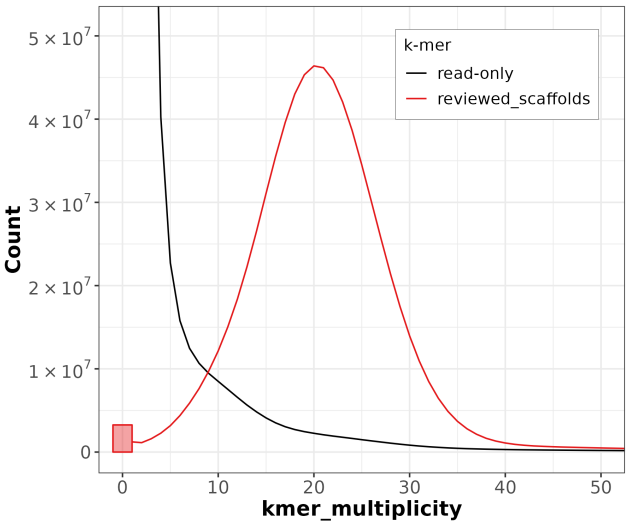


collapsed [LINK]

K-mer spectra of curated assembly

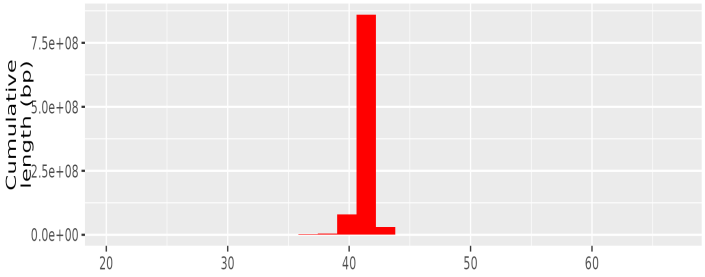


Distribution of k-mer counts per copy numbers found in asm

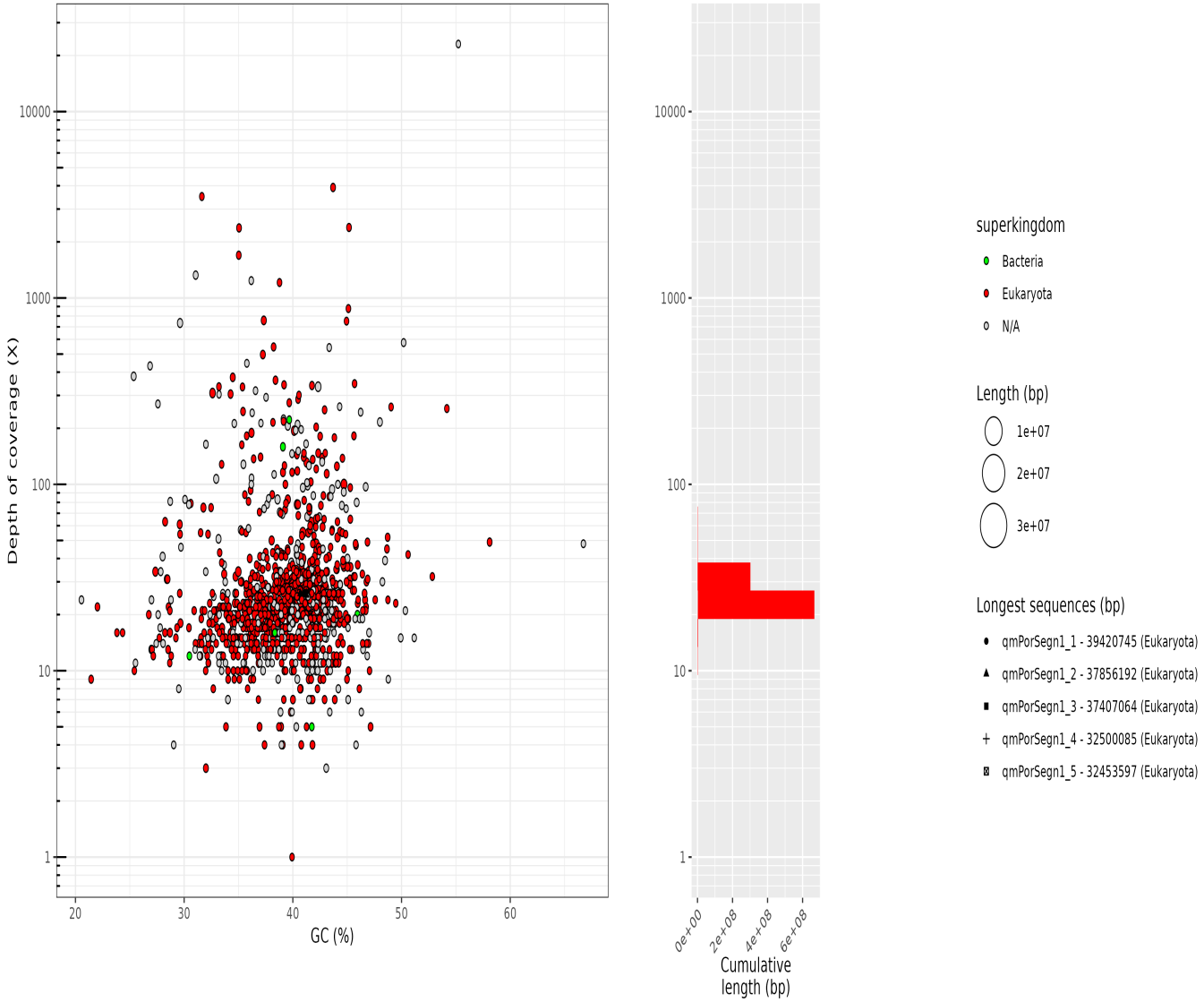


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	34	132

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: Lola Demirdjian
Affiliation: Genoscope

Date and time: 2025-04-05 17:28:49 CEST