

# ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	470445
ToLID	<b>qmPorSegn1.1</b>
Species	Portunus segnis
Class	Malacostraca
Order	Decapoda

Genome Traits	Expected	Observed
Haploid size (bp)	1,245,950,572	984,093,569
Haploid Number	52 (source: ancestor)	50
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

## EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.7.Q40

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

- . Observed Haploid size (bp) has >20% difference with Expected
- . Observed Haploid Number is different from Expected
- . Kmer completeness value is less than 90 for collapsed
- . More than 1000 gaps/Gbp for collapsed

### Curator notes

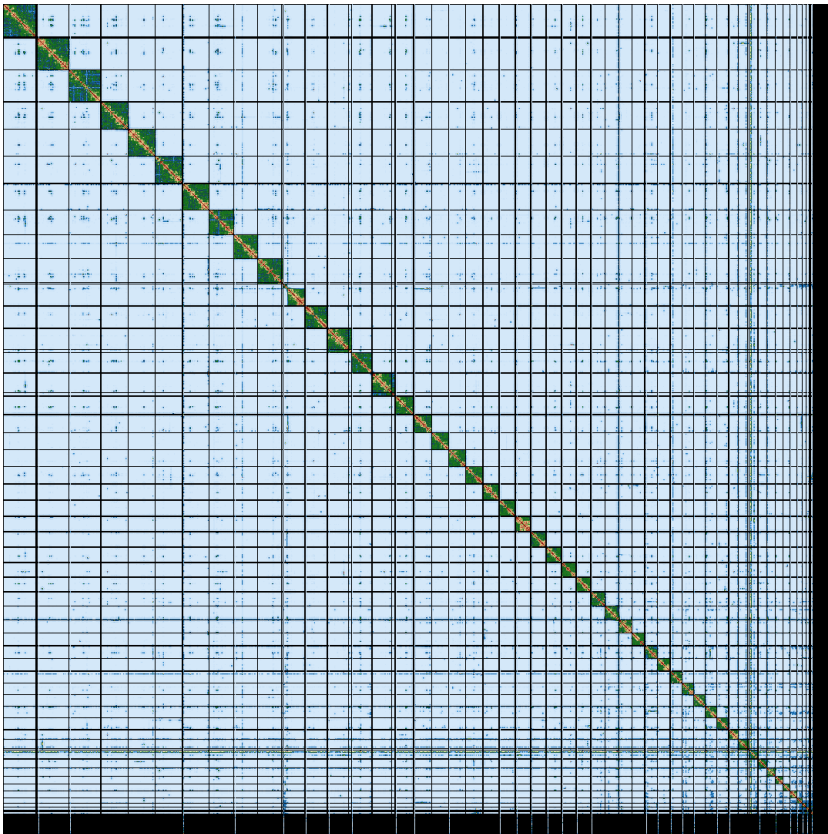
. Interventions/Gb: 280  
. Contamination notes: ""  
. Other observations: "The assembly of PORTUNUS SEGNIS (qmPorSegn1) is based on 35X ONT data and 910X Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, <https://www.erga-biodiversity.eu/>) via the Biodiversity Genomics 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	984,093,569
GC %	41.21	41.23
Gaps/Gbp	2,345.8	2,412.37
Total gap bp	234,900	243,900
Scaffolds	6,558	1,276
Scaffold N50	19,830,070	20,435,499
Scaffold L50	18	18
Scaffold L90	48	44
Contigs	8,907	3,650
Contig N50	925,711	939,202
Contig L50	276	267
Contig L90	1,450	1,320
QV	33.304	40.169
Kmer compl.	89.3201	88.5077
BUSCO sing.	93.3%	93.3%
BUSCO dupl.	0.7%	0.7%
BUSCO frag.	3.6%	3.4%
BUSCO miss.	2.4%	2.6%

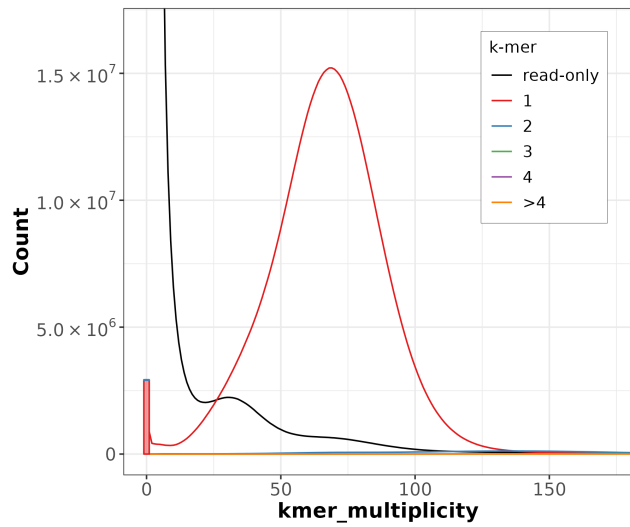
BUSCO: 5.4.3 (euk\_genome\_met, metaeuk) / Lineage: arthropoda\_odb10 (genomes:90, BUSCOs:1013)

# HiC contact map of curated assembly

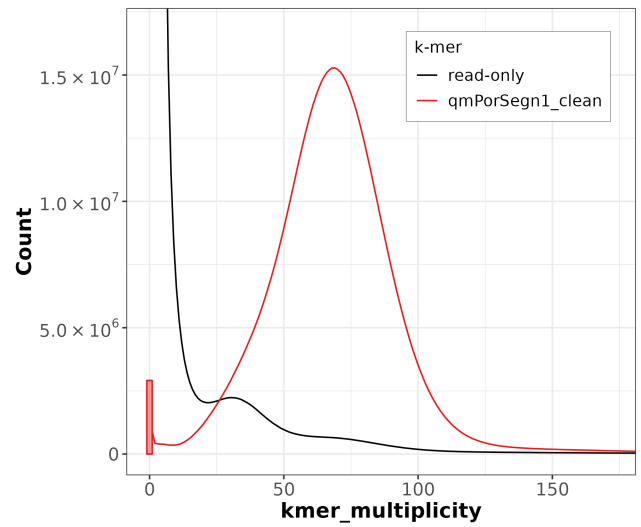


collapsed [\[LINK\]](#)

# K-mer spectra of curated assembly

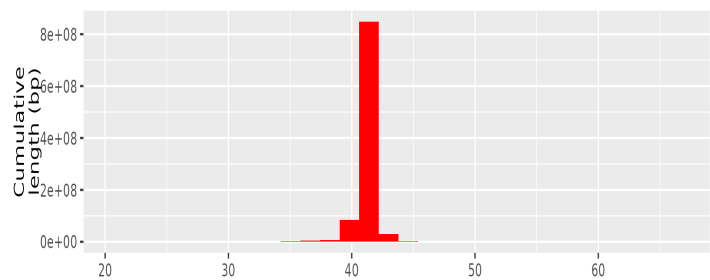


Distribution of k-mer counts per copy numbers found in asm

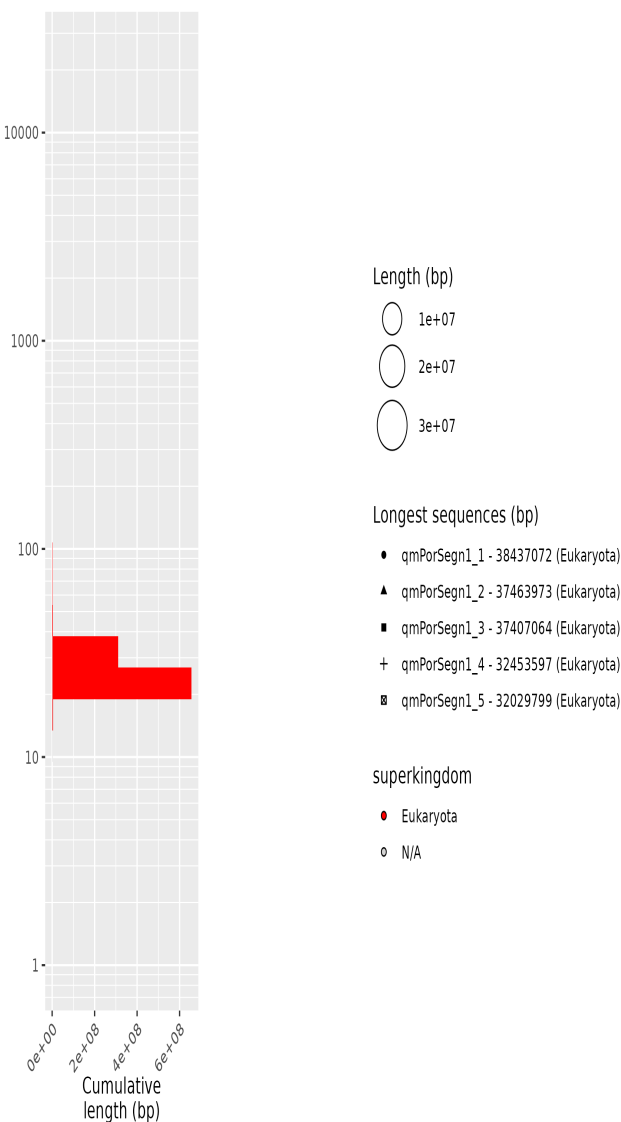
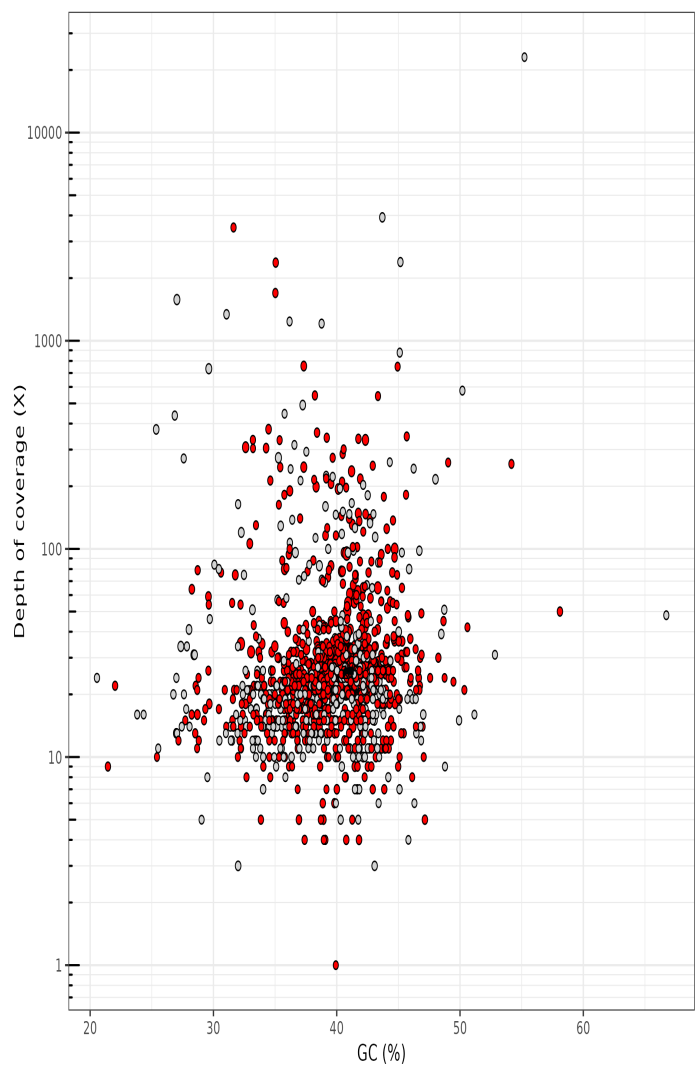


Distribution of k-mer counts coloured by their presence in reads/assemblies

# 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	ONT Hifi	Arima
Coverage	35	910

## Assembly pipeline

- **Flye**
  - |\_ *ver*: 2.9.2-b1786
  - |\_ *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: 2025-02-24 17:35:16 CET