ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1210415
ToLID	**wpGasClav1**
Species	Gastrolepidia clavigera
Class	Polychaeta
Order	Phyllodocida

Genome Traits	Expected	Observed
Haploid size (bp)	1,221,366,168	1,249,792,948
Haploid Number	10 (source: ancestor)	12
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q50

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

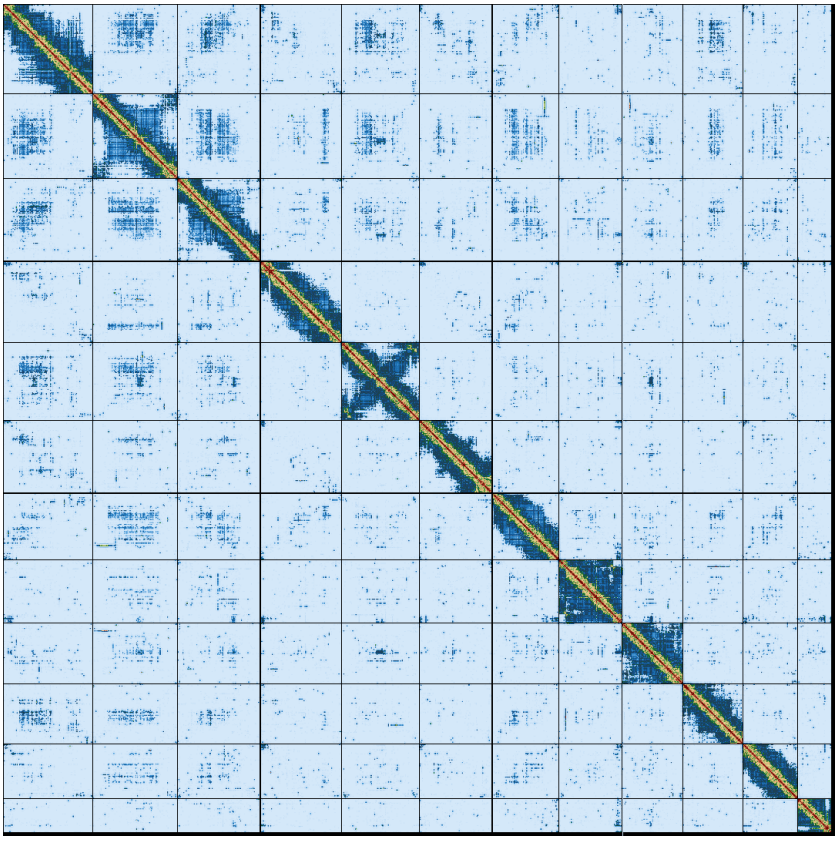
. Interventions/Gb: 116
. Contamination notes: ""
. Other observations: "The assembly of Gastrolepidia clavigera (wpGasClav1) is based on 63X PacBio data and 203X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial 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, 196 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 9.243 Mb (with the largest being 2.287 Mb). Additionally, 537 regions totaling 270.532 Mb (with the largest being 13.55 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, 41 haplotypic regions were removed, totaling 217Mb, (with the largest being 30Mb). 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	1,467,875,980	1,249,792,948
GC %	40.74	40.72
Gaps/Gbp	59.95	67.21
Total gap bp	8,800	12,000
Scaffolds	90	72
Scaffold N50	111,183,449	109,794,105
Scaffold L50	6	6
Scaffold L90	13	11
Contigs	178	156
Contig N50	20,883,306	22,540,000
Contig L50	24	20
Contig L90	68	55
QV	50.1531	50.2394
Kmer compl.	75.2587	66.8436
BUSCO sing.	75.2%	88.4%
BUSCO dupl.	15.0%	1.0%
BUSCO frag.	8.2%	8.4%
BUSCO miss.	1.6%	2.2%

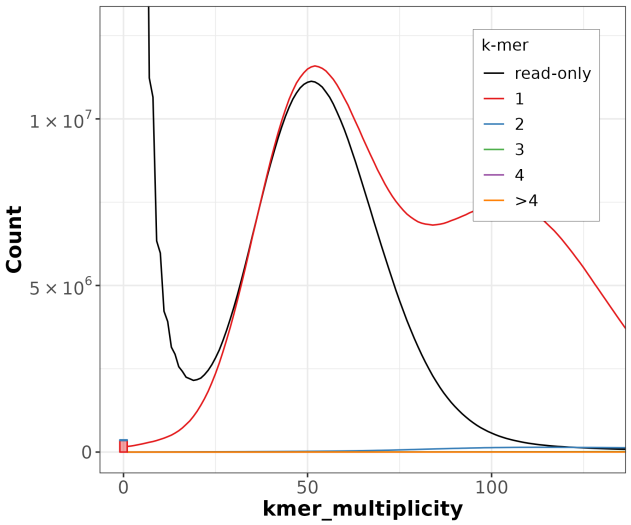
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: lophotrochozoa_odb12 (genomes:75, BUSCOs:1252)

HiC contact map of curated assembly

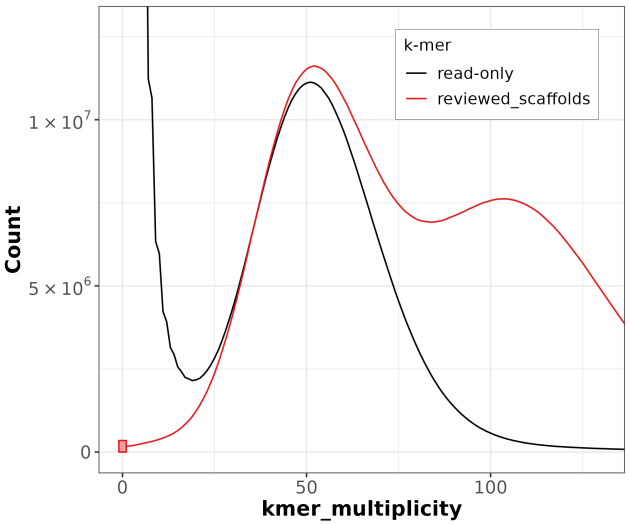


collapsed [LINK]

K-mer spectra of curated assembly

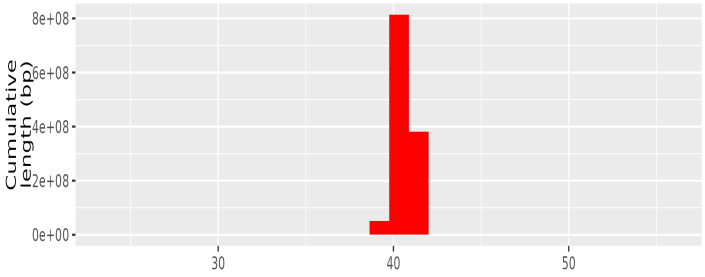


Distribution of k-mer counts per copy numbers found in asm

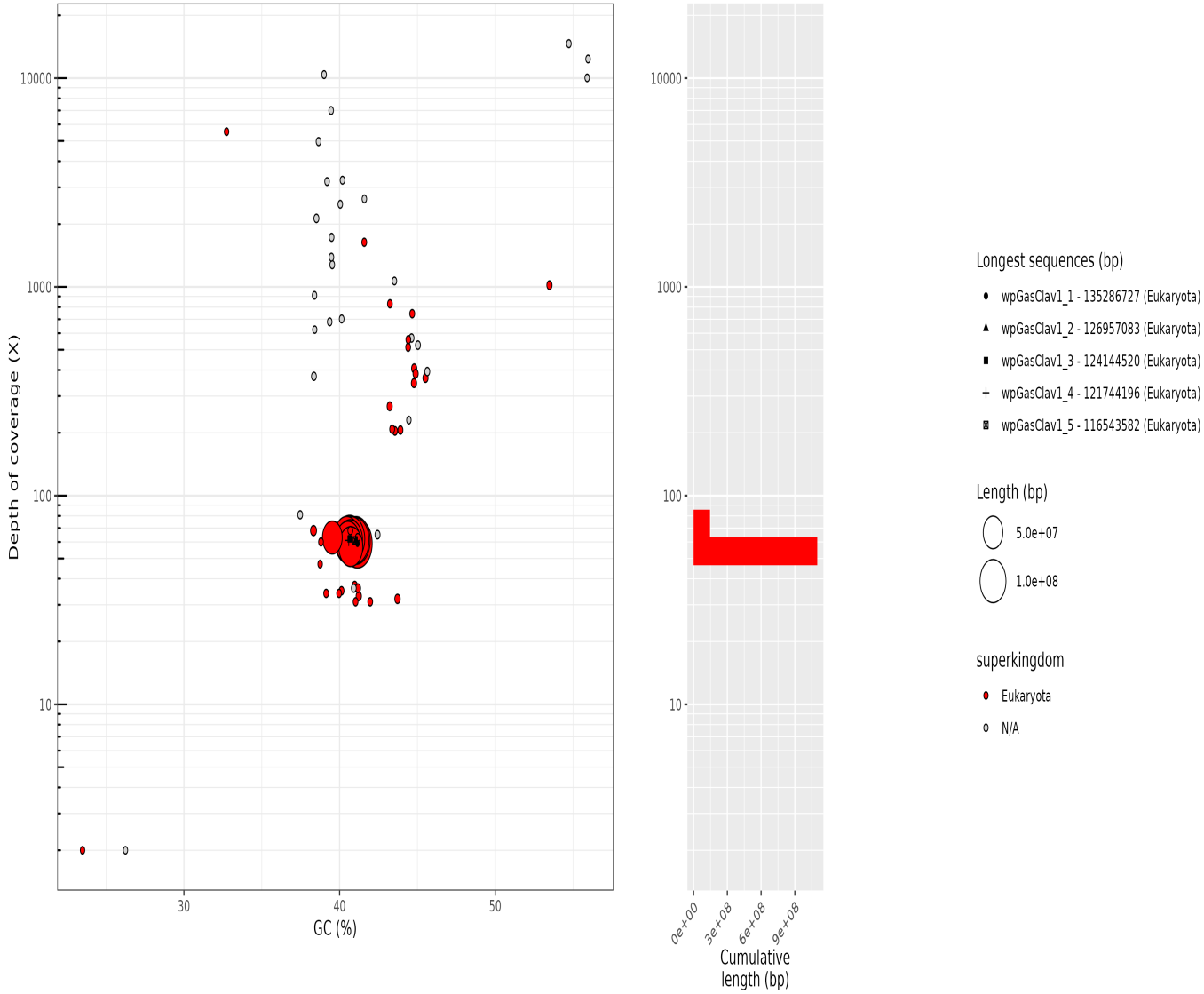


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	63	203

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: Arnaud Couloux
Affiliation: Genoscope

Date and time: 2025-07-24 00:06:41 CEST