ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1200667
ToLID	**xgDorVerr1**
Species	Doris verrucosa
Class	Gastropoda
Order	Nudibranchia

Genome Traits	Expected	Observed
Haploid size (bp)	981,646,056	1,223,411,306
Haploid Number	12 (source: ancestor)	13
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.8.Q42

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

Curator notes

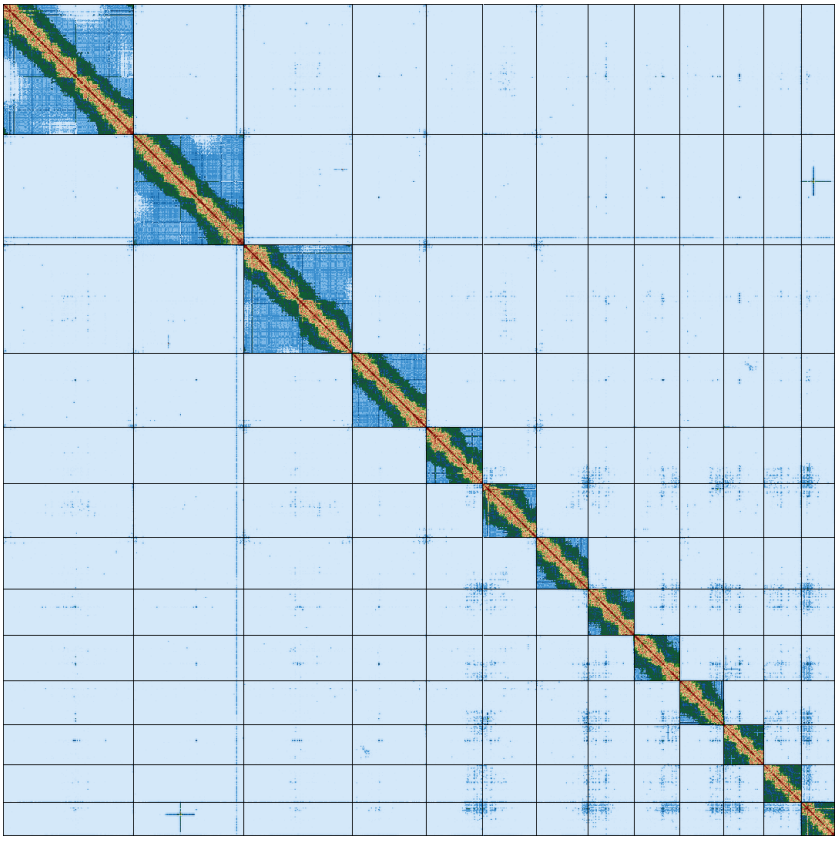
. Interventions/Gb: 60
. Contamination notes: ""
. Other observations: "The assembly of Doris verrucosa (xgDorVerr1) is based on 59X ONT data and 215X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial ONT assembly 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, 6 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 0.468 Mb (with the largest being 0.181 Mb). Additionally, 200 regions totaling 40.656 Mb (with the largest being 1.429 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, 200 haplotypic regions and 1 contaminant sequences were removed, totaling 40.656Mb and 0.023Mb, respectively (with the largest being 1.429Mb and 0.023Mb). 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	1,223,945,357	1,223,411,306
GC %	39.12	39.12
Gaps/Gbp	297.4	316.33
Total gap bp	36,400	43,300
Scaffolds	64	39
Scaffold N50	108,801,763	108,632,048
Scaffold L50	4	4
Scaffold L90	11	11
Contigs	428	426
Contig N50	5,756,812	5,756,812
Contig L50	60	60
Contig L90	213	213
QV	42.3613	42.3652
Kmer compl.	89.3881	89.376
BUSCO sing.	90.8%	97.7%
BUSCO dupl.	0.5%	0.9%
BUSCO frag.	6.4%	0.3%
BUSCO miss.	2.2%	1.1%

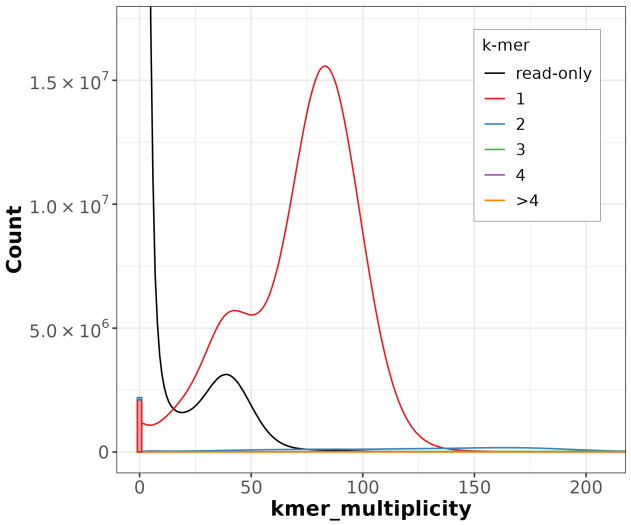
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: mollusca_odb12 (genomes:36, BUSCOs:4421)
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: mollusca_odb12 (genomes:36, BUSCOs:4421)

HiC contact map of curated assembly

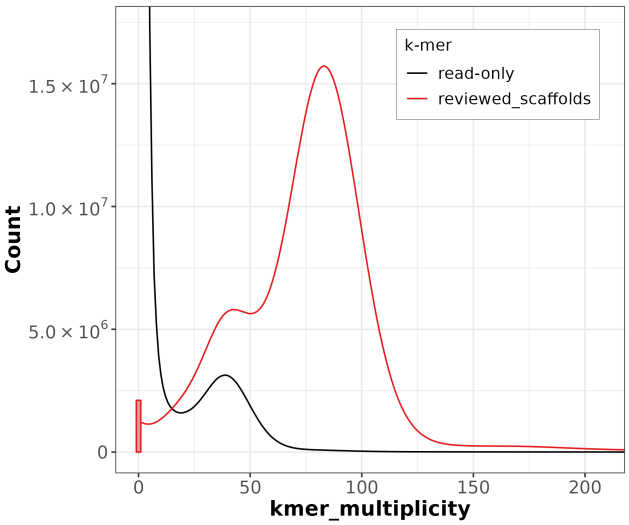


collapsed [LINK]

K-mer spectra of curated assembly

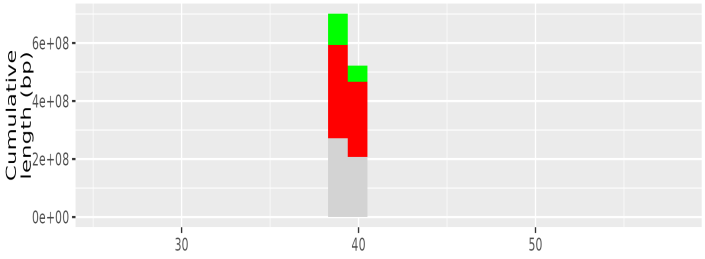


Distribution of k-mer counts per copy numbers found in asm

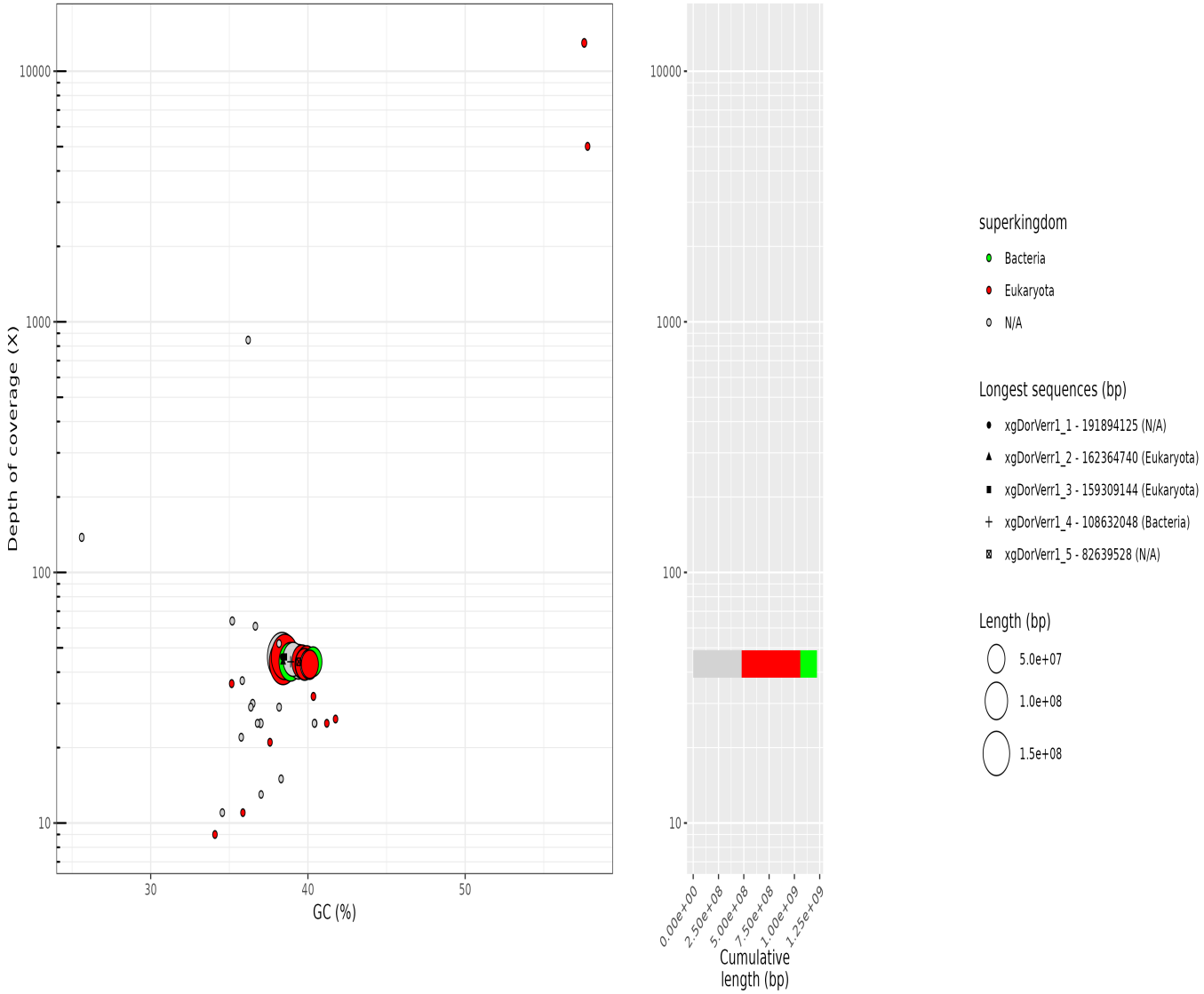


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



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	Long reads	Arima
Coverage	59	215

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: 2025-10-07 20:59:11 CEST