ERGA Assembly Report

v24.10.15

Tags: ERGA-Pilot

TxID	2040467
ToLID	**kaBotGaia**
Species	Botryllus gaiae
Class	Ascidiacea
Order	Stolidobranchia

Genome Traits	Expected	Observed
Haploid size (bp)	295,237,727	298,740,274
Haploid Number	16 (source: ancestor)	16
Ploidy	2 (source: ancestor)	2
Sample Sex	unknown	unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for pri: 6.7.Q35

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

. QV value is less than 40 for pri
. Kmer completeness value is less than 90 for pri

Curator notes

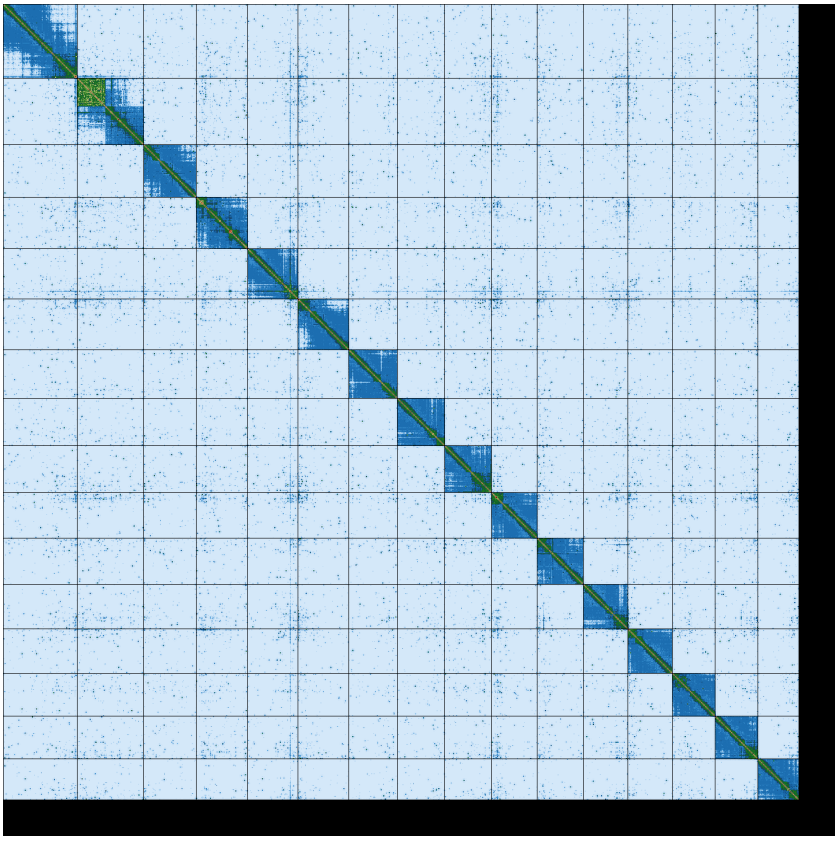
. Interventions/Gb: 96
. Contamination notes: "3 sequences were removed as contaminants."
. Other observations: "The assembly of species Botryllus gaiae (kaBotGaia) is based on 355X long read ONT data and 289X Arima HiC data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/). The assembly process included the following steps: ONT reads shorter than 3 kb were filtered out, thus the remaining reads -for a total of ~105 Gb- were assembled and polished using Necat. Contigs were subsequently filtered with purge_dups to remove haplotigs and obtain a more accurate, non-redundant assembly. Scaffolding was performed by aligning HiC reads to the purged contigs using the Omni-C mapping pipeline, followed by YaHS for scaffold construction. Contamination was checked using BlobTools. The scaffolds were processed through sanger_tol/curationpretext pipeline to generate the contact map, which was manually improved using PretextView. The presence of a mitochondrial genome contig/scaffold was not assessed. Chromosome-scale scaffolds confirmed by HiC data were named in order of size."

Quality metrics table

Metrics	Pre-curation pri	Curated pri
Total bp	304,686,464	298,740,274
GC %	41.17	41.18
Gaps/Gbp	626.87	826.81
Total gap bp	19,100	32,700
Scaffolds	432	377
Scaffold N50	16,538,291	17,039,967
Scaffold L50	8	8
Scaffold L90	16	15
Contigs	623	624
Contig N50	2,134,286	2,024,695
Contig L50	41	40
Contig L90	151	150
QV	35.6615	35.6654
Kmer compl.	69.5153	68.6206
BUSCO sing.	90.9%	91.5%
BUSCO dupl.	0.9%	0.9%
BUSCO frag.	1.9%	1.5%
BUSCO miss.	6.3%	6.1%

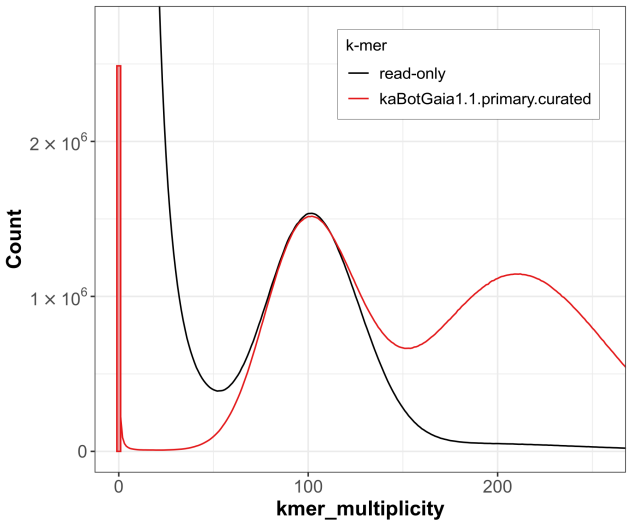
BUSCO: 5.8.2 (euk_genome_aug, augustus) / Lineage: metazoa_odb10 (genomes:65, BUSCOs:954)

HiC contact map of curated assembly

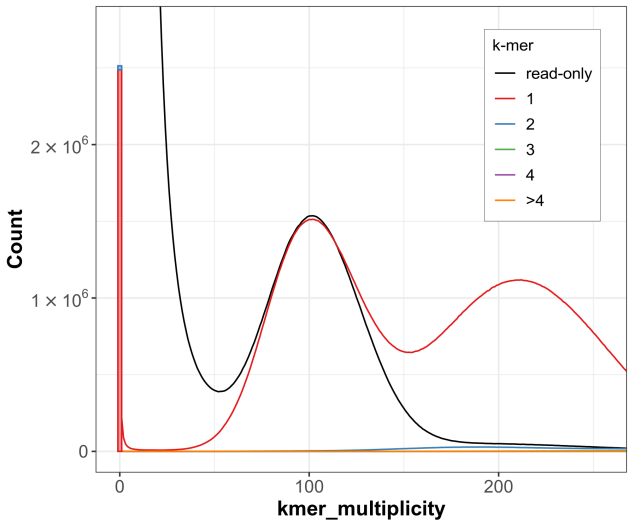


pri [LINK]

K-mer spectra of curated assembly

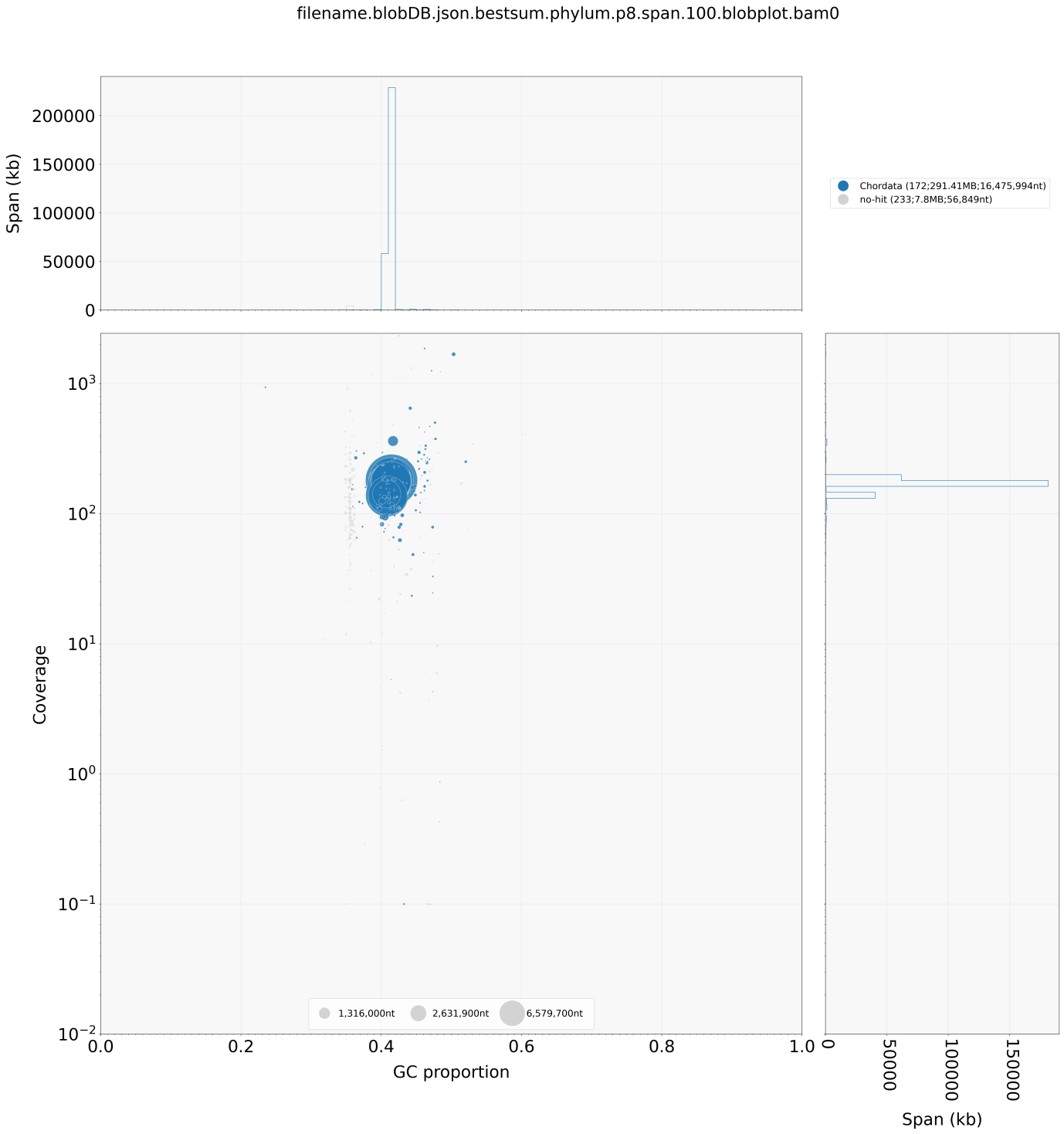


Distribution of k-mer counts coloured by their presence in reads/assemblies



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

Post-curation contamination screening



pri. 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	Arima HiC
Coverage	355x	289x

Assembly pipeline

- **Necat**
  |_ *ver:* 0.0.1
  |_ *key param:* NA
- **purge_dups**
  |_ *ver:* 1.2.5
  |_ *key param:* NA
- **YaHS**
  |_ *ver:* 1.2.2
  |_ *key param:* NA

Curation pipeline

- **sanger-tol/curationpretext**
  |_ *ver:* 1.5.0
  |_ *key param:* NA
- **PretextView**
  |_ *ver:* 1.0.3
  |_ *key param:* NA

Submitter: Ilenia Urso
Affiliation: UNIBA

Date and time: 2025-12-01 16:11:40 CET