

ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	521046
ToLID	kaMicPoly
Species	Microcosmus polymorphus
Class	Ascidacea
Order	Stolidobranchia

Genome Traits	Expected	Observed
Haploid size (bp)	323,716,733	320,678,343
Haploid Number	16 (source: ancestor)	12
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:

- . Observed Haploid Number is different from Expected
- . QV value is less than 40 for pri
- . Kmer completeness value is less than 90 for pri
- . BUSCO single copy value is less than 90% for pri
- . Assembly length loss > 3% for pri
- . More than 1000 gaps/Gbp for pri

Curator notes

- . Interventions/Gb: 233
- . Contamination notes: "28 sequences were removed as contaminants. In addition, one large scaffold (SUPER_12) was classified as Arthropoda. However, it was not removed because its coverage and GC content were consistent with the rest of the genome, and it showed no Hi-C contacts with other scaffolds."
- . Other observations: "The assembly of species *Microcosmus polymorphus* is based on 352X long read ONT data and 211X Arima HiC data generated as part of the ERGA initiative via the BGE project. The assembly process included the following steps: ONT reads shorter than 1 kb were filtered out, thus the remaining reads for a total of ~121 Gb were assembled using Necat with polishing performed using NextPolish. Contigs were subsequently filtered with purge_dups. 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,

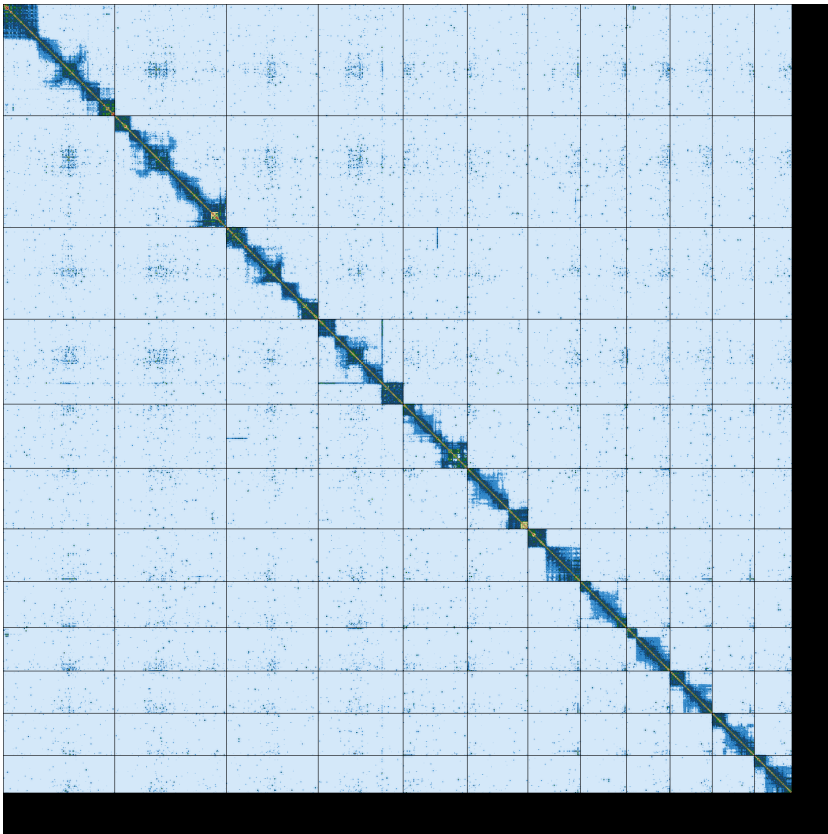
which was manually improved using PretextView. The presence of a mitochondrial genome contig/scaffold was not assessed. Chromosome-scale scaffolds confirmed by HiC data were named in order of size."

Quality metrics table

Metrics	Pre-curation pri	Curated pri
Total bp	369,422,115	320,678,343
GC %	43.04	43.08
Gaps/Gbp	1,580.85	1,836.73
Total gap bp	58,400	70,800
Scaffolds	1,150	1,052
Scaffold N50	19,963,143	24,703,516
Scaffold L50	6	5
Scaffold L90	28	11
Contigs	1,734	1,641
Contig N50	1,029,542	1,069,128
Contig L50	97	84
Contig L90	420	361
QV	36.019	35.5155
Kmer compl.	72.238	65.3982
BUSCO sing.	85.4%	89.1%
BUSCO dupl.	7.0%	2.5%
BUSCO frag.	2.2%	1.9%
BUSCO miss.	5.3%	6.5%

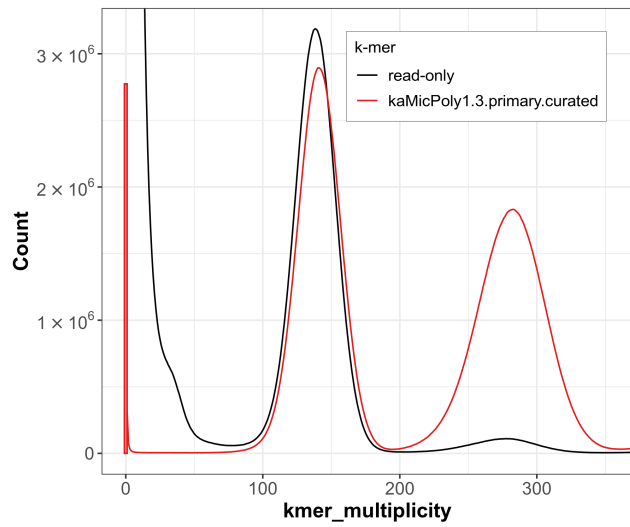
BUSCO: 5.8.2 (euk_genome_aug, augustus) / Lineage: metazoa_odb10 (genomes:65, BUSCOs:954)

HiC contact map of curated assembly

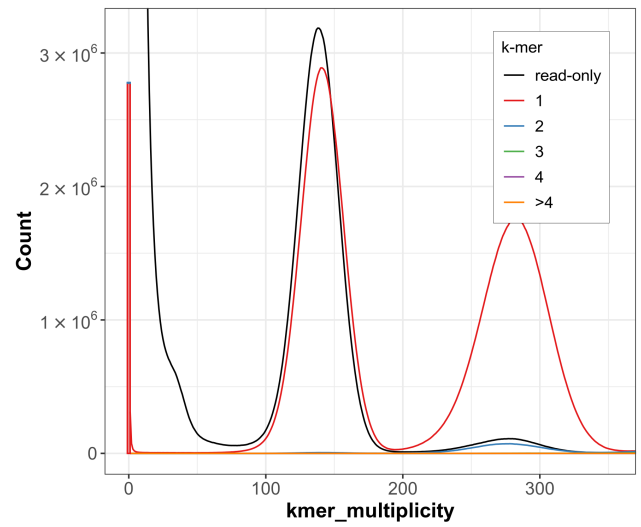


pri [\[LINK\]](#)

K-mer spectra of curated assembly



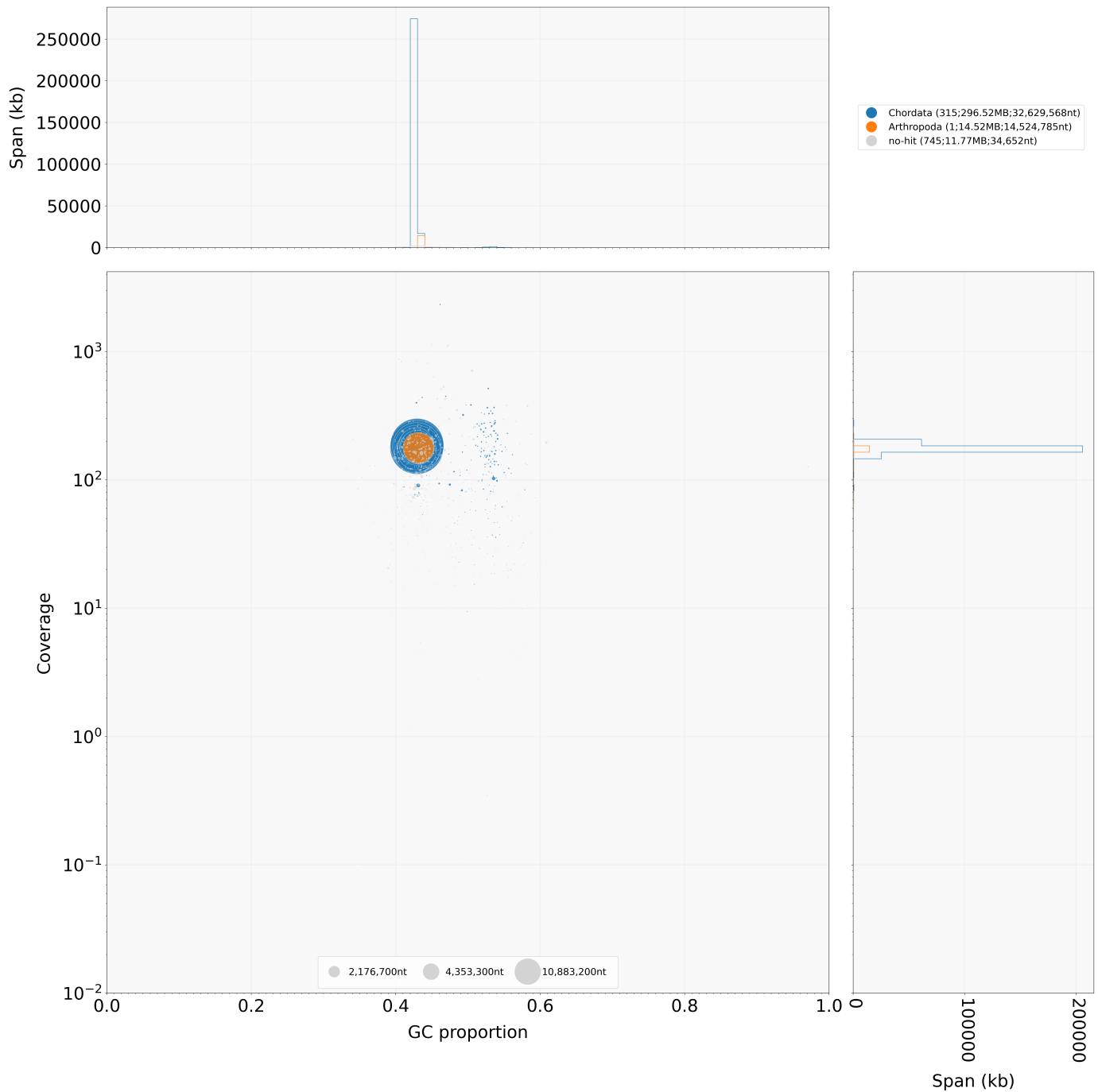
Distribution of k-mer counts coloured by their presence in reads/assemblies



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

Post-curation contamination screening

filename.blobDB.json.bestsum.phylum.p8.span.100.blobplot.bam0



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	352x	211x

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*: 1.0.5
 - |_ *key param*: NA

Submitter: Ilenia Urso

Affiliation: UNIBA

Date and time: 2025-12-18 09:27:33 CET