ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	76338
ToLID	**fThaBif2**
Species	Thalassoma bifasciatum
Class	Actinopteri
Order	Labriformes

Genome Traits	Expected	Observed
Haploid size (bp)	761,954,446	779,196,266
Haploid Number	24 (source: ancestor)	24
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q49

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

. Kmer completeness value is less than 90 for collapsed
. Assembly length loss > 3% for collapsed

Curator notes

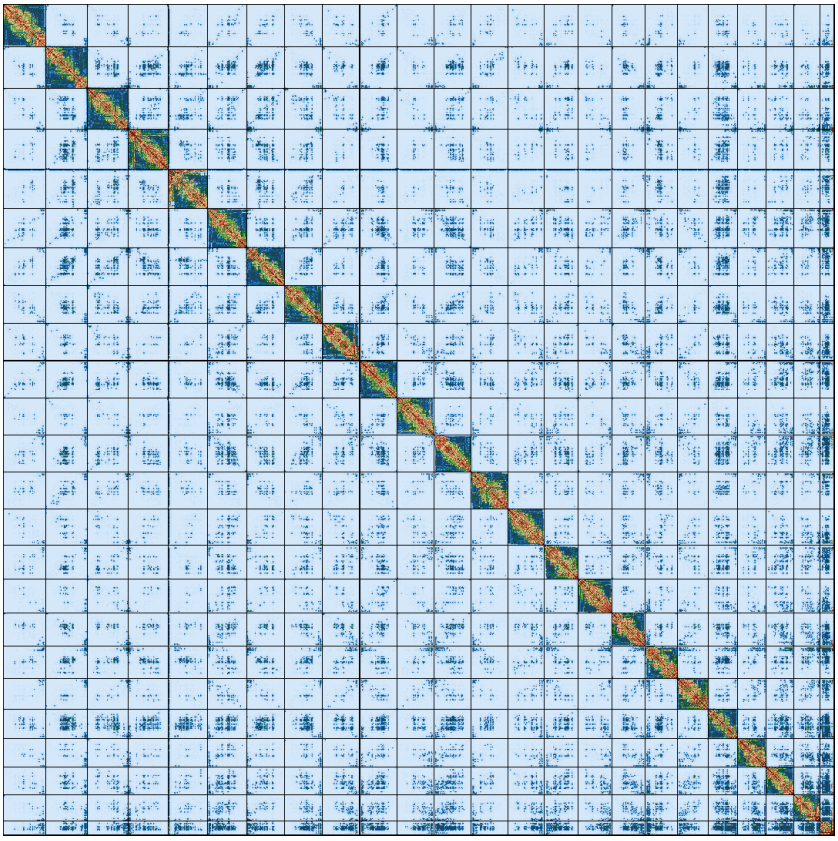
. Interventions/Gb: 15
. Contamination notes: ""
. Other observations: "The assembly of Thalassoma bifasciatum (fThaBif2) is based on 38X PacBio data and 157X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial 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, 6 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 0.152 Mb (with the largest being 0.045 Mb). Additionally, 264 regions totaling 20.709 Mb (with the largest being 1.819 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, 705 contaminant sequences were removed, totaling 32.25Mb (with the largest being 0.145Mb). 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	811,442,005	779,196,266
GC %	39.38	39.74
Gaps/Gbp	101.05	100.1
Total gap bp	8,800	9,500
Scaffolds	760	41
Scaffold N50	34,322,671	34,643,621
Scaffold L50	12	11
Scaffold L90	22	21
Contigs	828	119
Contig N50	30,888,556	30,888,556
Contig L50	12	12
Contig L90	27	25
QV	41.4907	49.0572
Kmer compl.	75.7037	75.5652
BUSCO sing.	97.1%	99.5%
BUSCO dupl.	0.2%	0.1%
BUSCO frag.	0.6%	0.1%
BUSCO miss.	2.0%	0.4%

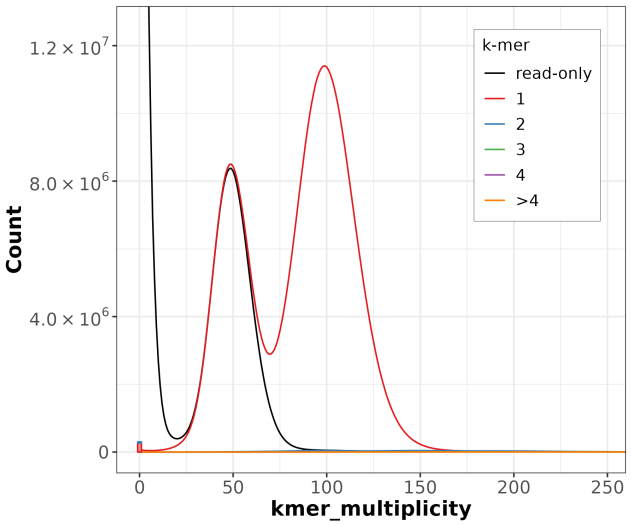
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: actinopterygii_odb12 (genomes:75, BUSCOs:7207)
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: actinopterygii_odb12 (genomes:75, BUSCOs:7207)

HiC contact map of curated assembly

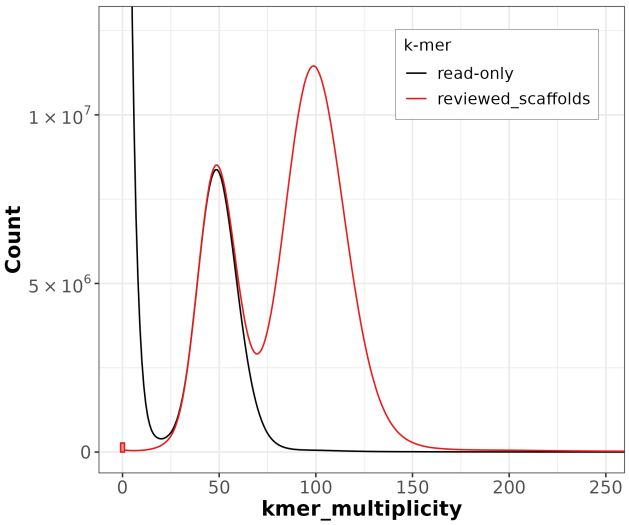


collapsed [LINK]

K-mer spectra of curated assembly

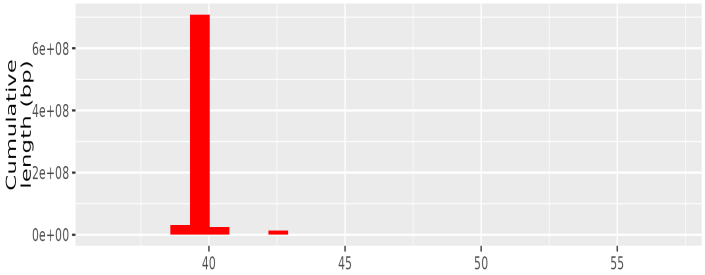


Distribution of k-mer counts per copy numbers found in asm

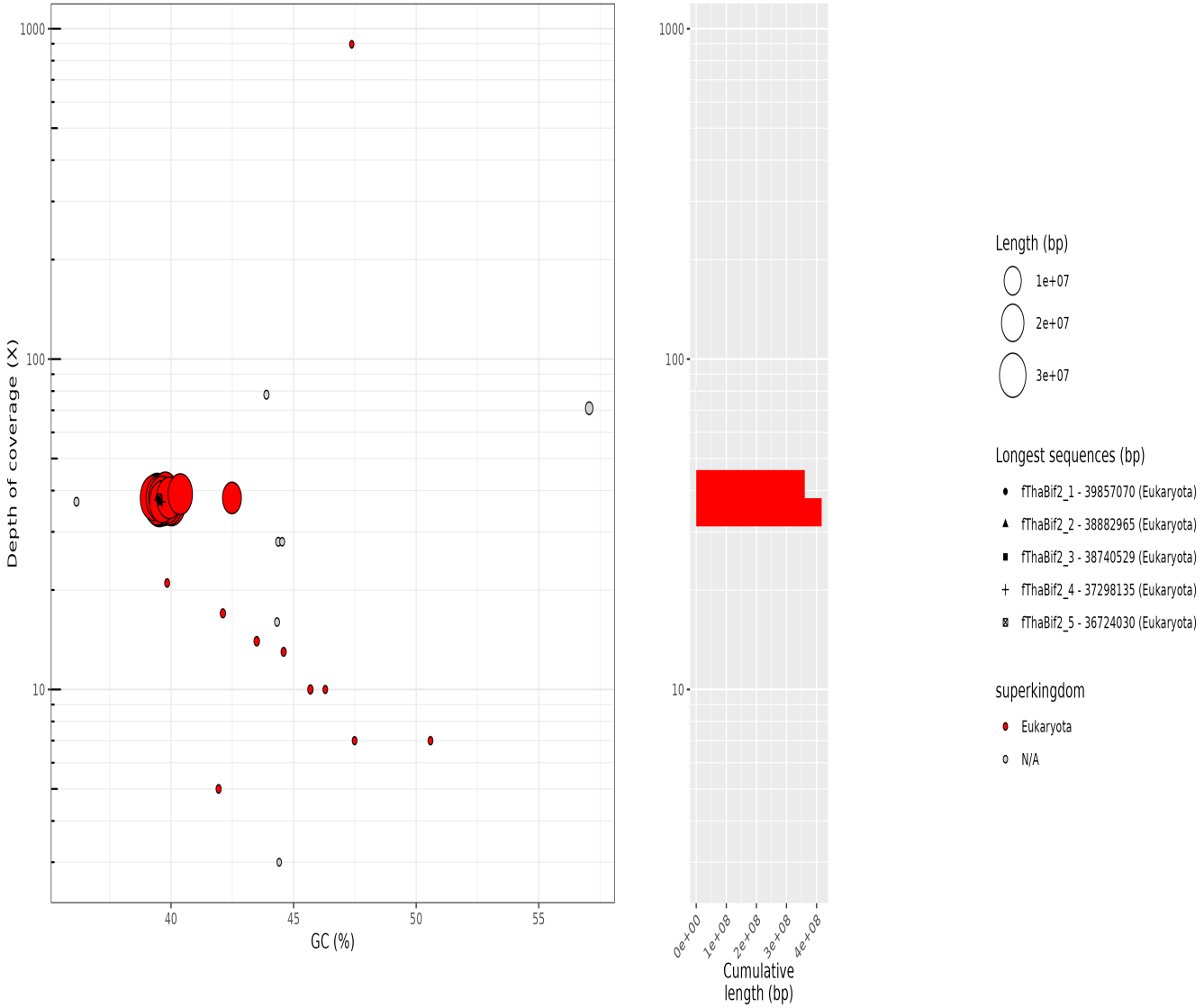


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	38	157

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: Jean-Marc Aury
Affiliation: Genoscope

Date and time: 2025-10-13 07:10:41 CEST