ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	257540
ToLID	**fGobFla1**
Species	Gobiusculus flavescens
Class	Actinopteri
Order	Gobiiformes

Genome Traits	Expected	Observed
Haploid size (bp)	769,472,362	836,700,789
Haploid Number	19 (source: ancestor)	23
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q47

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

Curator notes

. Interventions/Gb: 162
. Contamination notes: ""
. Other observations: "The assembly of Gobiusculus flavescens (fGobFla1.1) is based on 58X PacBio data and 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, 5 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 0.75 Mb (with the largest being 0.58 Mb). Additionally, 407 regions totaling 30.48 Mb (with the largest being 1.8 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, 7 haplotypic regions were removed, totaling 4.11 Mb (with the largest being 1.92 Mb). Subtelomeric regions were more fragmented, and the organization of contigs in these regions has lower confidence. Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. Scaffold_187 was initially tagged as bacterial by context software. However, Hi-C contact data and further alignments confirm that this contig is actually from the

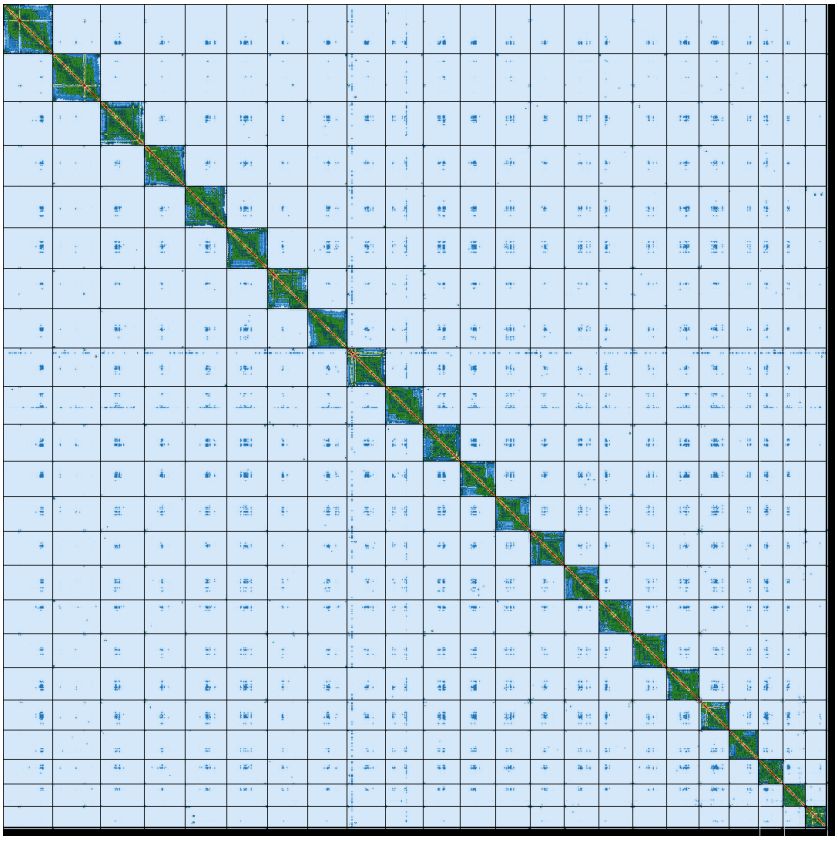
host genome. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	840,870,892	836,700,789
GC %	42.05	42.05
Gaps/Gbp	391.26	500.78
Total gap bp	32,900	52,900
Scaffolds	258	146
Scaffold N50	35,118,790	37,920,148
Scaffold L50	11	10
Scaffold L90	22	20
Contigs	587	565
Contig N50	17,987,913	17,987,913
Contig L50	18	18
Contig L90	138	137
QV	47.1537	47.1648
Kmer compl.	82.8661	82.7538
BUSCO sing.	91.0%	91.5%
BUSCO dupl.	0.9%	0.4%
BUSCO frag.	2.5%	2.4%
BUSCO miss.	5.6%	5.6%

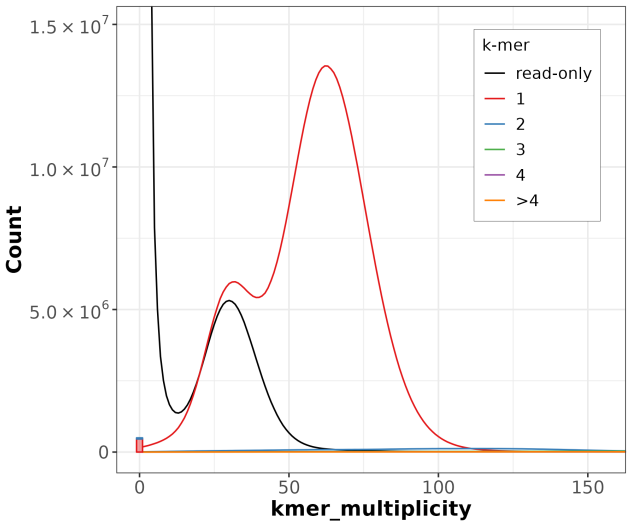
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: actinopterygii_odb12 (genomes:75, BUSCOs:7207)

HiC contact map of curated assembly

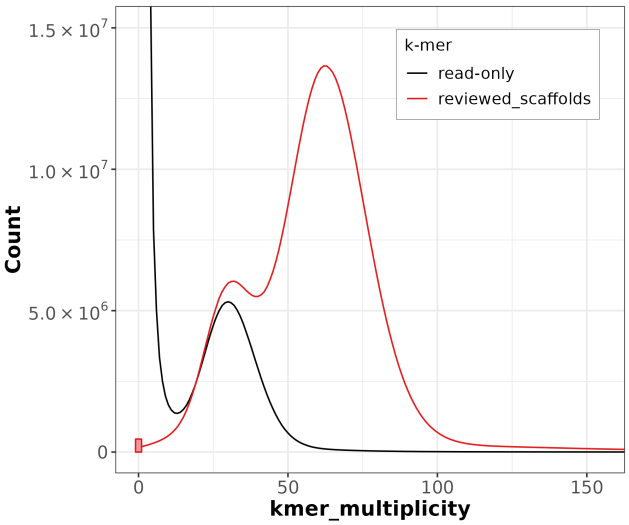


collapsed [LINK]

K-mer spectra of curated assembly

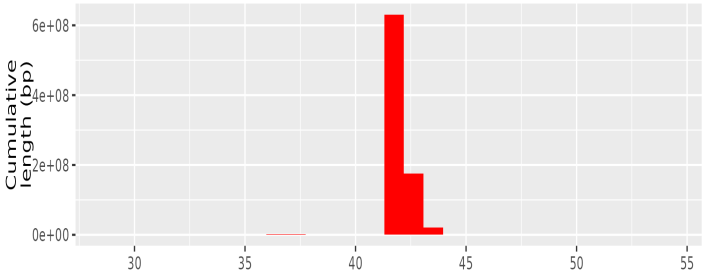


Distribution of k-mer counts per copy numbers found in asm

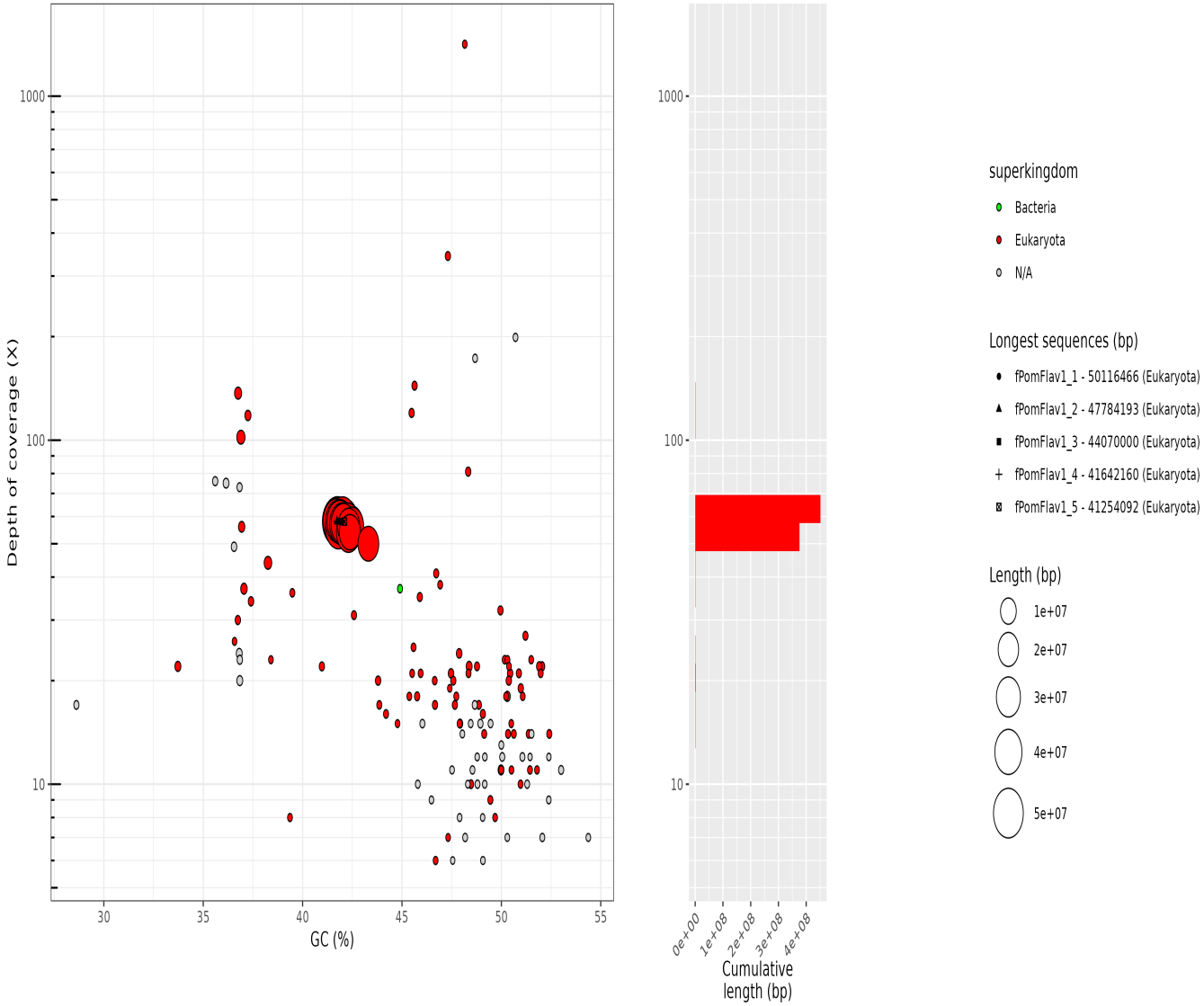


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	PACBIO Hifi	Arima
Coverage	58	188

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: 2025-05-09 01:42:15 CEST