ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	203316
ToLID	**fGobXan1**
Species	Gobius xanthocephalus
Class	Actinopteri
Order	Gobiiformes

Genome Traits	Expected	Observed
Haploid size (bp)	821,974,802	859,247,006
Haploid Number	21 (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.Q49

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
. Not 90% of assembly in chromosomes for collapsed

Curator notes

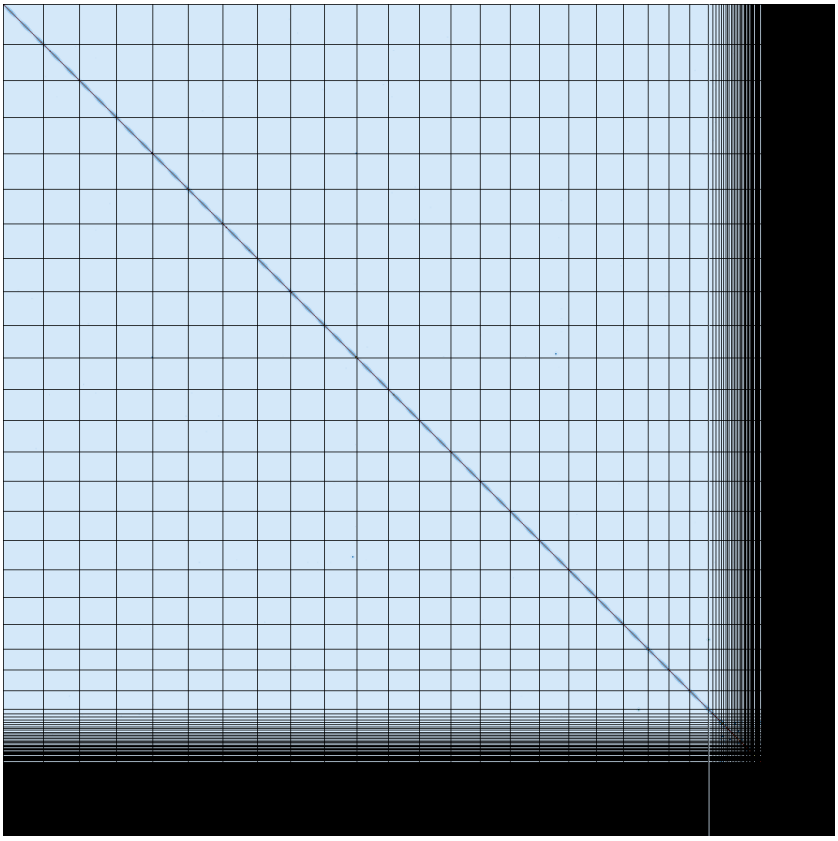
. Interventions/Gb: 5
. Contamination notes: ""
. Other observations: "The assembly of Gobius xanthocephalus (fGobXan1) is based on 71X PacBio data and 124X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial 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, 4 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 0.494 Mb (with the largest being 0.349 Mb). Additionally, 301 regions totaling 24.909 Mb (with the largest being 2.329 Mb) were identified as haplotypic duplications and removed. Mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 2 haplotypic regions and 13 contaminant sequences were removed, totaling 1.055Mb and 0.371Mb, respectively (with the largest being 0.814Mb and 0.047Mb). 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	860,697,163	859,247,006
GC %	41.94	41.94
Gaps/Gbp	152.2	150.13
Total gap bp	13,100	13,000
Scaffolds	1,033	1,020
Scaffold N50	32,203,740	32,203,740
Scaffold L50	11	12
Scaffold L90	48	49
Contigs	1,164	1,149
Contig N50	30,452,000	30,452,000
Contig L50	13	13
Contig L90	90	89
QV	49.1434	49.1538
Kmer compl.	76.2471	76.2254
BUSCO sing.	96.4%	96.8%
BUSCO dupl.	1.2%	0.8%
BUSCO frag.	0.7%	0.7%
BUSCO miss.	1.7%	1.7%

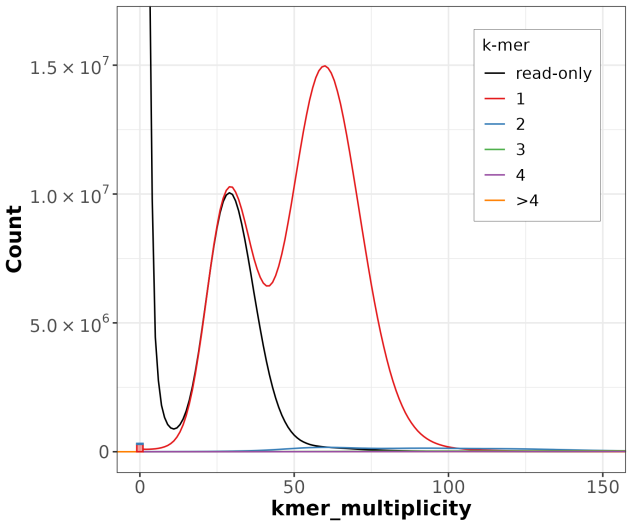
Warning! BUSCO versions or lineage datasets are not the same across results:
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: actinopterygii_odb12 (genomes:75, BUSCOs:7207)
BUSCO: 6.1.0 (euk_genome_min, miniprot) / Lineage: actinopterygii_odb12 (genomes:75, BUSCOs:7207)

HiC contact map of curated assembly

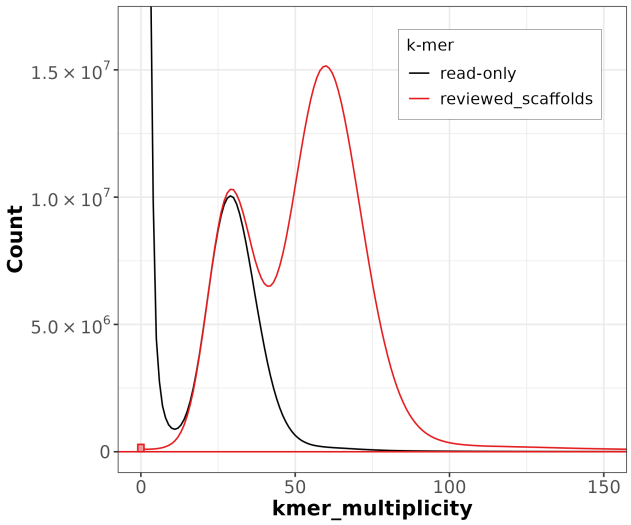


collapsed [LINK]

K-mer spectra of curated assembly

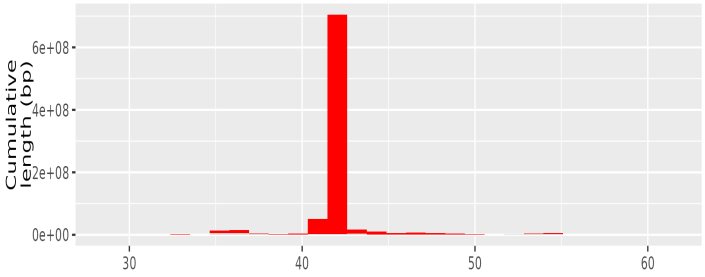


Distribution of k-mer counts per copy numbers found in asm

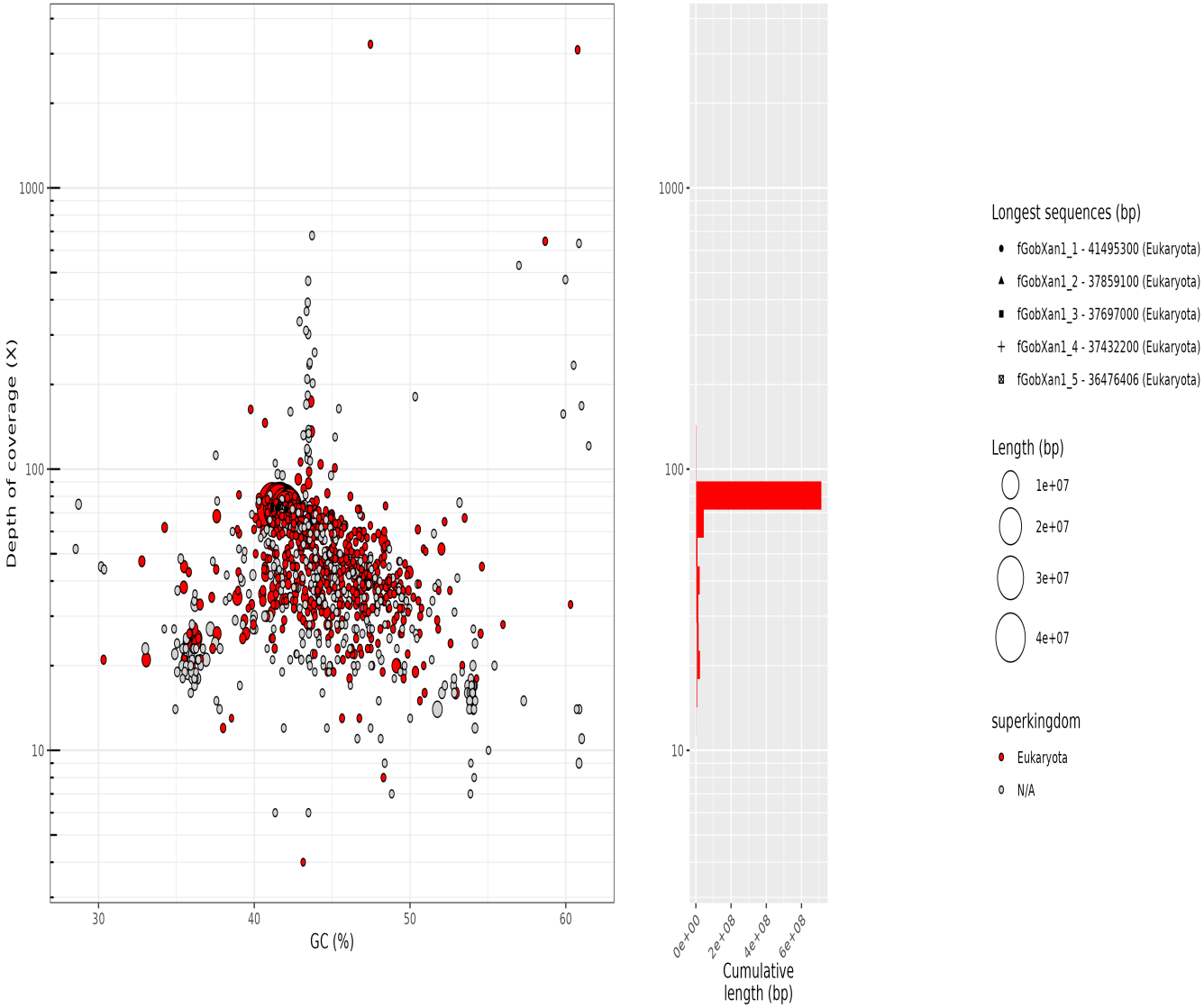


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	71	124

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: 2026-06-17 14:31:15 CEST