

ERGA Assembly Report

v24.10.15

Tags: ATLASEa[INVALID TAG]

TxID	584627
ToLID	fGymMil1
Species	Gymnothorax miliaris
Class	Actinopteri
Order	Anguilliformes

Genome Traits	Expected	Observed
Haploid size (bp)	2,072,788,673	2,054,988,274
Haploid Number	21 (source: ancestor)	21
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q48

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

- . Kmer completeness value is less than 90 for collapsed
- . BUSCO duplicated value is more than 5% for collapsed

Curator notes

. Interventions/Gb: 11
. Contamination notes: ""
. Other observations: "The assembly of Gymnothorax miliaris (fGymMil1) is based on 30X PacBio data and 128X 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. No contigs were found to be contaminants (bacterial, archaeal, or viral). Additionally, 229 regions totaling 74.488 Mb (with the largest being 3.925 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, 1 haplotypic region and 2 contaminant sequences were removed, totaling 2.162Mb and 0.332Mb, respectively (with the largest being 2.162Mb and 0.239Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

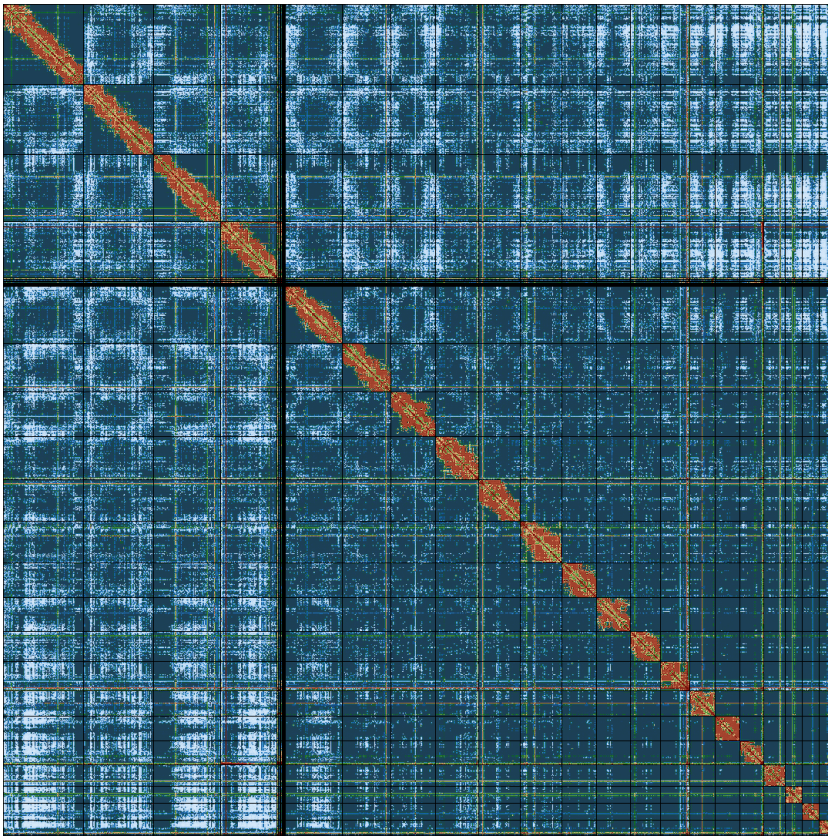
Metrics	Pre-curation collapsed	Curated collapsed
Total bp	2,057,463,523	2,054,988,274
GC %	41.72	41.72
Gaps/Gbp	31.59	36.98
Total gap bp	6,500	9,100
Scaffolds	53	43
Scaffold N50	109,888,096	109,951,085
Scaffold L50	7	7
Scaffold L90	18	17
Contigs	118	119
Contig N50	52,233,601	52,751,958
Contig L50	13	12
Contig L90	43	43
QV	48.9417	48.9464
Kmer compl.	82.4092	82.389
BUSCO sing.	91.5%	90.1%
BUSCO dupl.	6.8%	6.6%
BUSCO frag.	0.2%	1.4%
BUSCO miss.	1.4%	2.0%

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)

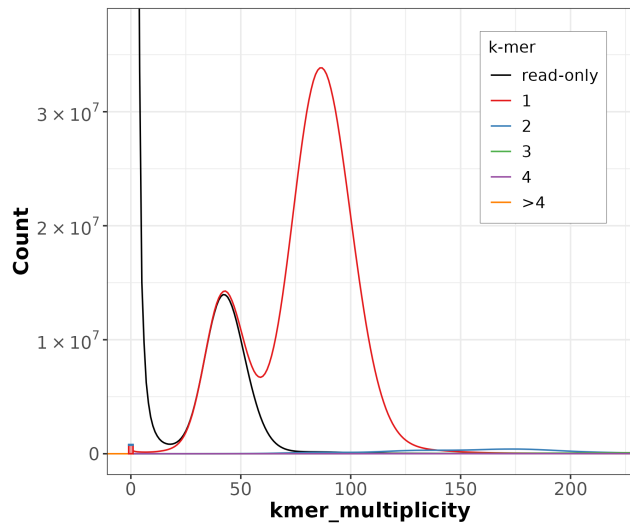
BUSCO: 5.8.2 (euk_genome_min, miniprot) / Lineage: actinopterygii_odb12 (genomes:75, BUSCOs:7207)

HiC contact map of curated assembly

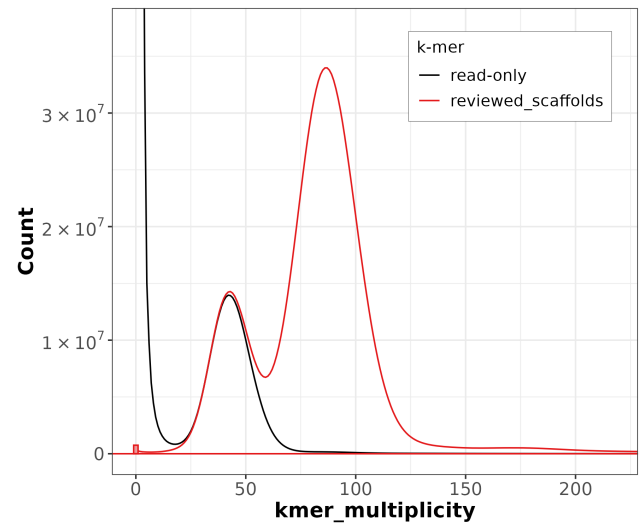


collapsed [\[LINK\]](#)

K-mer spectra of curated assembly

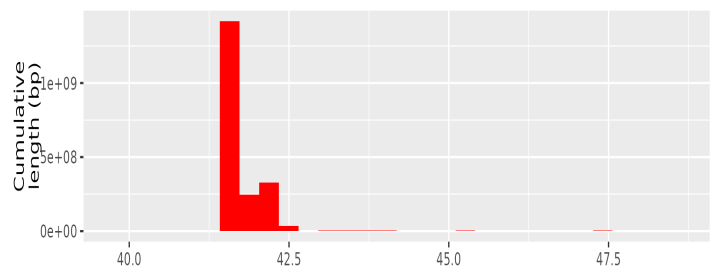


Distribution of k-mer counts per copy numbers found in asm

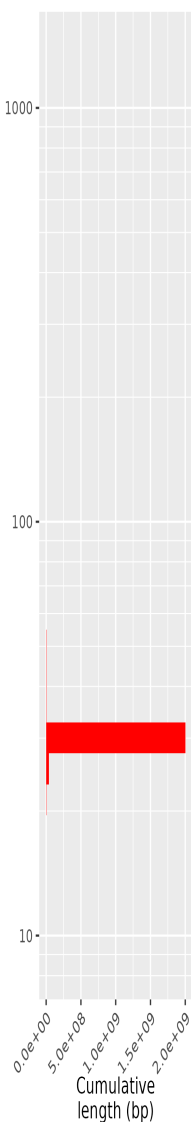
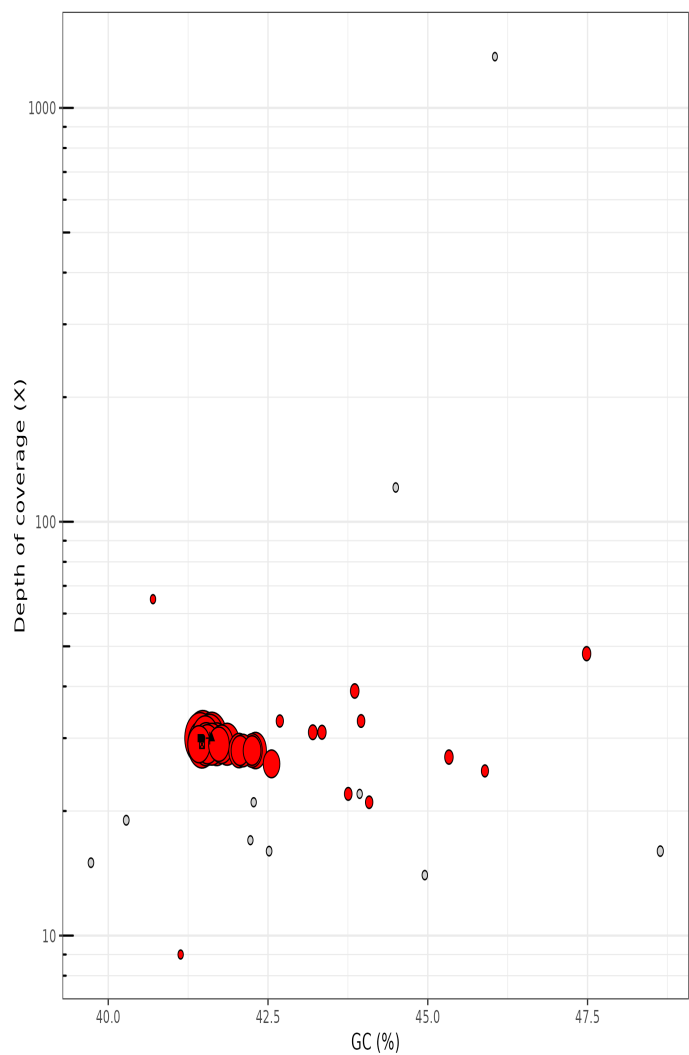


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



- Length (bp)
- 5.0e+07
 - 1.0e+08
 - 1.5e+08
- superkingdom
- Eukaryota
 - N/A
- Longest sequences (bp)
- fGymMii1_1 - 199514236 (Eukaryota)
 - ▲ fGymMii1_2 - 171619469 (Eukaryota)
 - fGymMii1_3 - 165597936 (Eukaryota)
 - + fGymMii1_5 - 141231199 (Eukaryota)
 - ▣ fGymMii1_4 - 139850289 (Eukaryota)

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	30	129

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: Emilie Teodori

Affiliation: Genoscope

Date and time: 2026-07-24 21:15:17 CEST