ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	1446102
ToLID	**wpGlyFall1.1**
Species	Glycera fallax
Class	Polychaeta
Order	Phyllodocida

Genome Traits	Expected	Observed
Haploid size (bp)	3,452,143,121	4,176,881,704
Haploid Number	10 (source: ancestor)	19
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.8.Q53

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

. Observed Haploid size (bp) has >20% difference with Expected
. Observed Haploid Number is different from Expected

. Kmer completeness value is less than 90 for collapsed
. More than 1000 gaps/Gbp for collapsed

Curator notes

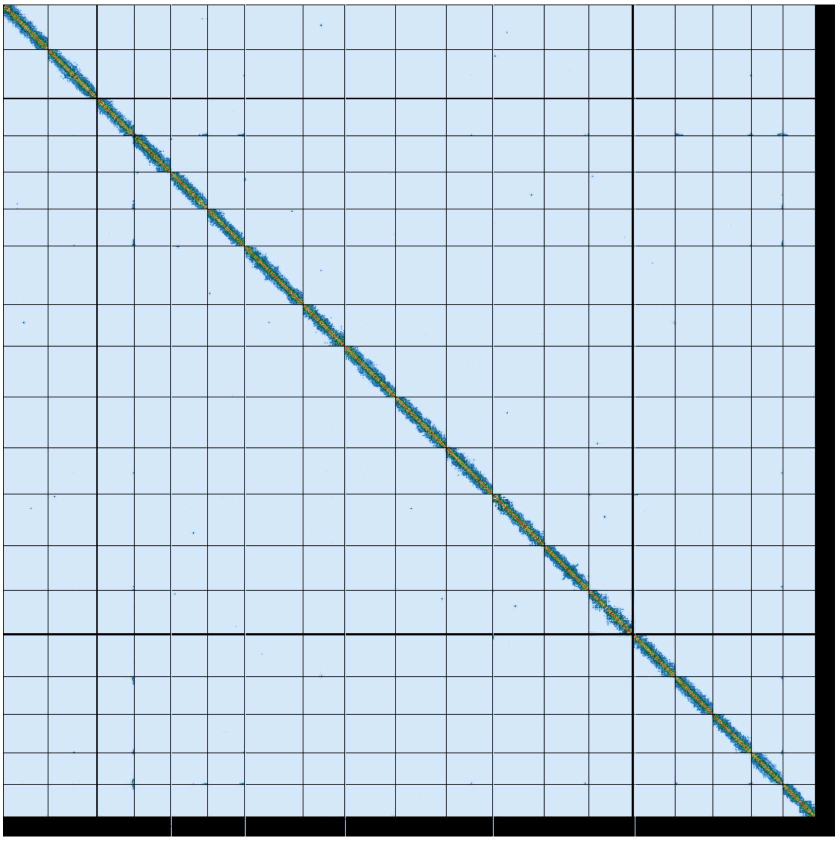
. Interventions/Gb: 104
. Contamination notes: ""
. Other observations: "The assembly of Glycera fallax (wpGlyFall1.1) is based on 39X PacBio data and Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial PacBio 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, 62 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 17.68 Mb (with the largest being 3.28 Mb). Additionally, 6,043 regions totaling 1,168 Mb (with the largest being 4.00 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, 40 haplotypic regions and 3 contaminant sequences were removed, totaling 41.95 Mb and 0.25 Mb, respectively (with the largest being 3.38 Mb and 0.15 Mb). 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	4,213,213,160	4,176,881,704
GC %	44.29	44.29
Gaps/Gbp	1,198.14	1,215.74
Total gap bp	504,800	522,500
Scaffolds	1,114	931
Scaffold N50	231,445,062	219,108,530
Scaffold L50	7	9
Scaffold L90	16	18
Contigs	6,162	6,009
Contig N50	1,497,747	1,507,316
Contig L50	861	848
Contig L90	2,981	2,938
QV	43.5219	53.3594
Kmer compl.	69.2895	69.9981
BUSCO sing.	95.0%	95.4%
BUSCO dupl.	1.5%	1.0%
BUSCO frag.	2.3%	2.3%
BUSCO miss.	1.2%	1.3%

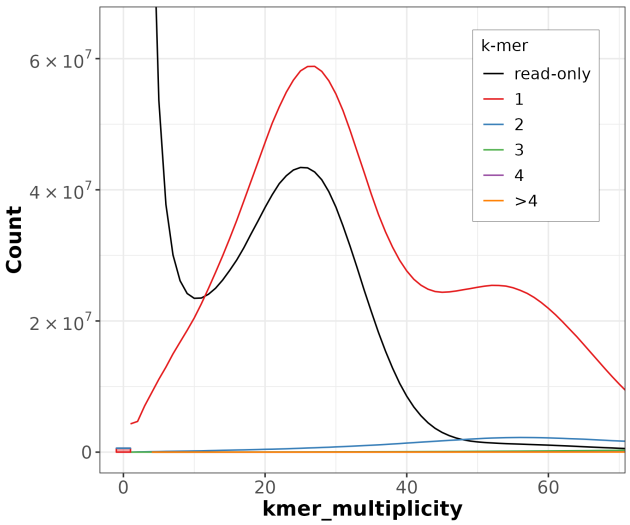
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: metazoa_odb10 (genomes:65, BUSCOs:954)

HiC contact map of curated assembly

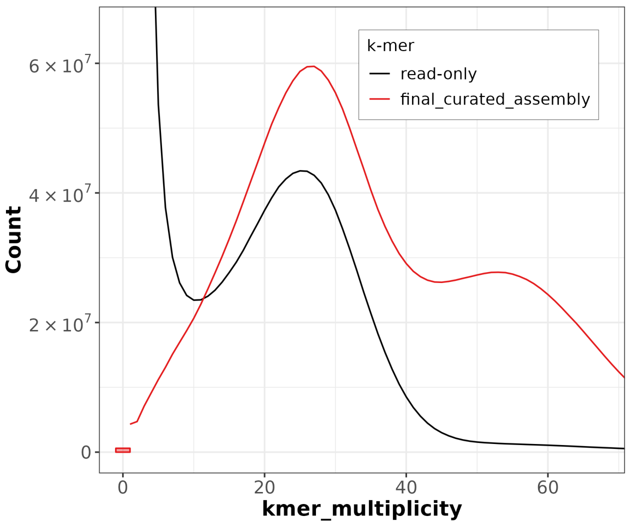


collapsed [LINK]

K-mer spectra of curated assembly

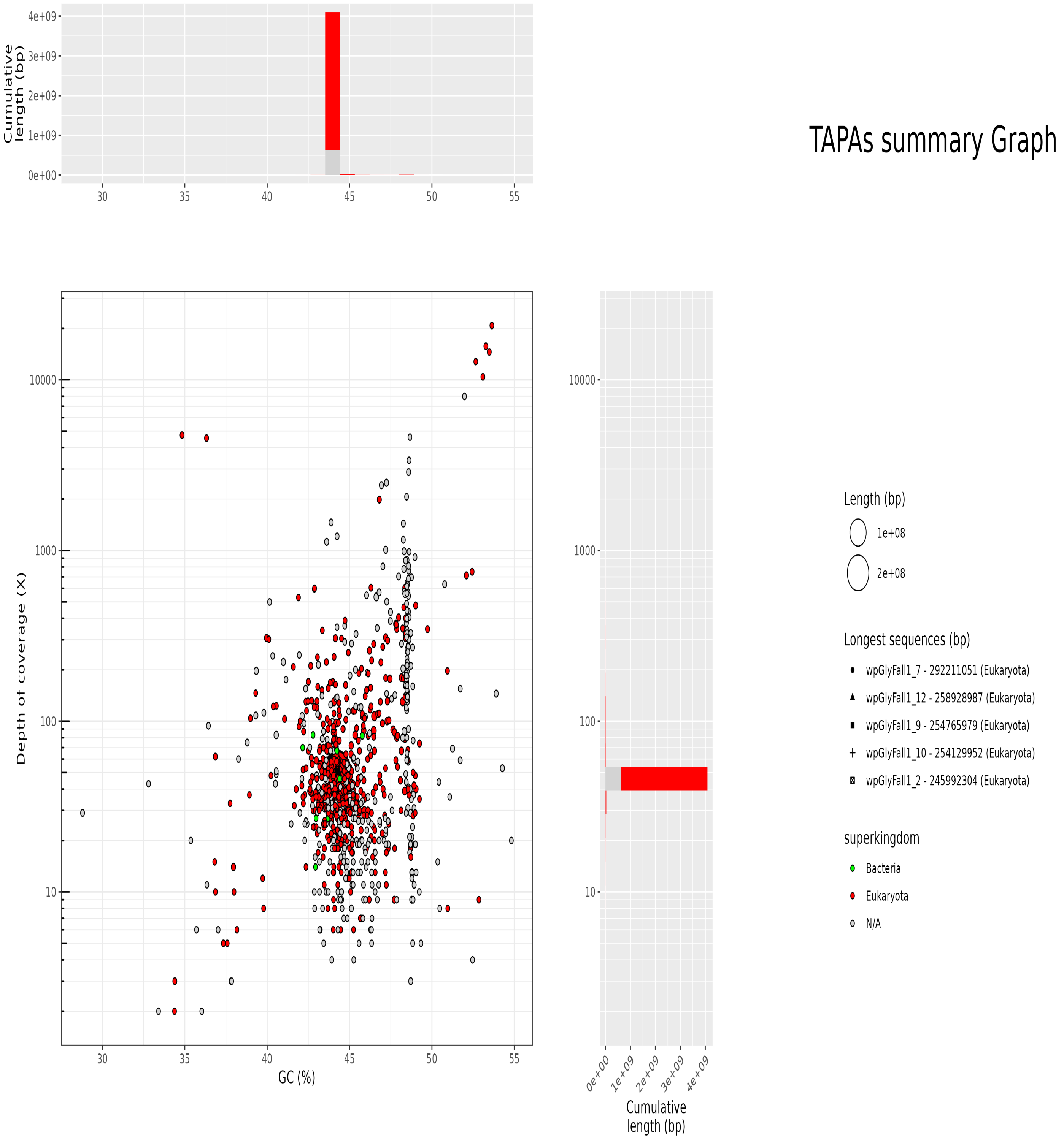


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	39	30

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: 2024-12-15 15:17:56 CET