ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1534734
ToLID	**daXanOrie1**
Species	Xanthium orientale
Class	Magnoliopsida
Order	Asterales

Genome Traits	Expected	Observed
Haploid size (bp)	2,269,223,707	2,258,156,991
Haploid Number	18 (source: direct)	18
Ploidy	4 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q68

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

. Observed Ploidy is different from Expected

. BUSCO duplicated value is more than 5% for collapsed

Curator notes

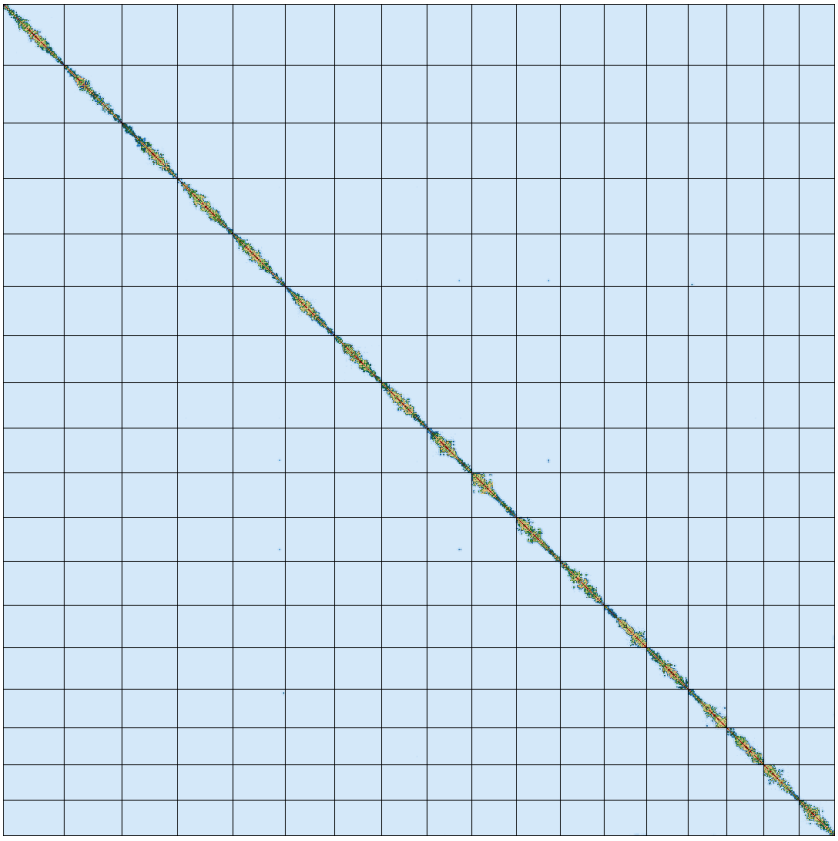
. Interventions/Gb: 1
. Contamination notes: ""
. Other observations: "The assembly of Xanthium orientale (daXanOrie1.2) is based on 40X of PacBio data and Arima Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe project (BGE, https://biodiversitygenomics.eu/). 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, 7 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 0.4 Mb (with the largest being 0.15 Mb). Additionally, 276 regions totaling 15.5 Mb (with the largest being 0.36 Mb) were identified as haplotypic duplications and removed. Mitochondrial and chloroplast genomes was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. 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	2,258,077,340	2,258,156,991
GC %	37.03	37.03
Gaps/Gbp	7.53	6.2
Total gap bp	1,700	1,400
Scaffolds	27	26
Scaffold N50	134,603,727	124,155,882
Scaffold L50	6	8
Scaffold L90	13	16
Contigs	44	40
Contig N50	98,804,340	98,804,340
Contig L50	9	9
Contig L90	21	21
QV	67.9927	68.0388
Kmer compl.	99.6146	99.6132
BUSCO sing.	93.3%	93.3%
BUSCO dupl.	5.8%	5.8%
BUSCO frag.	0.2%	0.2%
BUSCO miss.	0.7%	0.7%

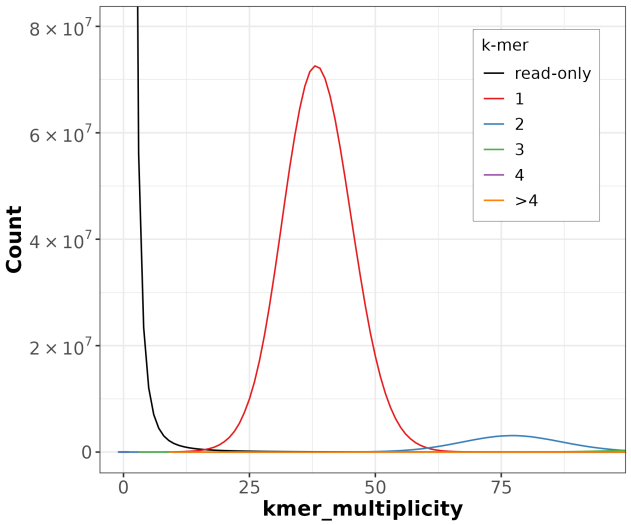
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: embryophyta_odb10 (genomes:50, BUSCOs:1614)

HiC contact map of curated assembly

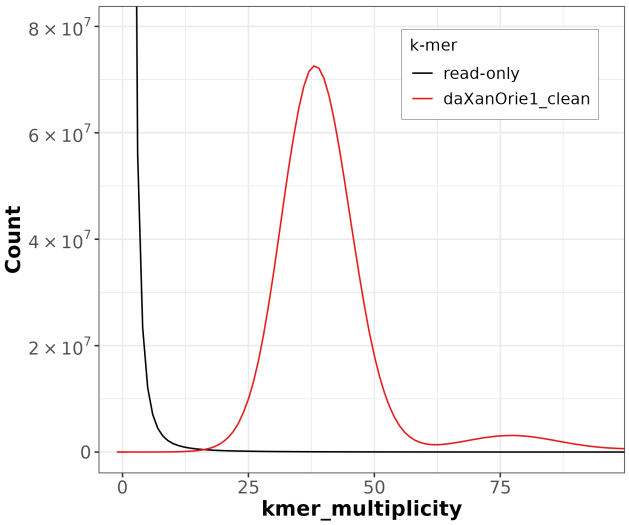


collapsed [LINK]

K-mer spectra of curated assembly

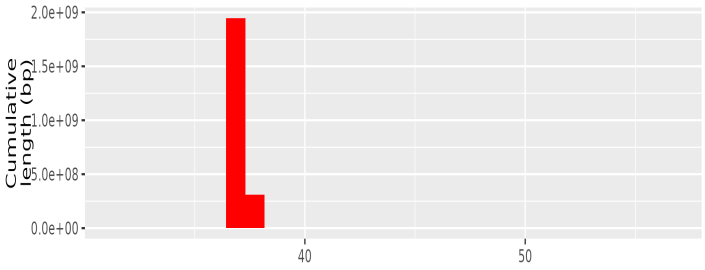


Distribution of k-mer counts per copy numbers found in asm

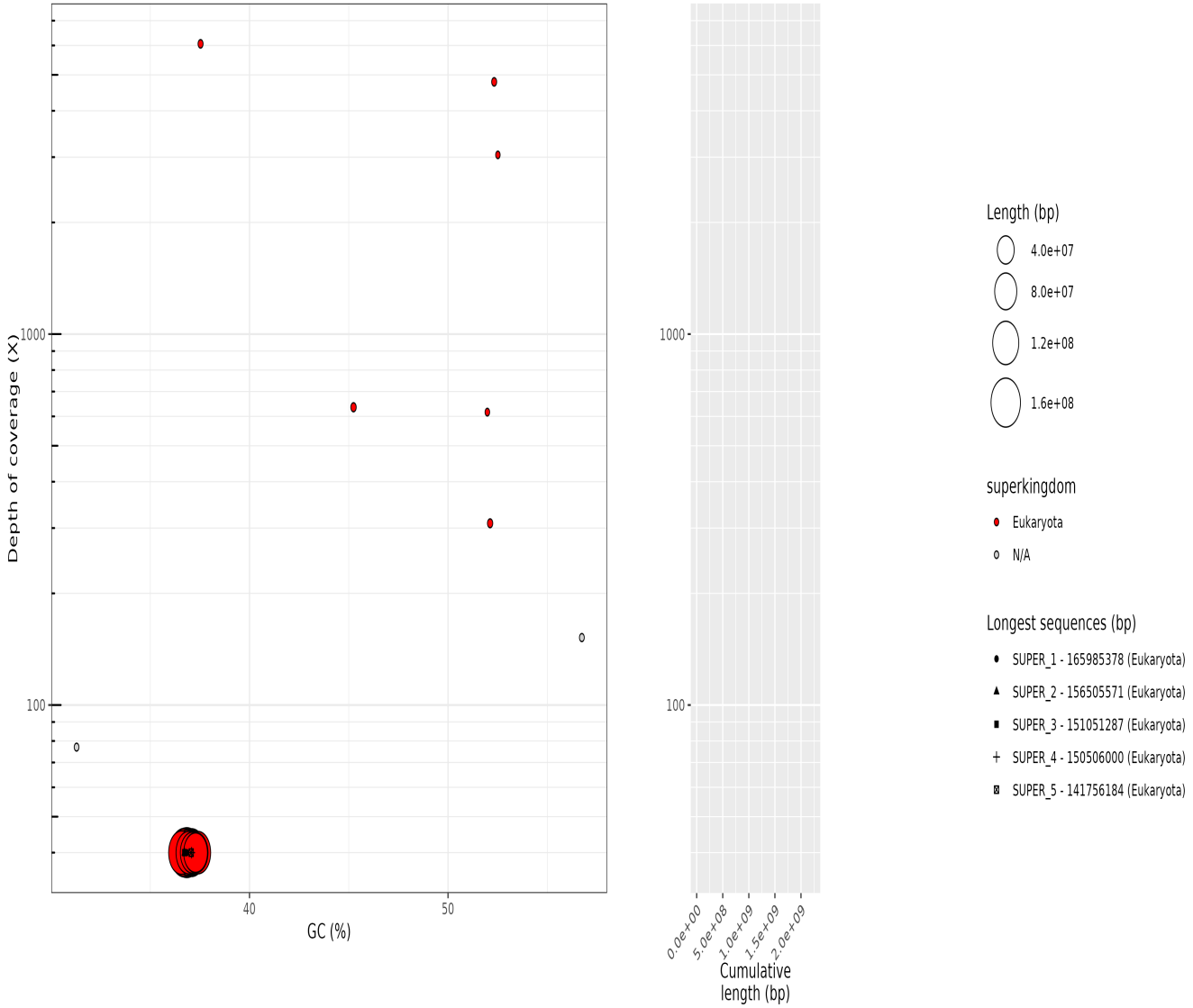


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	Omnic
Coverage	40	10

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: Adama Ndar
Affiliation: Genoscope

Date and time: 2025-01-23 05:04:09 CET