ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	71827
ToLID	**dcArmPung1**
Species	Armeria pungens
Class	Magnoliopsida
Order	Caryophyllales

Genome Traits	Expected	Observed
Haploid size (bp)	4,328,332,935	4,329,991,736
Haploid Number	9 (source: direct)	9
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 8.8.Q72

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
. Assembly length loss > 3% for collapsed

Curator notes

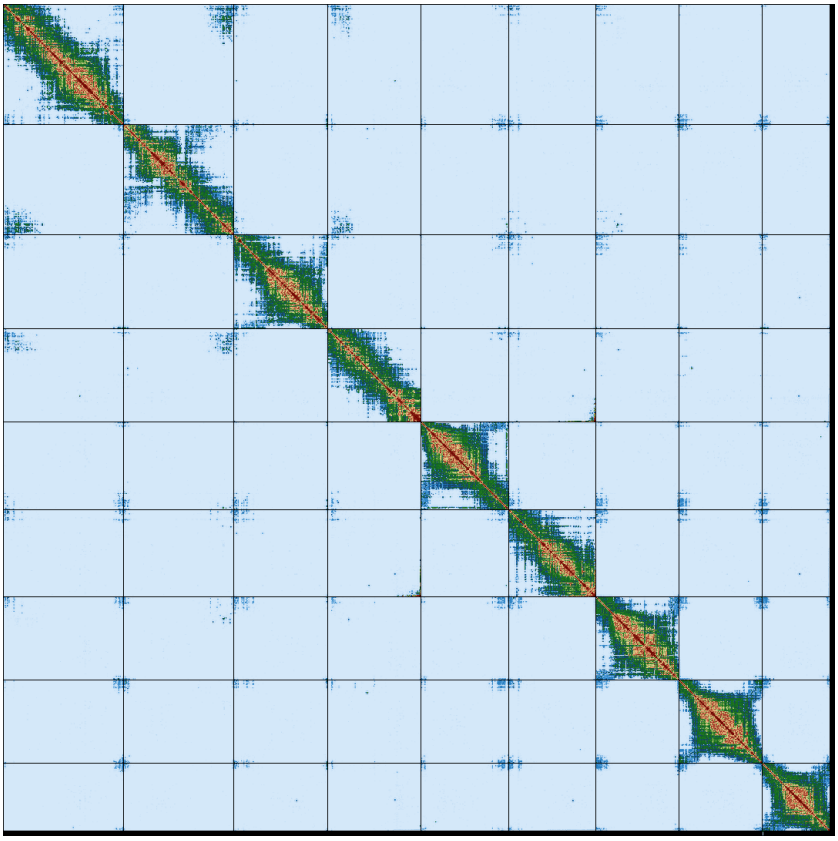
. Interventions/Gb: 10
. Contamination notes: ""
. Other observations: "The assembly of Armeria pungens (dcArmPung1) is based on 35X PacBio data and Omni-C Hi-C data generated as part of the European Reference Genome Atlas (ERGA, https://www.erga-biodiversity.eu/) via the Biodiversity Genomics Europe 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 (sequences > 1Mb were reincorporated to the assembly), and Hi-C-based scaffolding with YaHS. No contamination was detected. Additionally, 1066 regions totaling 73 Mb (with the largest being 911 Kb) were identified as haplotypic duplications and removed. The mitochondrial and chloroplast genomes were assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 26 haplotypic regions were removed, totaling 139 Mb (with the largest being 22 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	4,469,371,627	4,329,991,736
GC %	38.43	38.44
Gaps/Gbp	10.96	9.7
Total gap bp	4,900	5,000
Scaffolds	83	66
Scaffold N50	638,378,407	483,591,384
Scaffold L50	3	4
Scaffold L90	7	8
Contigs	132	108
Contig N50	181,861,489	181,861,489
Contig L50	9	9
Contig L90	29	25
QV	72.4497	72.2601
Kmer compl.	85.1393	83.5255
BUSCO sing.	88.1%	90.2%
BUSCO dupl.	8.3%	6.1%
BUSCO frag.	1.2%	1.3%
BUSCO miss.	2.4%	2.4%

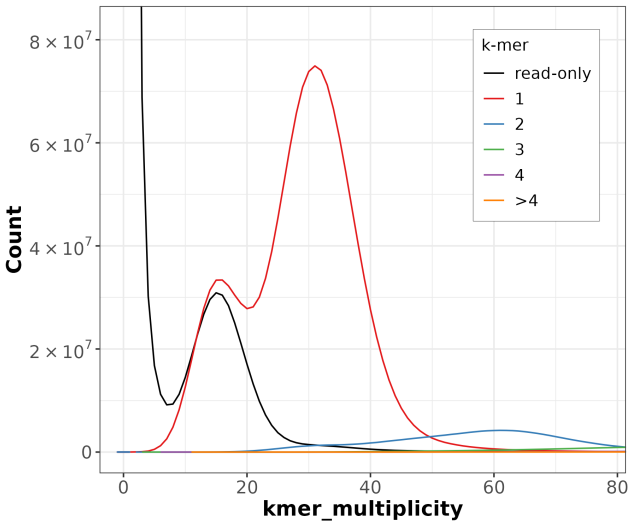
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: embryophyta_odb10 (genomes:50, BUSCOs:1614)

HiC contact map of curated assembly

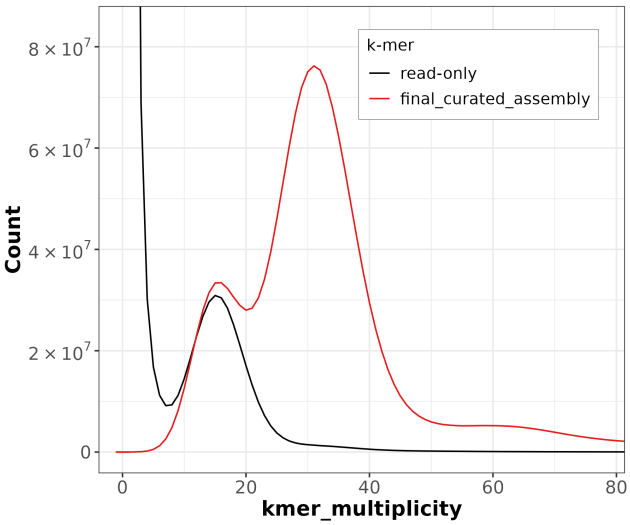


collapsed [LINK]

K-mer spectra of curated assembly

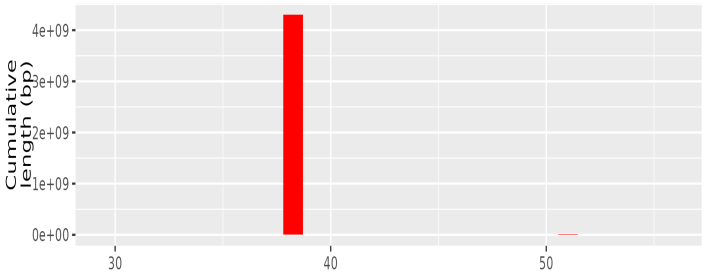


Distribution of k-mer counts per copy numbers found in asm

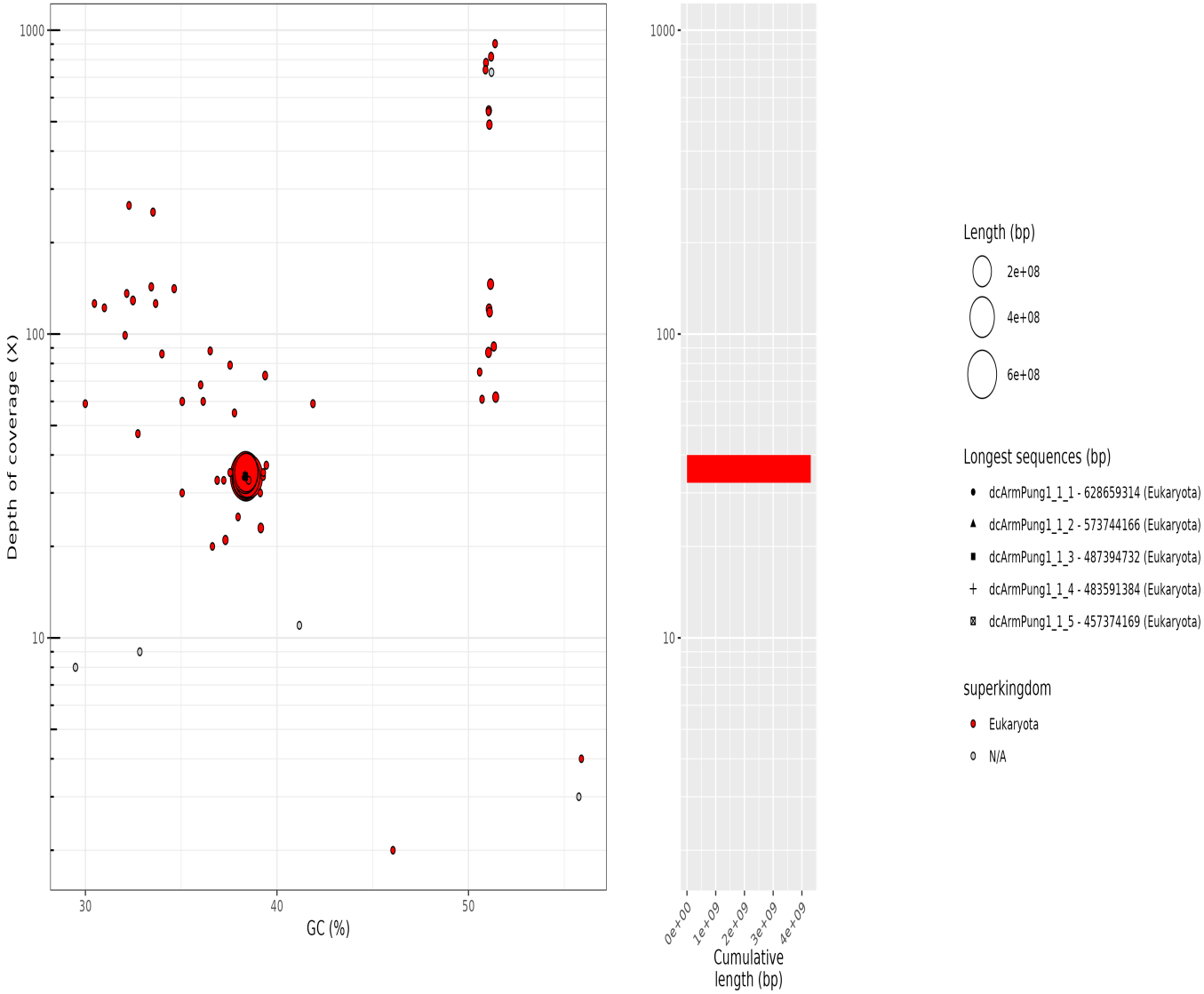


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	35	60

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: Caroline Belser
Affiliation: Genoscope

Date and time: 2025-02-20 12:03:51 CET