ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	184181
ToLID	**ddPriHirs2**
Species	Primula hirsuta
Class	Magnoliopsida
Order	Ericales

Genome Traits	Expected	Observed
Haploid size (bp)	1,561,081,177	1,521,706,300
Haploid Number	32 (source: direct)	31
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q66

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
. BUSCO single copy 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: 64
. Contamination notes: ""
. Other observations: "The assembly of PRIMULA HIRSUTA (ddPriHirs2) is based on 37X PacBio data and 126X 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, and Hi-C-based scaffolding with YaHS. In total, 10 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 4 Mb (with the largest being 2.7 Mb). Additionally, 1,492 regions totaling 53 Mb (with the largest being 993,547 pb) 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, 30 haplotypic regions were removed, totaling 54,740,569 pb (with the largest being 14,250,946 pb).

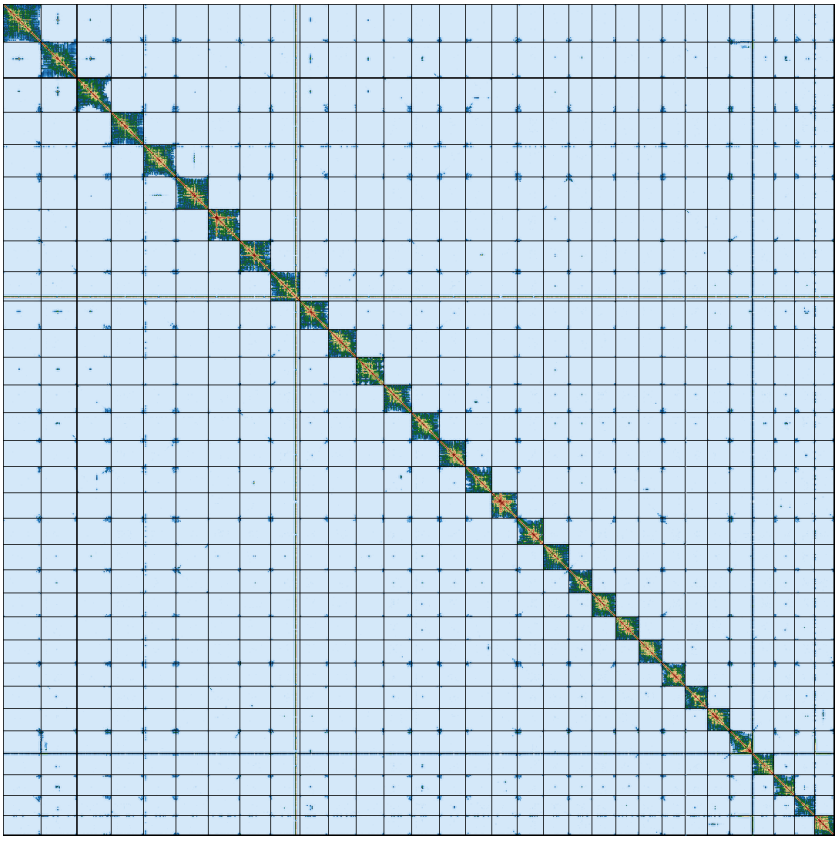
Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	1,575,465,818	1,521,706,300
GC %	35.33	35.32
Gaps/Gbp	0	65.72
Total gap bp	0	13,500
Scaffolds	190	74
Scaffold N50	23,035,158	49,900,556
Scaffold L50	22	14
Scaffold L90	69	27
Contigs	190	174
Contig N50	23,035,158	22,102,891
Contig L50	22	24
Contig L90	69	72
QV	66.5821	66.8709
Kmer compl.	76.8142	75.3927
BUSCO sing.	57.6%	59.0%
BUSCO dupl.	37.5%	36.2%
BUSCO frag.	3.1%	3.2%
BUSCO miss.	1.7%	1.7%

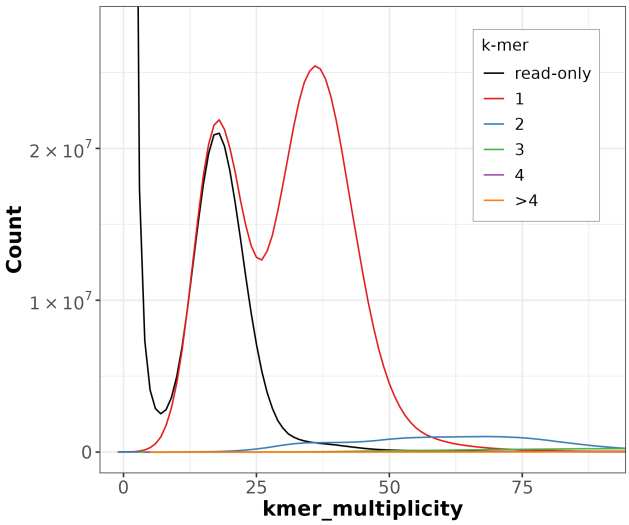
BUSCO: 5.8.2 (euk_genome_met, metaeuk) / Lineage: eudicotyledons_odb12 (genomes:76, BUSCOs:2805)

HiC contact map of curated assembly

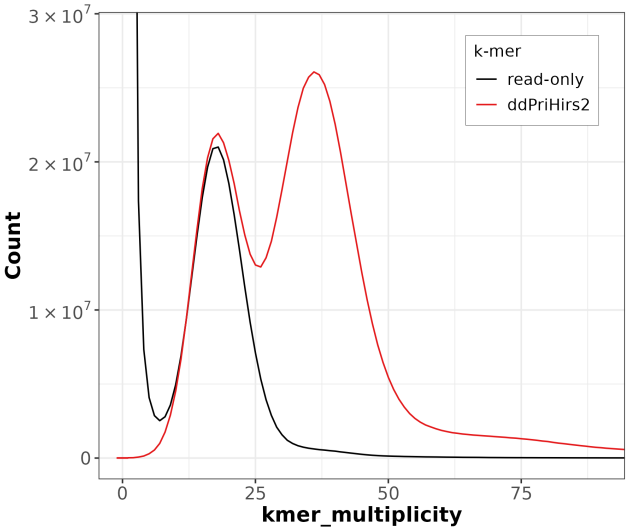


collapsed [LINK]

K-mer spectra of curated assembly

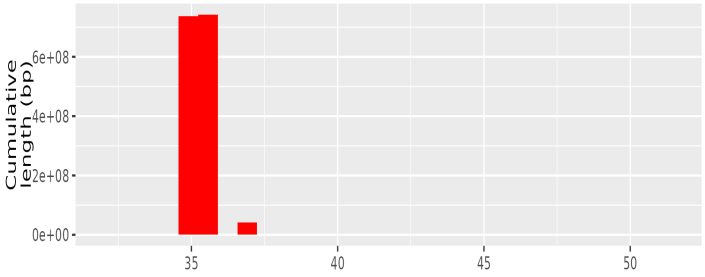


Distribution of k-mer counts per copy numbers found in asm

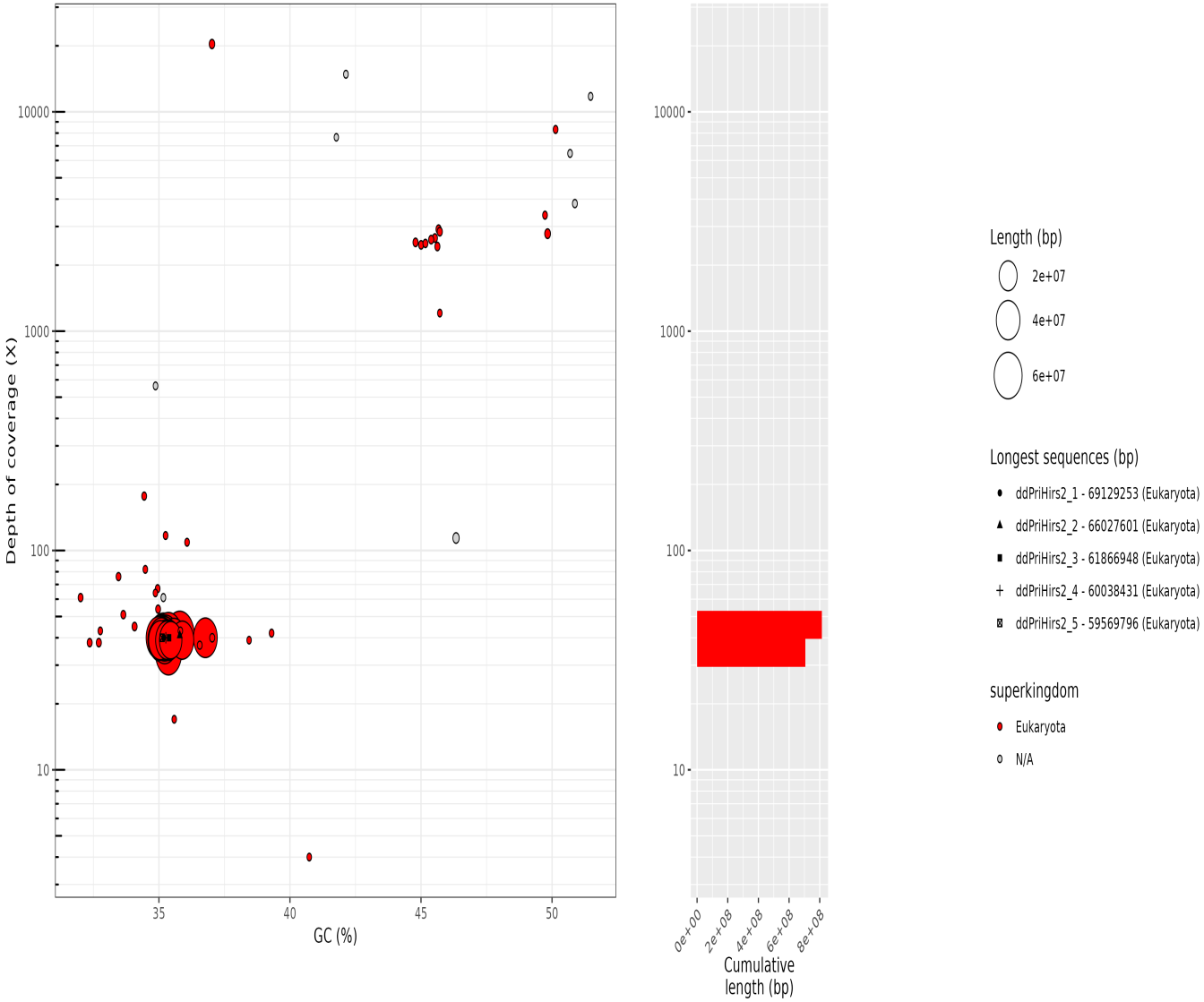


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	37	126

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: Lola Demirdjian
Affiliation: Genoscope

Date and time: 2025-06-11 12:48:33 CEST