

ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	213841
ToLID	drStaPinn1.1
Species	Staphylea pinnata
Class	Magnoliopsida
Order	Crossosomatales

Genome Traits	Expected	Observed
Haploid size (bp)	1,626,848,065	1,563,064,433
Haploid Number	12 (source: direct)	13
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.8.Q60

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 duplicated value is more than 5% for collapsed

Curator notes

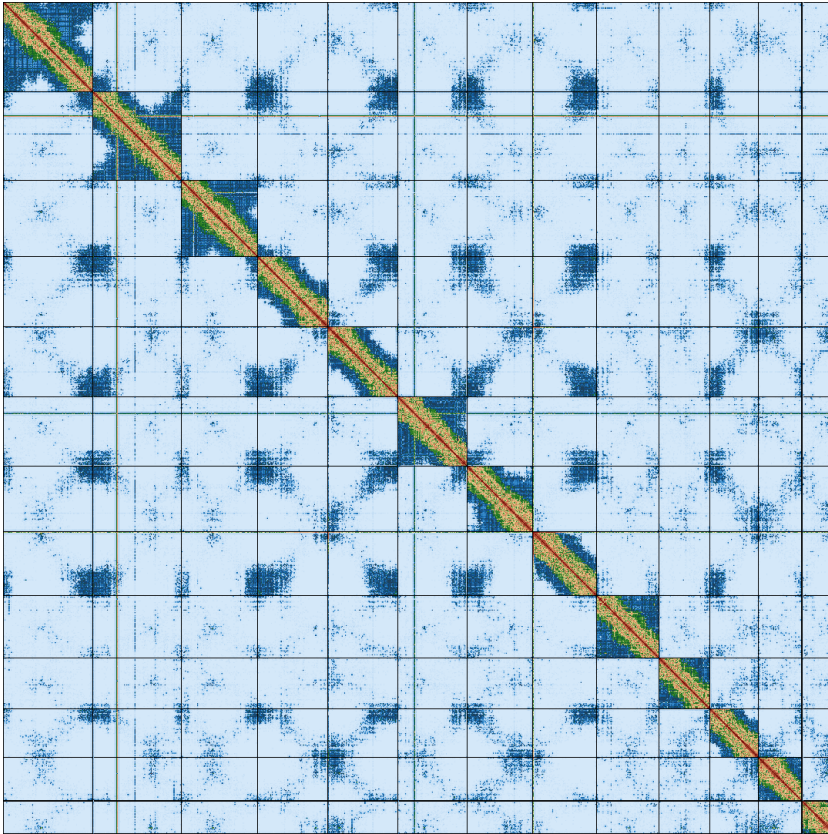
- . Interventions/Gb: 8
- . Contamination notes: ""
- . Other observations: "The assembly of STAPHYLEA PINNATA (drStaPinn1) is based on 50X PacBio data and 137X Omni-C Hi-C data generated as part of the European Reference Genome Atlas (ERGA, <https://www.erga-biodiversity.eu/>) via the Biodiversity Genomics Europe project (BGE, <https://biodiversitygenomics.eu/>). The assembly process included the following steps: initial PacBio assembly generation with Hifiiasm, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 5 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 390 Kb (with the largest being 277 Kb). Additionally, 650 regions totaling 102 Kb 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, no regions were tagged as allelic duplications but 3 regions as contaminants (151 Kb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size and contigs were already mostly chromosome-scale. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	1,562,129,863	1,563,064,433
GC %	35.67	35.68
Gaps/Gbp	0	15.99
Total gap bp	0	2,600
Scaffolds	47	23
Scaffold N50	79,008,546	130,005,016
Scaffold L50	7	6
Scaffold L90	21	11
Contigs	47	48
Contig N50	79,008,546	71,159,830
Contig L50	7	8
Contig L90	21	23
QV	60.7974	60.8091
Kmer compl.	89.6396	89.6852
BUSCO sing.	90.4%	90.4%
BUSCO dupl.	6.4%	6.4%
BUSCO frag.	0.4%	0.4%
BUSCO miss.	2.8%	2.8%

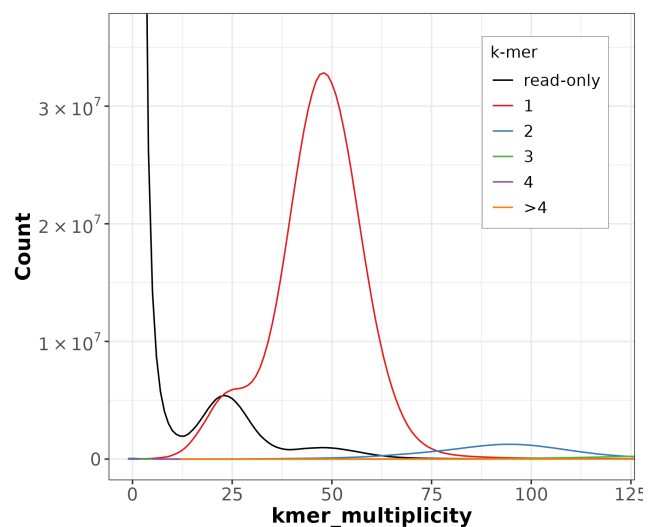
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: embryophyta_odb10 (genomes:50, BUSCOs:1614)

HiC contact map of curated assembly

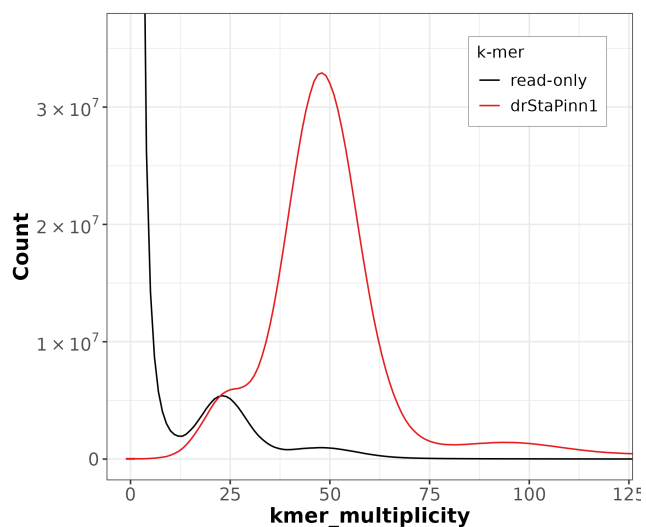


collapsed [\[LINK\]](#)

K-mer spectra of curated assembly

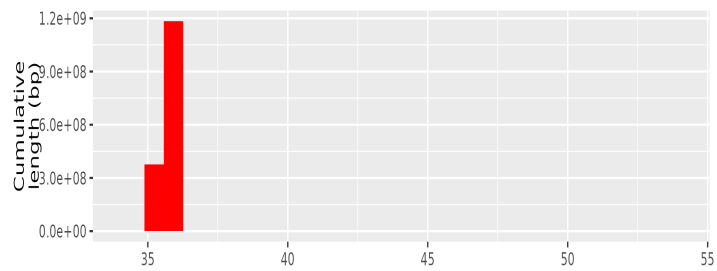


Distribution of k-mer counts per copy numbers found in asm

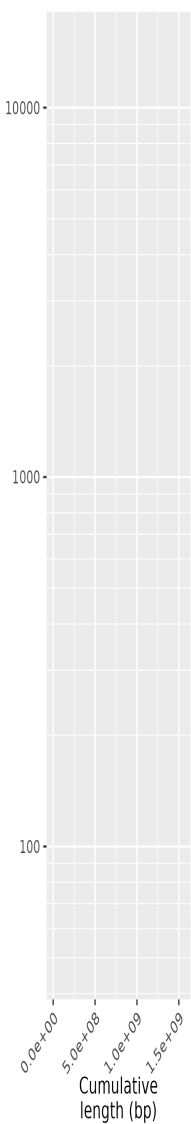
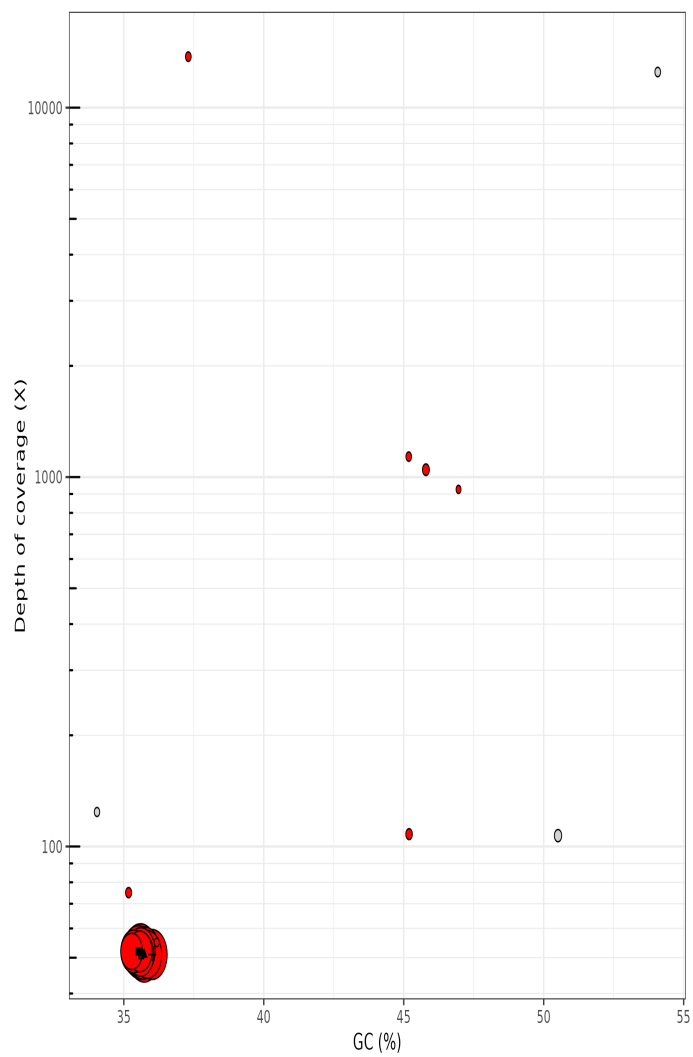


Distribution of k-mer counts coloured by their presence in reads/assemblies

Post-curation contamination screening



TAPAs summary Graph



Length (bp)

- 4.0e+07
- 8.0e+07
- 1.2e+08
- 1.6e+08

Longest sequences (bp)

- drStaPinn1_1 - 168484217 (Eukaryota)
- drStaPinn1_2 - 166806080 (Eukaryota)
- drStaPinn1_3 - 142608644 (Eukaryota)
- drStaPinn1_4 - 132689066 (Eukaryota)
- drStaPinn1_5 - 130443848 (Eukaryota)

superkingdom

- Eukaryota
- N/A

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	50	137

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: 2025-02-05 16:32:47 CET