

# ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	3086330
ToLID	<b>drCotNebr1</b>
Species	<i>Cotoneaster nebrodensis</i>
Class	Magnoliopsida
Order	Rosales

Genome Traits	Expected	Observed
Haploid size (bp)	1,029,637,717	2,837,986,727
Haploid Number	34 (source: direct)	87
Ploidy	4 (source: ancestor)	2
Sample Sex	Unknown	Unknown

## EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q70

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
- . Observed Ploidy is different from Expected
- . BUSCO single copy value is less than 90% for collapsed
- . BUSCO duplicated value is more than 5% for collapsed

## Curator notes

- . Interventions/Gb: 148
- . Contamination notes: ""
- . Other observations: "The assembly of *Cotoneaster nebrodensis* (drCotNebr1) is based on 20X PacBio data and 98X OmniC 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/>). This genome is decaploid and expected to have an AAABB structure. The assembly was performed using PacBio and Hi-C data, with the -n\_hap parameter set to 2. Several tests were performed with this option, and only with this configuration were we able to retrieve the five chromosomes for each haplotype: in general, hap1 contains two A copies (often fused into a single scaffold) and one B chromosome, while hap2 contains one A and one B chromosome. The assembly process included the following steps: initial PacBio assembly generation with Hifiasm, removal of contaminant sequences using Context, and Hi-C-based scaffolding with YaHS. The mitochondrial and chloroplastic genomes were assembled using OATK. Finally, the primary assembly was analyzed and

manually improved using Pretext. During manual curation, 895 contaminant sequences were removed, totaling 38.13 Mb (the largest being 0.51 Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size (average size of the 3 A and 2 B chromosomes), and similarly, A and B copies were named according to their size. Note that the assembly contains more than the expected  $5 \times 17 = 85$  chromosomes, as the two expected B chromosome 1 copies are split into four chromosomes; consequently, the final assembly contains a total of 87 chromosomes. "

# Quality metrics table

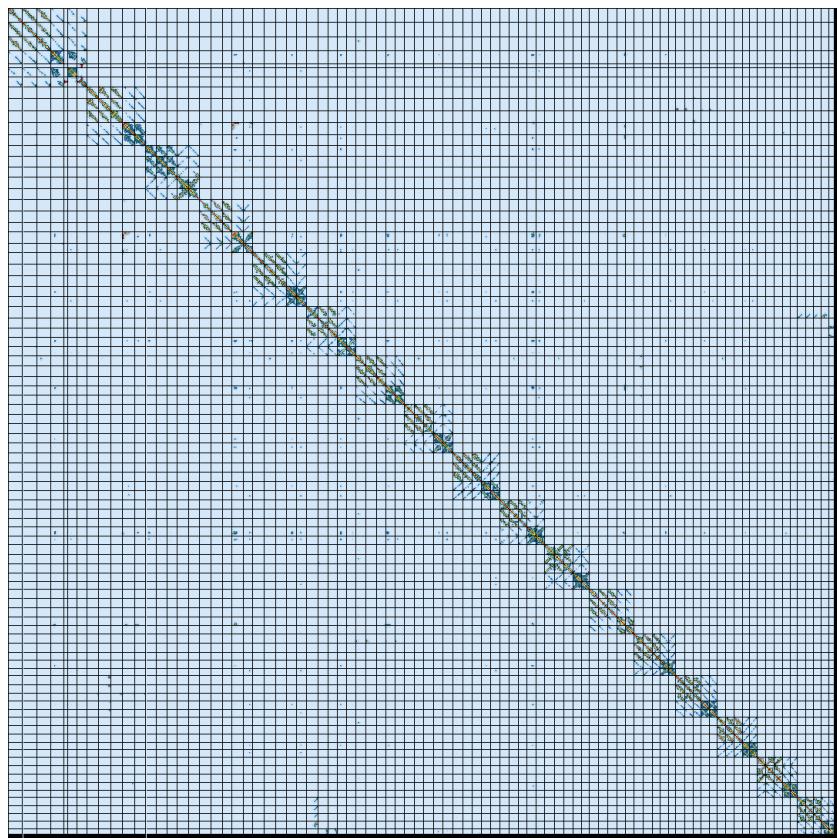
Metrics	Pre-curation collapsed	Curated collapsed
Total bp	1,455,938,951	2,837,986,727
GC %	38.34	38.17
Gaps/Gbp	156.6	153.63
Total gap bp	22,800	67,800
Scaffolds	620	421
Scaffold N50	72,466,578	32,471,515
Scaffold L50	9	39
Scaffold L90	21	77
Contigs	848	857
Contig N50	25,638,989	11,316,000
Contig L50	24	73
Contig L90	67	257
QV	65.8183	70.7191
Kmer compl.	78.602	99.3787
BUSCO sing.	3.8%	0.1%
BUSCO dupl.	91.9%	97.6%
BUSCO frag.	0.1%	0.0%
BUSCO miss.	4.2%	2.3%

Warning! BUSCO versions or lineage datasets are not the same across results:

BUSCO: 5.8.2 (euk\_genome\_met, metaeuk) / Lineage: rosaceae\_odb12 (genomes:5, BUSCOs:10071)

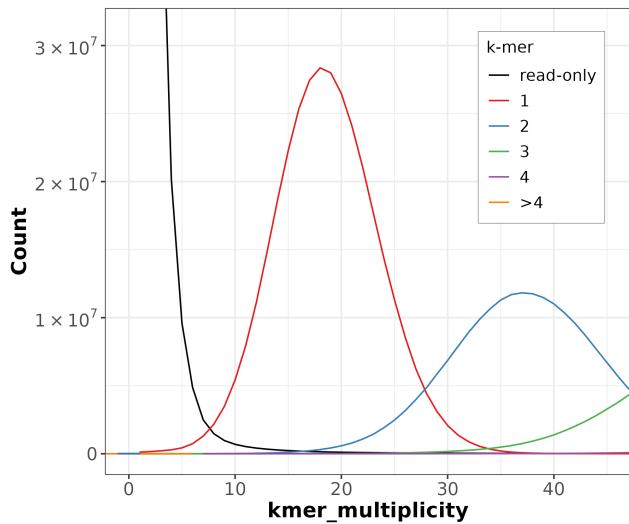
BUSCO: 6.0.0 (euk\_genome\_min, miniprot) / Lineage: rosaceae\_odb12 (genomes:5, BUSCOs:10071)

# HiC contact map of curated assembly

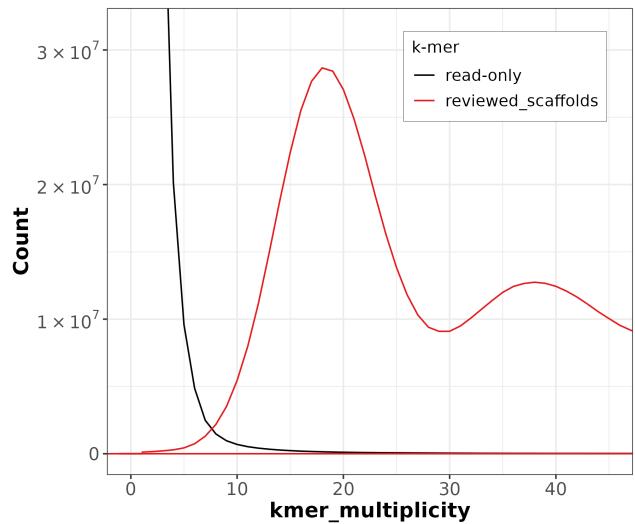


collapsed [\[LINK\]](#)

# K-mer spectra of curated assembly

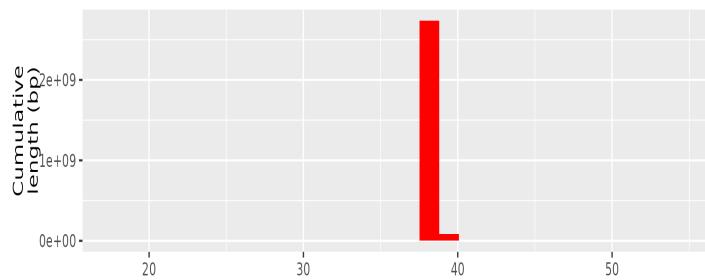


Distribution of k-mer counts per copy numbers found in asm



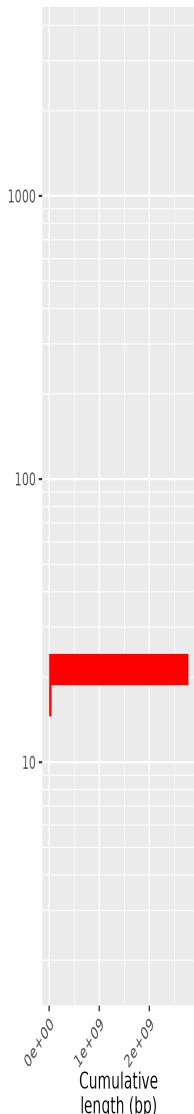
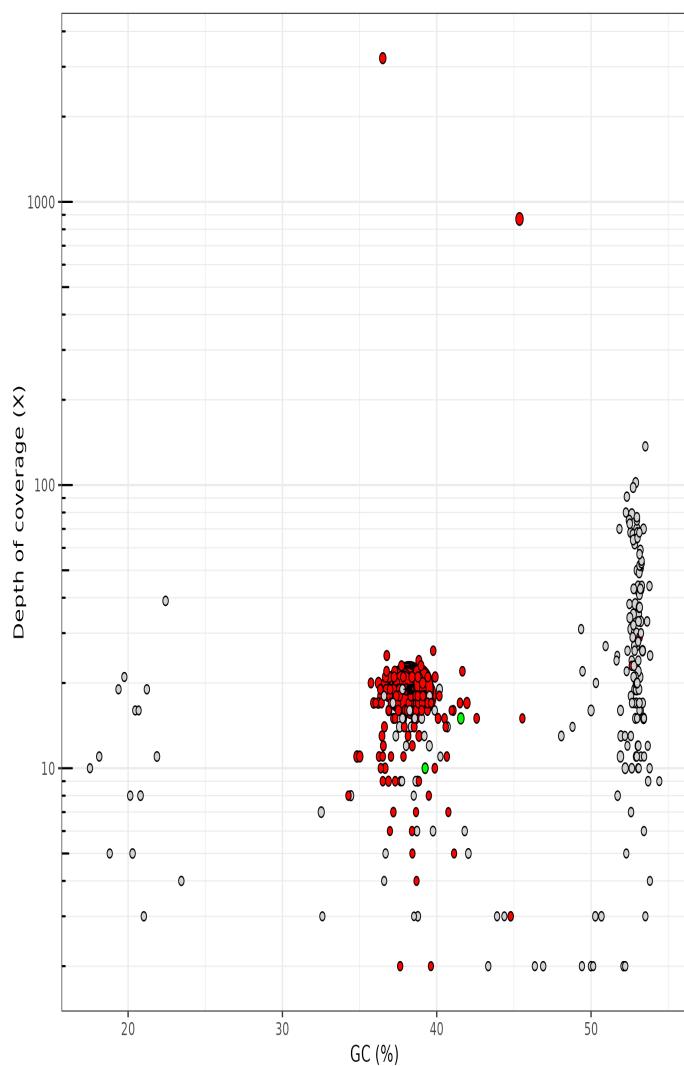
Distribution of k-mer counts coloured by their presence in reads/assemblies

# Post-curation contamination screening



TAPAs summary Graph

(3 0X contigs have been hidden)



## superkingdom

- Bacteria
- Eukaryota
- N/A

## Length (bp)

- 1e+07
- 2e+07
- 3e+07
- 4e+07

## Longest sequences (bp)

- drCotNebr1\_1 - 48856001 (Eukaryota)
- ▲ drCotNebr1\_2 - 48518616 (Eukaryota)
- drCotNebr1\_3 - 48140259 (Eukaryota)
- + drCotNebr1\_4 - 45437404 (Eukaryota)
- ▣ drCotNebr1\_11 - 43783242 (Eukaryota)

**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	Long reads	Omnic
Coverage	53	264

## 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 Menguy

Affiliation: Genoscope

Date and time: 2026-02-10 14:20:07 CET