ERGA Assembly Report

v24.02.09_beta

Tags: ERGA-Pilot

ToLID	**ddCorAlbu1**
Species	**Corema album**
Class	Magnoliopsida
Order	Ericales

Genome Traits	Expected	Observed
Haploid size (bp)	628,534,562	548,232,794
Haploid Number	13 (source: direct)	13
Ploidy	2 (source: direct)	2
Sample Sex	NA	unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for pri: 7.7.Q67

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

. Observed sex is different from Sample sex

. Kmer completeness value is less than 90 for pri
. Assembly length loss > 3% for pri

Curator notes

. Interventions/Gb: None
. Contamination notes: "Lots of fungal sequences and one insect detected. These were removed before scaffolding"
. Other observations: "PacBio and HiC data were from different specimens and the HiC was made of a mix of male and female parts of the plant. There were a lot of retained duplicates after the initial contig assembly and purge-dups was found to over-purge, removing quite a few genes. As such, large haplotids were manually identifed during curation and purge-dups was run after curation to remove JUNK, REPEAT and HIGHCOV sequenes. The 500kb mitochria and 100kb chloroplast were also assembled using oatk. There were a lot of fragments from the mito and chloroplast in the genome that were manually removed."

Quality metrics table

| Metrics	Pre-curation pri	Curated pri
Total bp	683,082,712	548,232,794
GC %	39.3	38.7
Gaps/Gbp	414.3	51.07
Total gap bp	28,300	2,800
Scaffolds	890	66
Scaffold N50	47,309,436	40,943,406
Scaffold L50	5	6
Scaffold L90	13	12
Contigs	1,173	94
Contig N50	17,254,601	23,686,000
Contig L50	12	9
Contig L90	48	25
QV	59.2597	67.0605
Kmer compl.	86.8115	84.7354
BUSCO sing.	91.4%	93.9%
BUSCO dupl.	5.2%	2.8%
BUSCO frag.	0.6%	0.7%
BUSCO miss.	2.8%	2.6%

BUSCO 5.4.7 Lineage: eudicots_odb10 (genomes:31, BUSCOs:2326)

HiC contact map of curated assembly

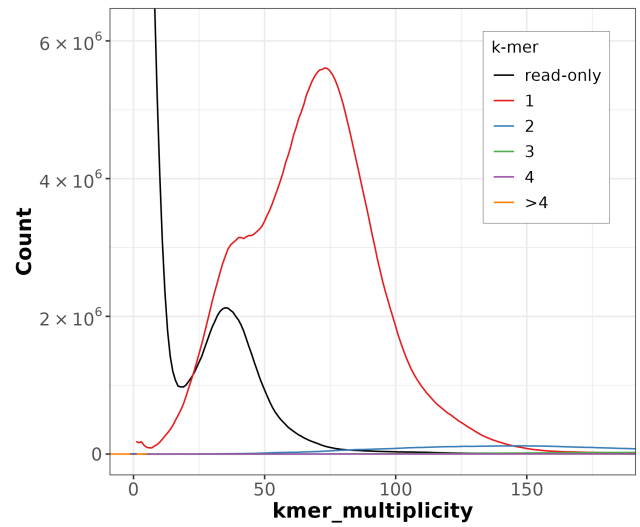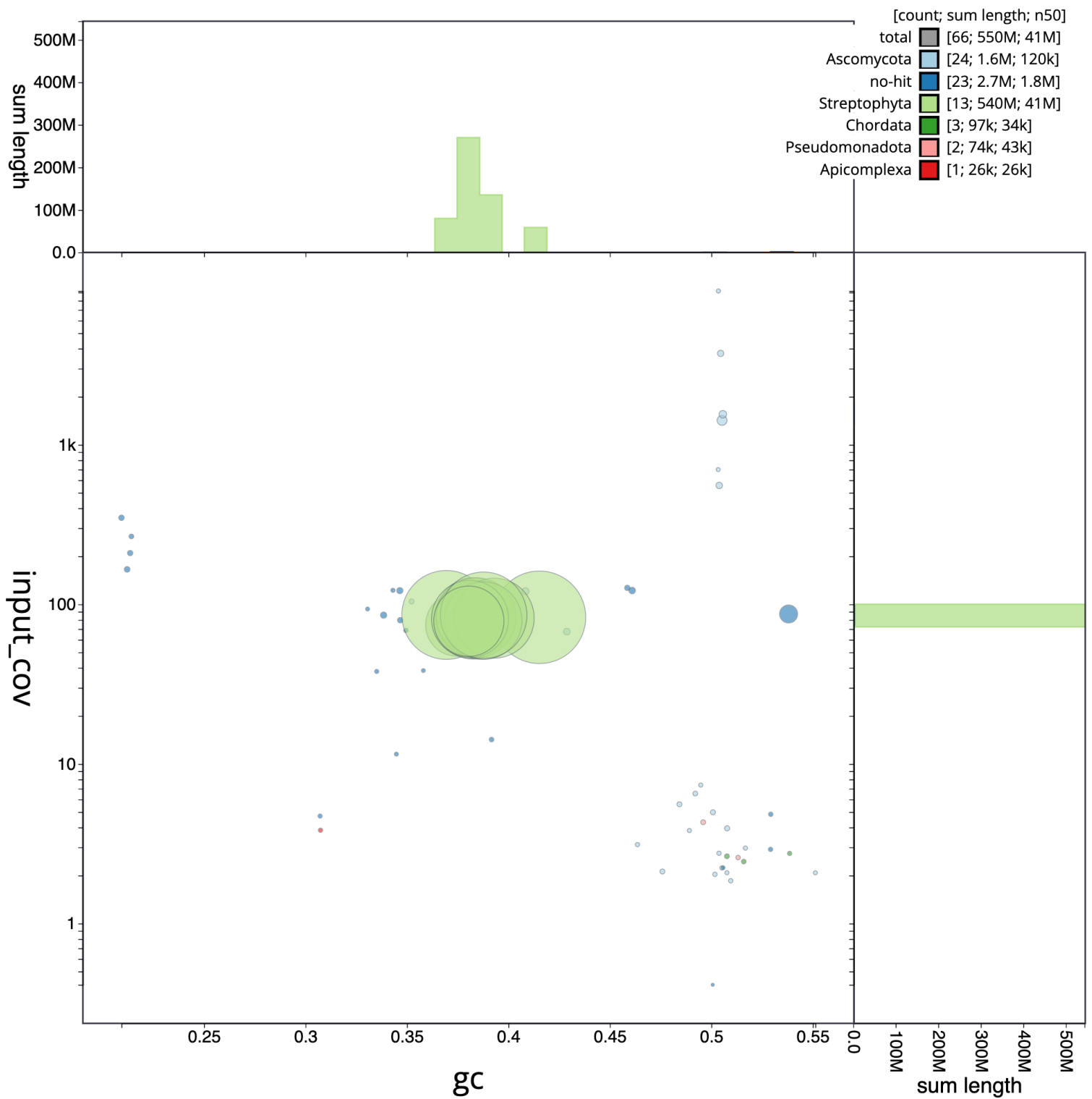**pri**
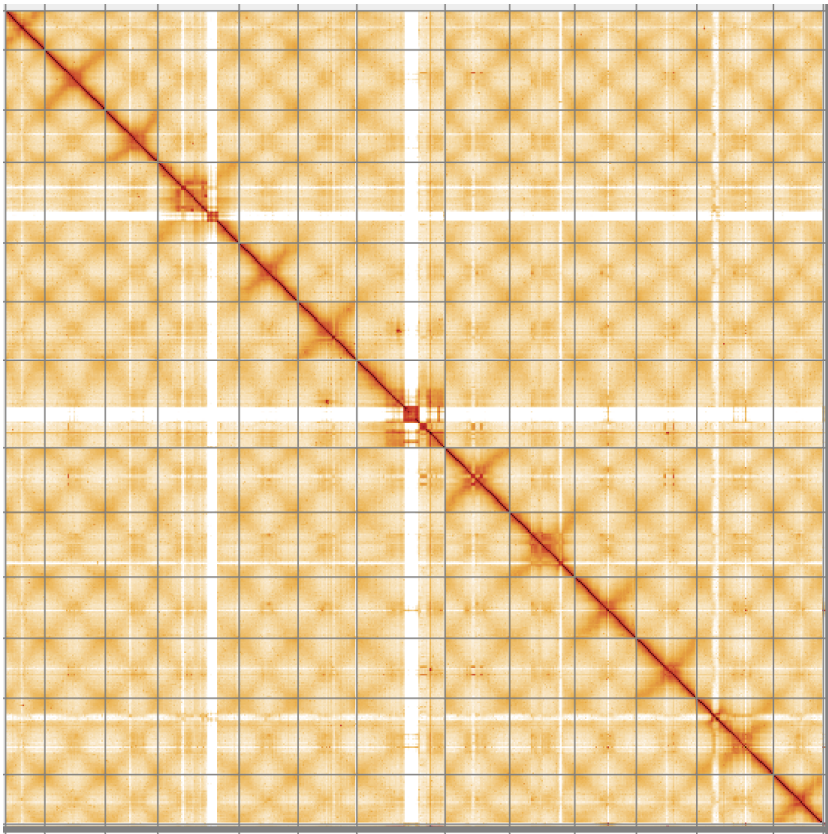
K-mer spectra of curated assembly
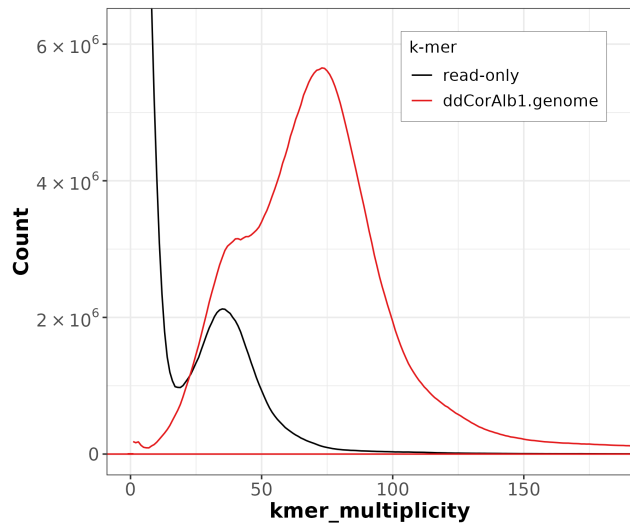
Distribution of k-mer counts coloured by
their presence in reads/assemblies

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

Post-curation contamination screening

pri. 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	Illumina Hic
Coverage	76X	50X

Assembly pipeline

- **Hifiasm**
    |_ *ver:* 0.19.8
    |_ *key param:* -l 2
- **yahs**
    |_ *ver:* 1.2
    |_ *key param:* NA
- **fcs-gx**
    |_ *ver:* 0.5.0
    |_ *key param:* NA
- **purge-dups**
    |_ *ver:* 1.2.6
    |_ *key param:* NA

Curation pipeline

- **higlass**
    |_ *ver:* 0.6.1
    |_ *key param:* NA
- **pretext**
    |_ *ver:* 0.2.5
    |_ *key param:* NA
- **blobtoolkit**
    |_ *ver:* 4.1.0
    |_ *key param:* NA

Submitter: Tom Brown
Affiliation: IZW

Date and time: 2024-03-14 18:53:12 CET