ERGA Assembly Report

v24.10.15

Tags: ERGA-BGE

TxID	1172132
ToLID	**cbLewAcum8.1**
Species	Lewinskya acuminata
Class	Bryopsida
Order	Orthotrichales

Genome Traits	Expected	Observed
Haploid size (bp)	273,294,980	256,705,263
Haploid Number	6 (source: ancestor)	6
Ploidy	1 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 7.7.Q54

The following metrics were automatically flagged as below EBP recommended standards or 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: 85
. Contamination notes: ""
. Other observations: "The assembly of LEWINSKYA ACUMINATA (cbLewAcum8) is based on 69X PacBio data and 202X Arima 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 Nextdenovo, removal of contaminant sequences using Context, removal of haplotypic duplications using purge_dups, and Hi-C-based scaffolding with YaHS. In total, 32 regions totaling 4 Mb were identified as haplotypic duplications and removed. Additionally, 114 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 23 Mb (with the largest being 4 Mb). The mitochondrial genome was assembled using OATK. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 8 haplotypic regions and 5 contaminant sequences were removed, totaling 423,178 pb and 456,681 pb, respectively (with the largest being 180,566 pb and 162,112 pb). There is a large haplotypic inversion between ~3.79-4.36Mb (chr 1). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size and

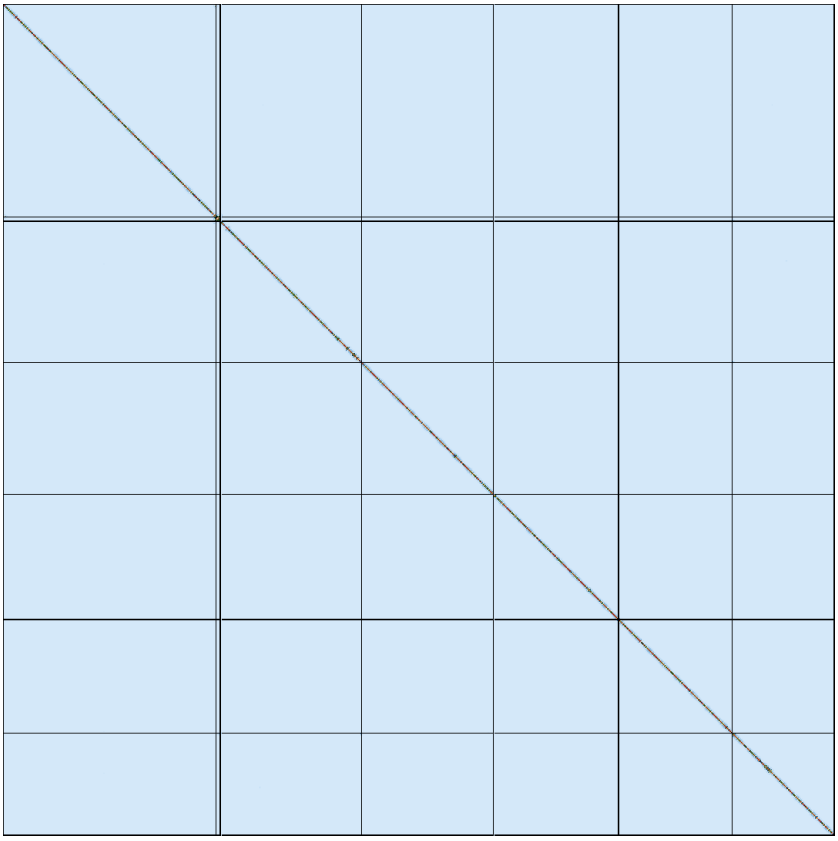
contigs were already mostly chromosome-scale. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	257,229,379	256,705,263
GC %	37.5	37.51
Gaps/Gbp	0	148.03
Total gap bp	0	4,700
Scaffolds	63	16
Scaffold N50	10,772,754	40,771,022
Scaffold L50	7	3
Scaffold L90	22	6
Contigs	63	54
Contig N50	10,772,754	11,461,491
Contig L50	7	7
Contig L90	22	20
QV	54.3675	54.4078
Kmer compl.	97.548	97.4963
BUSCO sing.	77.1%	77.0%
BUSCO dupl.	5.7%	5.7%
BUSCO frag.	2.0%	2.0%
BUSCO miss.	15.2%	15.3%

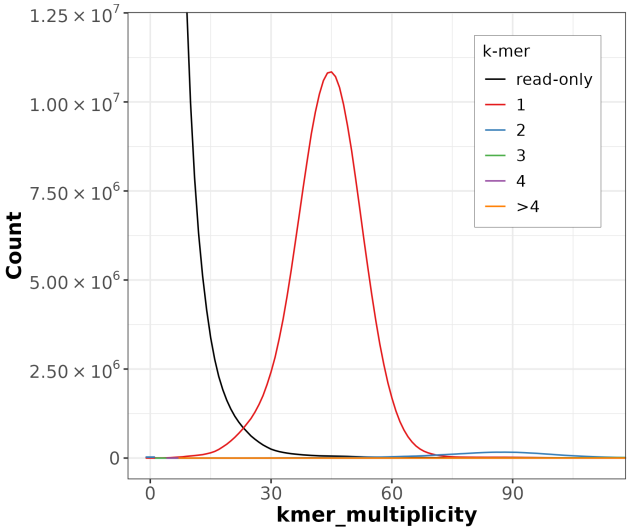
BUSCO: 5.4.3 (euk_genome_met, metaeuk) / Lineage: embryophyta_odb10 (genomes:50, BUSCOs:1614)

HiC contact map of curated assembly

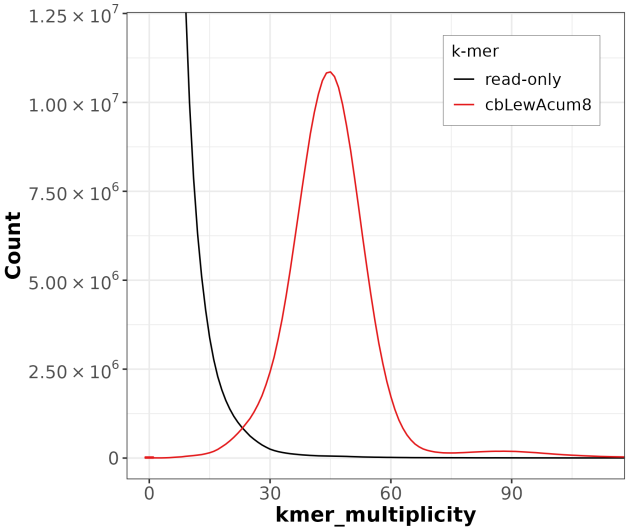


collapsed [LINK]

K-mer spectra of curated assembly

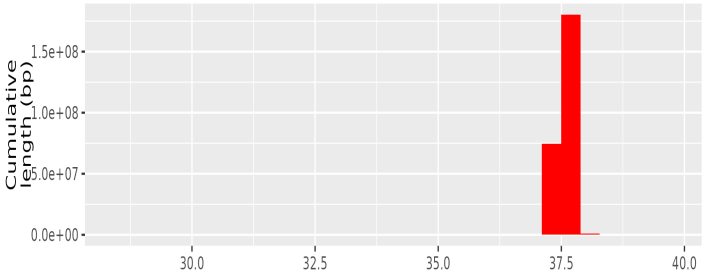


Distribution of k-mer counts per copy numbers found in asm

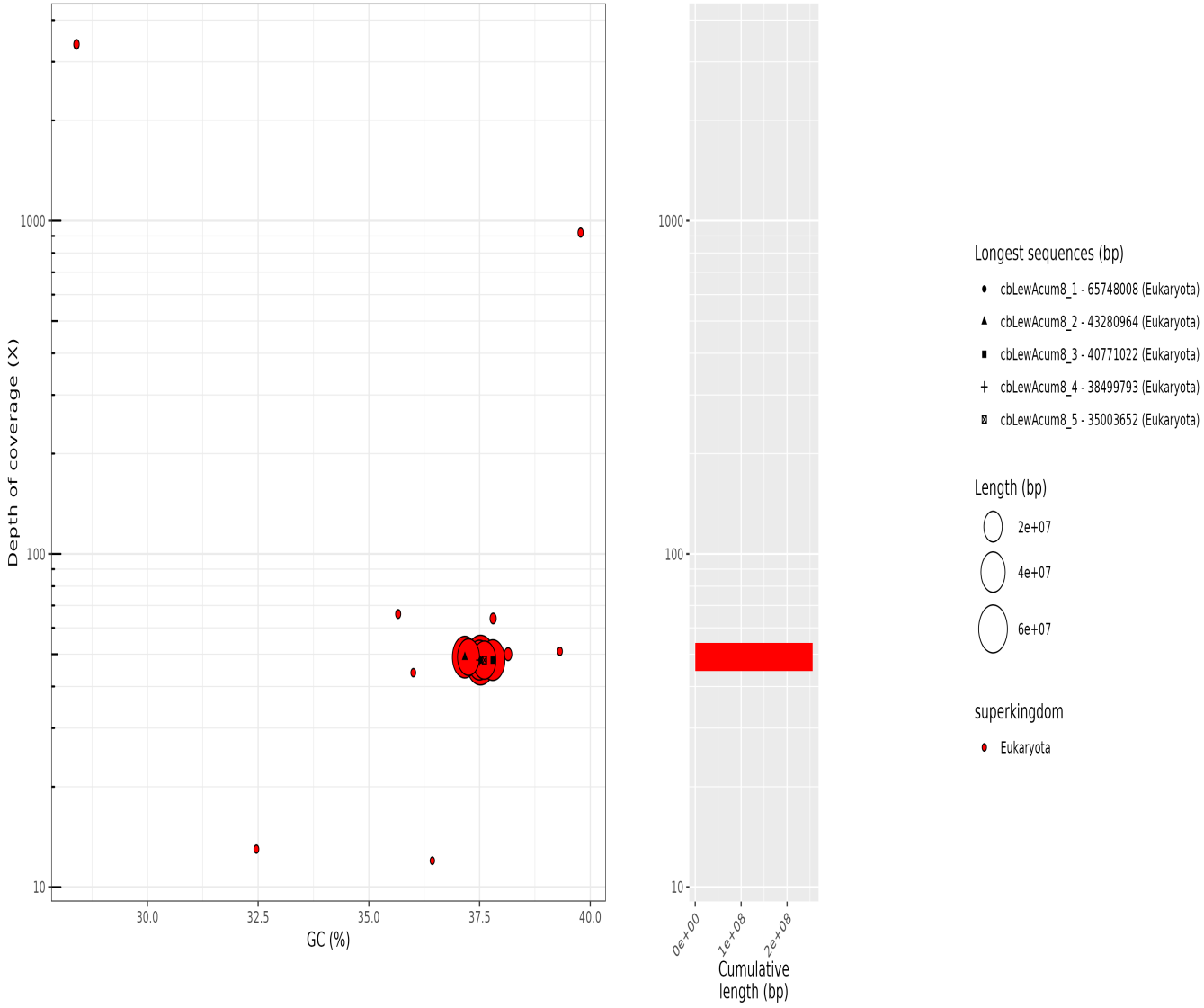


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	Arima
Coverage	69	202

Assembly pipeline

- **Nextdenovo**
  |_ *ver:* 2.5.1
  |_ *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: 2025-02-17 15:43:37 CET