ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3046
ToLID	**ucDunSali1**
Species	Dunaliella salina
Class	Chlorophyceae
Order	Chlamydomonadales

Genome Traits	Expected	Observed
Haploid size (bp)	572,986,696	495,732,735
Haploid Number	7 (source: ancestor)	17
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 6.7.Q44

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
. Assembly length loss > 3% for collapsed
. Not 90% of assembly in chromosomes for collapsed

Curator notes

. Interventions/Gb: 292
. Contamination notes: ""
. Other observations: "The assembly of Dunaliella salina str. RCC6369 (ucDunSali1) is based on 39.99X ONT data and 117.21X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr). The assembly process included the following steps: initial ONT 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, 106 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 30.409 Mb (with the largest being 6.116 Mb). Additionally, 874 regions totaling 42.78 Mb (with the largest being 1.089 Mb) were identified as haplotypic duplications and removed. The mitochondrial genome was assembled using ptGAUL and shows an unusual size compared to currently available reference. The chloroplast genome was assembled using ptGAUL. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, 44 haplotypic regions and 56 contaminant sequences were removed, totaling 12.942Mb and

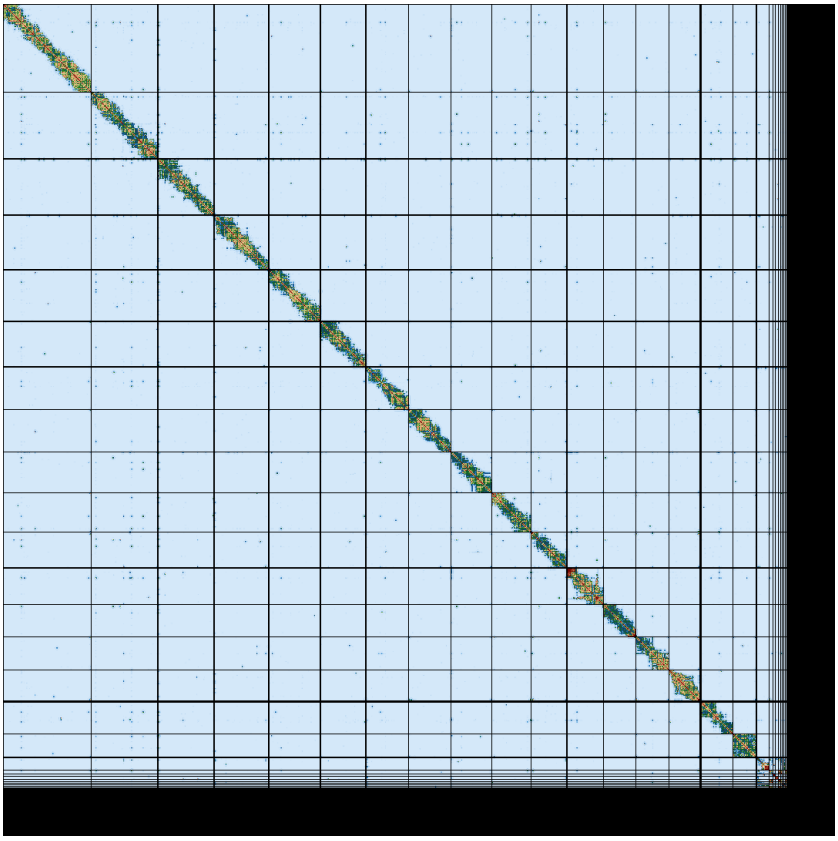
3.096Mb, respectively (with the largest being 1.449Mb and 0.361Mb). Chromosome-scale scaffolds confirmed by Hi-C data were named in order of size. "

Quality metrics table

Metrics	Pre-curation collapsed	Curated collapsed
Total bp	511,693,571	495,732,735
GC %	49.3	49.3
Gaps/Gbp	644.92	728.21
Total gap bp	33,000	43,800
Scaffolds	509	404
Scaffold N50	25,503,345	25,227,183
Scaffold L50	8	8
Scaffold L90	29	18
Contigs	839	765
Contig N50	2,776,926	2,776,926
Contig L50	57	55
Contig L90	249	231
QV	44.6811	44.8542
Kmer compl.	64.0438	63.093
BUSCO sing.	84.8%	86.3%
BUSCO dupl.	2.4%	0.8%
BUSCO frag.	1.5%	1.6%
BUSCO miss.	11.3%	11.3%

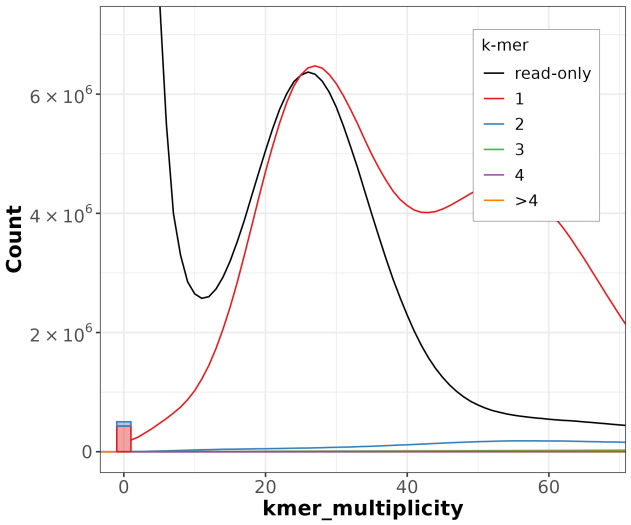
BUSCO: 6.1.0 (euk_genome_min, miniprot) / Lineage: chlorophyceae_odb12.2 (genomes:17, BUSCOs:2709)

HiC contact map of curated assembly

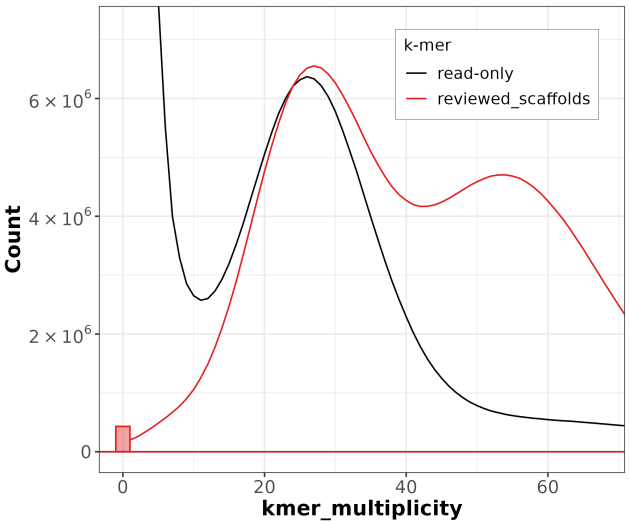


collapsed [LINK]

K-mer spectra of curated assembly

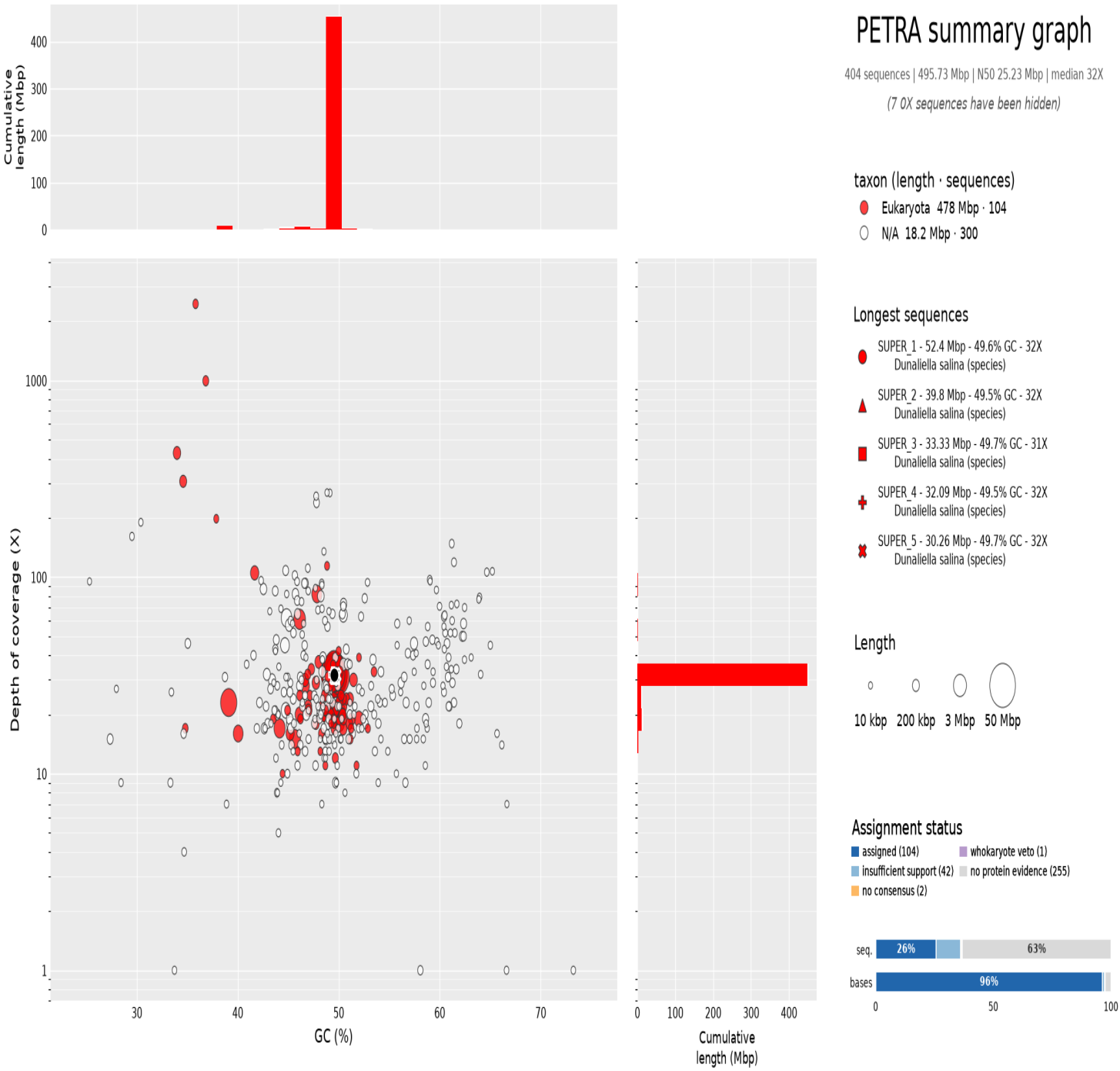


Distribution of k-mer counts per copy numbers found in asm



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

Post-curation contamination screening



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	Arima
Coverage	41	120

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:* 0.2.4
  |_ *key param:* NA
- **PretextView**
  |_ *ver:* 1.0.7
  |_ *key param:* NA

Submitter: Caroline Menguy
Affiliation: Genoscope

Date and time: 2026-10-07 23:52:18 CEST