ERGA Assembly Report

v24.10.15

Tags: ATLASea[INVALID TAG]

TxID	3451511
ToLID	**ucNepSpea1**
Species	Nephroselmis sp. RCC6846
Class	Nephroselmidophyceae
Order	Nephroselmidales

Genome Traits	Expected	Observed
Haploid size (bp)	146,666,756	113,885,183
Haploid Number	7 (source: ancestor)	33
Ploidy	2 (source: ancestor)	2
Sample Sex	Unknown	Unknown

EBP metrics summary and curation notes

Obtained EBP quality metric for collapsed: 5.6.Q37

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

. QV value is less than 40 for collapsed
. Kmer completeness value is less than 90 for collapsed
. More than 1000 gaps/Gbp for collapsed

Curator notes

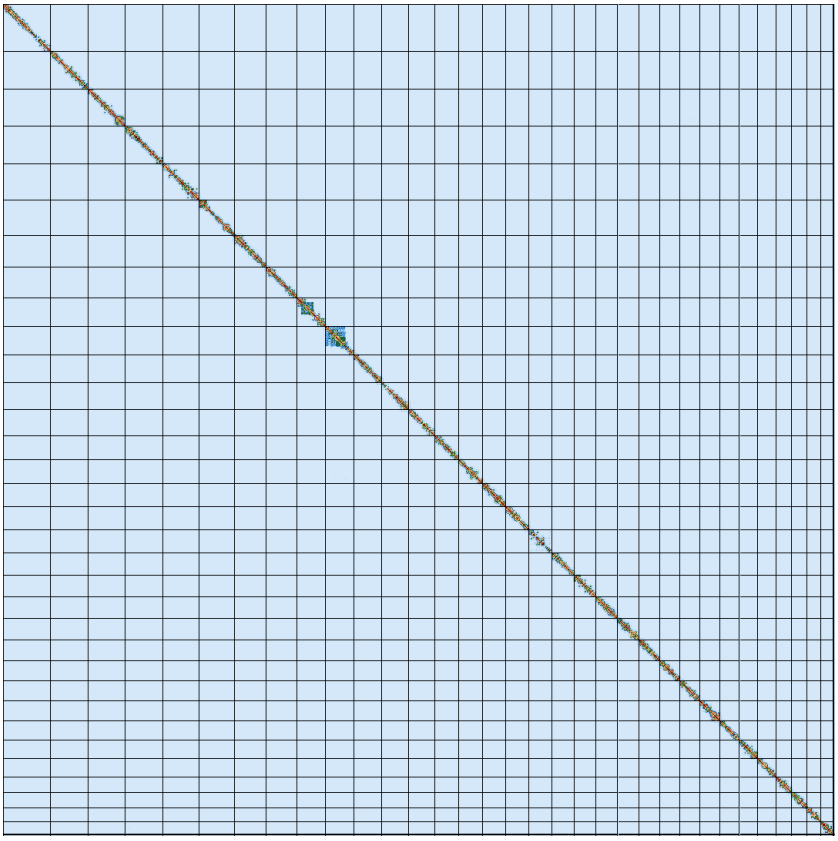
. Interventions/Gb: 686
. Contamination notes: ""
. Other observations: "The assembly of Nephroselmis sp. RCC6846 (ucNepSpea1) is based on 91X PacBio data and 87X Arima Hi-C data generated as part of the ATLASea programme (https://www.atlasea.fr).The assembly process included the following steps: initial PacBio 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, 532 contigs were identified as contaminants (bacterial, archaeal, or viral), totaling 79 Mb (with the largest being 4.2Mb). Additionally, 71 regions totaling 2.5 Mb (with the largest being 0.089 Mb) were identified as haplotypic duplications and removed. The mitochondrial and chloroplastic genomes were assembled using oatk. Finally, the primary assembly was analyzed and manually improved using Pretext. During manual curation, no supplementary haplotypic regions were removed. 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	113,826,063	113,885,183
GC %	63.67	63.66
Gaps/Gbp	3,786.48	4,012.81
Total gap bp	43,100	50,600
Scaffolds	59	38
Scaffold N50	3,451,288	3,568,584
Scaffold L50	11	13
Scaffold L90	27	28
Contigs	490	495
Contig N50	396,784	388,139
Contig L50	78	84
Contig L90	288	295
QV	37.3737	37.3968
Kmer compl.	70.0932	70.1047
BUSCO sing.	92.1%	92.1%
BUSCO dupl.	1.6%	1.7%
BUSCO frag.	1.8%	1.7%
BUSCO miss.	4.5%	4.5%

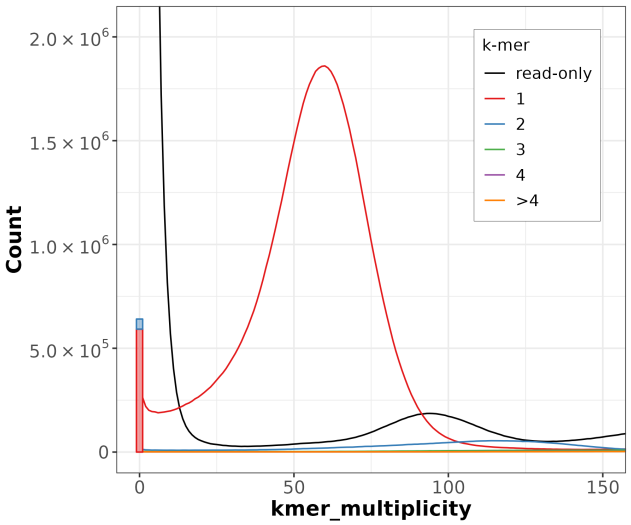
BUSCO: 6.0.0 (euk_genome_min, miniprot) / Lineage: chlorophyta_odb12 (genomes:39, BUSCOs:1523)

HiC contact map of curated assembly

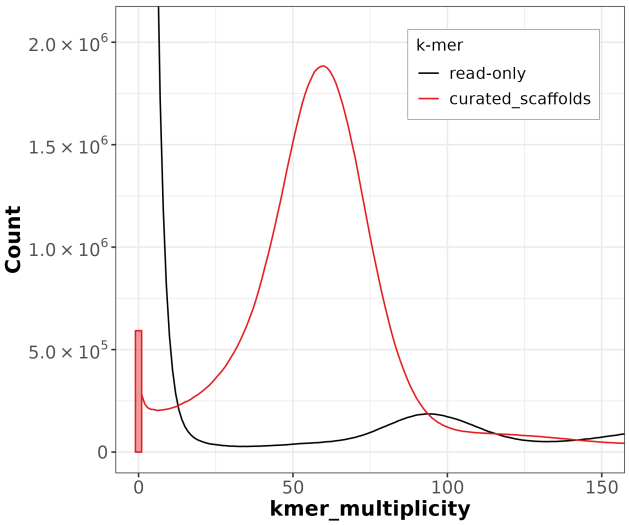


collapsed [LINK]

K-mer spectra of curated assembly

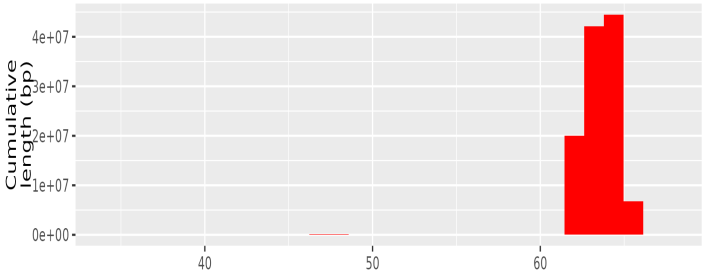


Distribution of k-mer counts per copy numbers found in asm

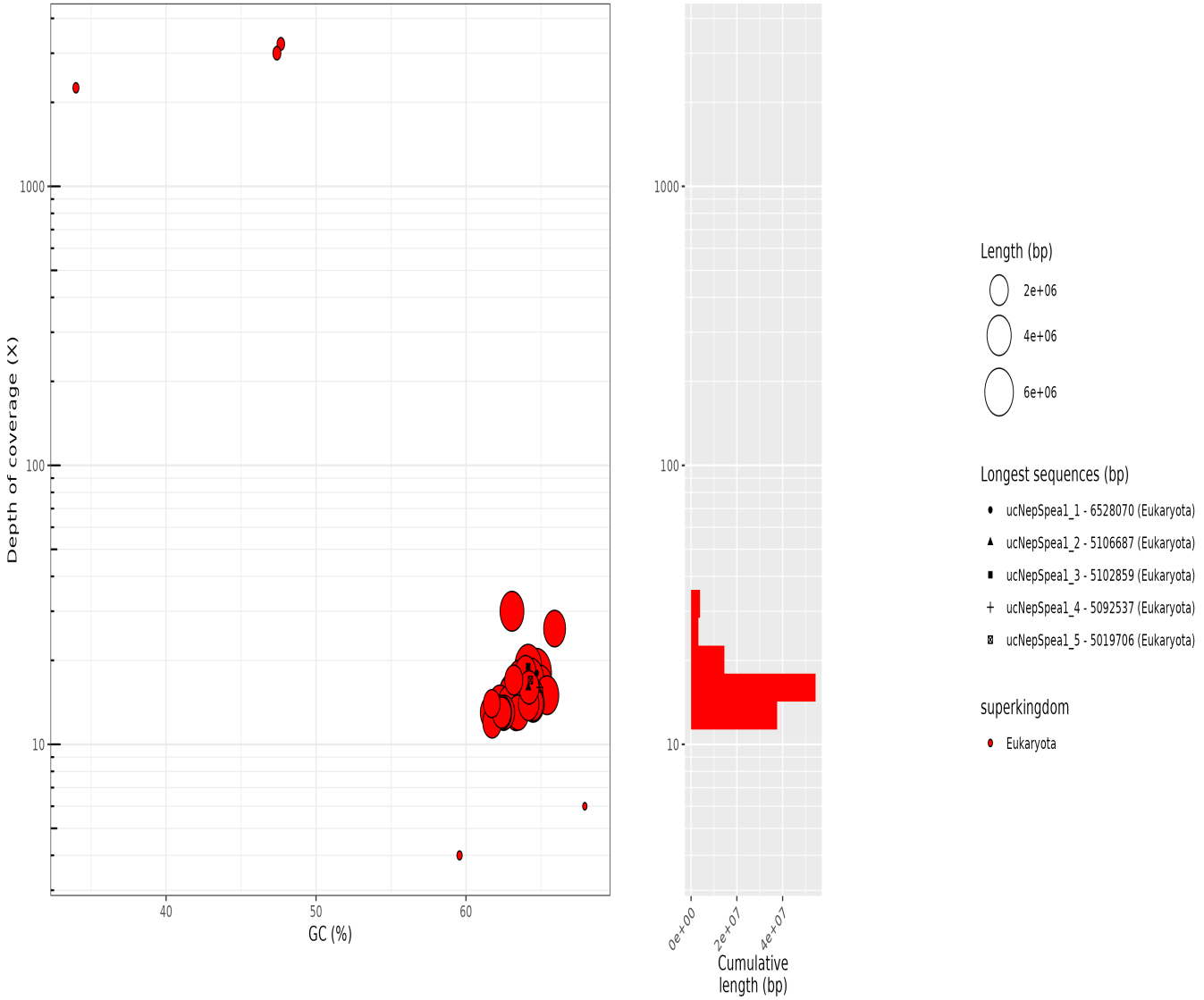


Distribution of k-mer counts coloured by their presence in reads/assemblies

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	Long reads	Arima
Coverage	91	87

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: 2025-10-29 16:42:34 CET