ChloroScan

Environmental plastid genome recovery pipeline

Algae are essenial to our global environment

They are involved in oxygen production, maintaining marine food chain and manufacturing valuable materials for cosmetics and industry.

But we have extremely limited genomic resources to investigate their characteristics, including their plastid genomes. There is also a shortage of computational working pipelines to mine the data.

To mitigate this, we present ChloroScan:
a Snakemake-based working pipeline for recovering plastid genomes from environmental samples known to contain abundant algal plastid sequences.

Step 1

Predict and isolate the plastid contigs via a deep neural network

Step 2

Cluster contigs into plastid genome bins configured by a plastid marker gene database

Step 3

Predict the taxonomy of binned plastid contigs based on predicted amino acid

Step 4

Summarize the former steps' results into multiple format for refinement and predict genes

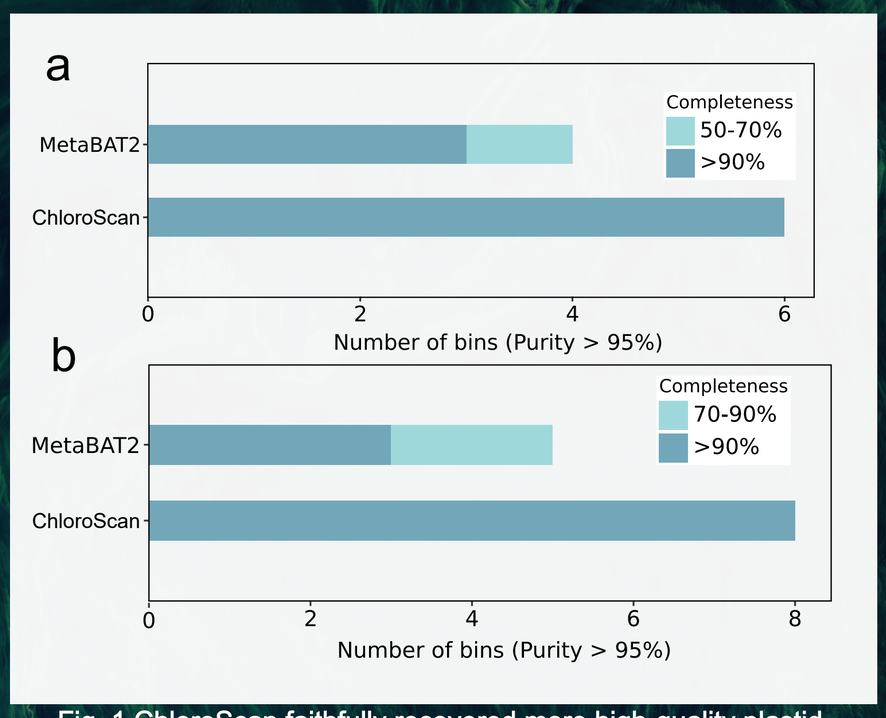

Fig. 1 ChloroScan faithfully recovered more high-quality plastid genomes than the baseline tool MetaBAT2 in two simulated samples.

Performance

We compared the binning (step 2) outcome with a baseline binning tool: MetaBAT2, using two simulated environmental shotgun sequencing datasets (metagenomes), with 12 plastid genomes simulated.

We found ChloroScan outperformed MetaBAT2 in recovering more high-quality genomes (with less contaminations), by incorporating marker gene information, adding more biological meaning to clustering process.

New algae genome identified!

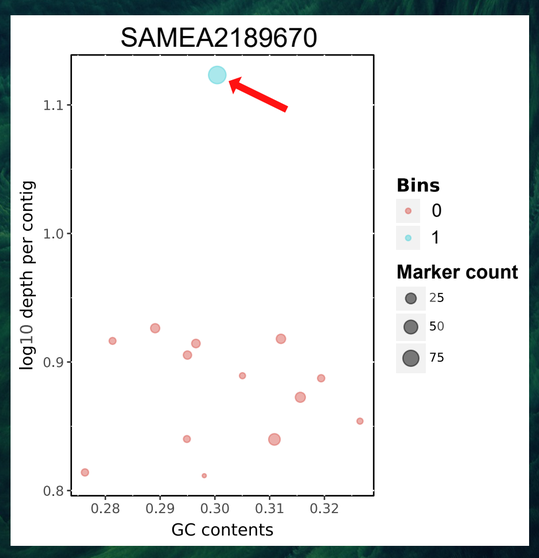

Fig. 2 log-transformed average read depth and GC percentage of each contig in the sample with novel bin (labelled with red arrow).

We ran ChloroScan on 4 real marine metagenomes. In total, we recovered 16 plastid genomes.

One of these is from a plastid with a previously unknown genome from the ochrophyte phylum, with highly contiguous sequences (Fig. 2).

This came from a sample taken just off the coast of Algeria (Fig. 3).

Its closest known organism is a kind of terrestrial alga (Fig. 3, inset).

This genome may belong to an undescribed golden algae lineage.

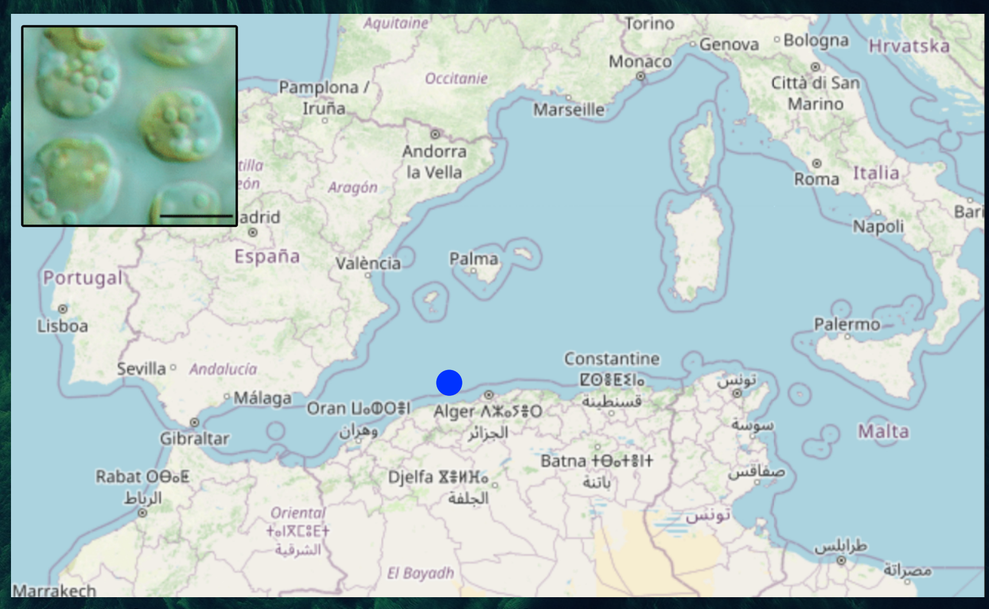

Fig. 3 We found this novel plastid genome from the sample collected by scientists from Tara Oceans expedition near Algeria coast (blue dot). The *rbcL* sequence resembles a terrestrial alga (*kremastochrysopsis austriaca*, inset).

Background: phytoplankton (photosynthetic microorganisms, including algae) bloom in Baltic sea. Credit: European Space Agency.

THE UNIVERSITY OF MELBOURNE

Yuhao Tong[1], Robert Turnbull[2], Vanessa R. Marcelino[1,3], Heroen Verbruggen[1,4]

1. School of Computing and Information System, University of Melbourne
2. Melbourne Data Analytics Platform, University of Melbourne
3. Peter Doherty Institute for Infection and Immunity, University of Melbourne
4. Institute of Agrochemistry and Food Technology, Spanish National Research Council (CSIC), Spain.

Read the preprint
https://bit.ly/4nZ3XnA