Title: plastiC: A pipeline for recovery and characterization of plastid genomes from metagenomic datasets

Authors: Ellen S. Cameron[a,b,*], Mark L. Blaxter[b], Robert D. Finn[a]

[a] EMBL-EBI European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[b] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

* Corresponding Author:
E-mail: escameron@ebi.ac.uk (E.S. Cameron)

Abstract

Summary

Most metagenomic workflows have been developed for identification of prokaryotic and eukaryotic community constituents, but tools for identification of plastid genomes are lacking. *plastiC* is a workflow that allows users to identify plastid genomes in metagenome assemblies, assess completeness, and predict taxonomic association from diverse environmental sources.

Availability and Implementation

plastiC is a Snakemake workflow available at https://github.com/Finn-Lab/plastiC.

Supplementary information

Supplementary information available online.

1. Introduction

Culture-independent sequencing of environmental samples, often referred to as metagenomics, has revolutionised the field of microbial ecology, allowing taxonomic and functional characterization of microbial communities from diverse environmental sources. A variety of bioinformatic tools and pipelines are available for the generation and quality estimation of prokaryotic and eukaryotic metagenomic assembled genomes (MAGs) (Parks *et al.*, 2015; Saary *et al.*, 2020). Organellar genomes are essential to the functioning of eukaryotic organisms but are poorly assessed in current workflows. Plastid (or chloroplast) genomes are of particular interest because of their roles in primary production, yet the diversity of plastids present in metagenomic data is usually ignored and constrains the full understanding of community structure and function. Most existing MAG methods are rooted in single-copy marker gene (SCMG) approaches. However, SCMG approaches are not easily transferred to the genomes of plastid organelles due to reduced genome size, variable gene content (especially driven by organellar-nuclear gene transfer) and large repetitive regions. Thus, alternative methods for quality evaluation are required. Here, we present *plastiC*, a Snakemake (Mölder *et al.*, 2021) workflow for the identification and evaluation of plastids in metagenomic samples available at www.github.com/Finn-Lab/plastiC.

2. Workflow

plastiC is designed to analyse contigs generated from metagenomic assembly workflows (Figure 1). Contigs from these assemblies can be sorted into bins using intrinsic sequence and coverage features, and these bins processed to identify those that putatively correspond to prokaryotic and eukaryotic MAGs. These MAGs are typically processed to identify likely taxonomic affiliation derived from marker gene placement, and their quality (*i.e.* completeness and contamination) assessed. While *plastiC* can be run on raw assemblies, we recommend removing contigs from bins/MAGs determined as "high quality" before analysis. Putative plastid contigs are identified using *Tiara* (Karlicki *et al.*, 2022). The assembly is subsequently (re-)binned using *metaBAT2* (Kang *et al.*, 2019) with a reduced bin size of 50 kb to ensure that putative plastid bins are retained. The resulting bins subsequently are scanned for plastid contigs as identified by *Tiara*, and bins containing more than a user-specified threshold (default: 90%) of plastid sequences are retained as probable plastid bins.

Machine learning algorithms trained on KEGG module completeness were recently demonstrated as an effective approach for estimating metagenomic bin completeness (Chklovski *et al.*, 2022) and we adapted this approach in *plastiC*. *Prodigal* (Hyatt *et al.*, 2010) is used to predict genes in the probable plastid bins, and these predictions are compared to the UniRef100 database (Bateman *et al.*, 2021) using *diamond blastp* (Buchfink *et al.*, 2021) to obtain KEGG annotations (Kanehisa *et al.*, 2016). plastiC calculates KEGG module coverage to estimate completeness (see Supplementary Methods). Due to the presence of shared metabolic pathways and genes between plastid and bacterial chromosomes, bacterial bins may score as highly complete, and this finding supports the requirement to filter assemblies to remove bacterial MAGs before running *plastiC*.

Importantly, *plastiC* also identifies the potential eukaryotic source of the plastid. *plastiC* initially performs a taxonomic classification on the plastid contigs using *CAT* (von Meijenfeldt *et al.*, 2019) and additionally searches for taxonomic marker genes using *hmmer* (for *rbcL*) (Eddy, 2022) and *barrnap* (for rRNA loci) (Seeman, 2022). These loci can be used for further taxonomic analyses. As outputs, *plastiC* produces report files summarising bin

span, estimated completeness, and taxonomic allocation for each putative plastid bin, as well as FASTA files of the marker loci identified.

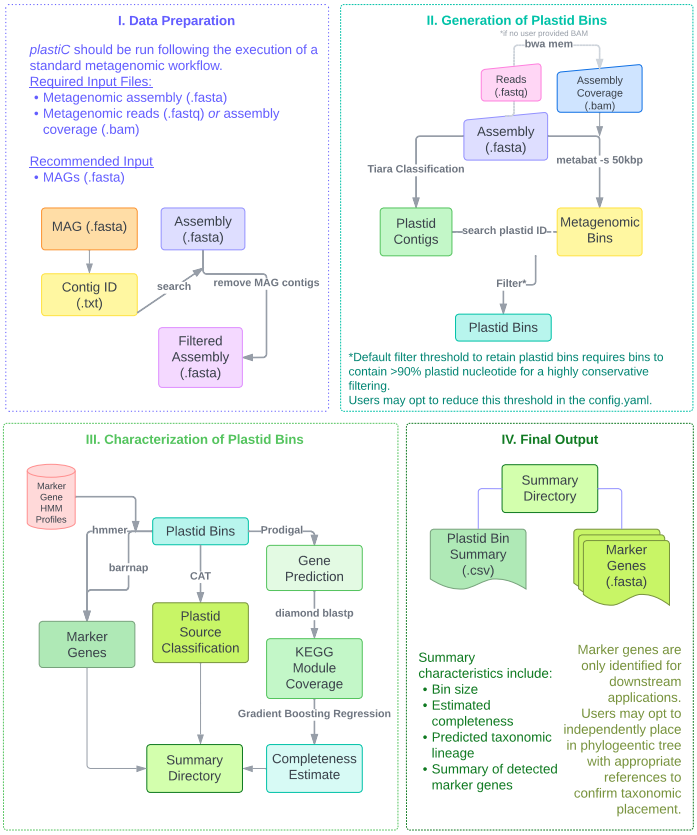

Figure 1: Data preparation and overall workflow for *plastic.* Users provide metagenomic assemblies from which putative plastid bins are identified. A summary report and FASTA files of identified marker loci are generated for each sample.

3. Discussion

Currently, metagenomic characterization of plastid genomes is limited to the identification of plastid contigs using *Tiara*. However, only identifying plastid contigs prevent full disentanglement of metagenomic datasets, including the potential presence of multiple plastid genomes arising from different sources. *plastiC* uniquely provides users with the opportunity to characterize plastid genomes including separation into independent plastid genomes and source classification to inform on present eukaryotic photosynthetic organisms. Notably, this is the first tool to provide a completeness estimate for plastid genomes from metagenomic sources. The high copy number and small genome size of plastids may permit easier recovery and serve as a proxy for the identification of the presence of eukaryotes which may not have sufficient genomic signal to be recovered as a high-quality eukaryotic MAG, as well as the potential to expand reference databases with metagenomic derived plastids. *plastiC* will support characterization of plastids in diverse ecological contexts including, for example, symbiotic interactions (e.g., lichens; see Supplementary Case Study), complex marine ecosystems and characterisation of components of herbivore diets.

4. References

Bateman,A. *et al.* (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, **49**, D480–D489.

Buchfink,B. *et al.* (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*, **18**, 366–368.

Chklovski,A. *et al.* (2022) CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using 1 machine learning 2 3 4. *bioRxiv*.

Eddy, S. (2022) HMMER: Biosequence analysis using profile hidden Markov models. http://hmmer.org

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*.

Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*, **44**, D457–D462.

Kang,D.D. *et al.* (2019) MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **2019**.

Karlicki,M. *et al.* (2022) Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, **38**, 344–350.

von Meijenfeldt,F.A.B. *et al.* (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol*, **20**.

Mölder,F. *et al.* (2021) Sustainable data analysis with Snakemake. *F1000Res*, **10**, 33.

Parks,D.H. *et al.* (2015) CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, **25**, 1043–1055.

Saary,P. *et al.* (2020) Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol*, **21**.

Seeman, T. (2022). barrnap 0.9: rapid ribosomal RNA prediction. https://github.com/tseemann/barrnap