Data in Brief 11 (2017) 273–276

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Reference datasets of *tuf*A and UPA markers to identify algae in metabarcoding surveys

Vanessa Rossetto Marcelino*, Heroen Verbruggen

School of BioSciences, University of Melbourne, Melbourne, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 21 December 2016
Received in revised form
15 January 2017
Accepted 6 February 2017
Available online 13 February 2017

Keywords:
Metabarcoding
Ostreobium
*tuf*A
RDP classifier
UPA
Reference sequences

ABSTRACT

The data presented here are related to the research article "Multi-marker metabarcoding of coral skeletons reveals a rich microbiome and diverse evolutionary origins of endolithic algae" (Marcelino and Verbruggen, 2016) [1]. Here we provide reference datasets of the elongation factor Tu (*tuf*A) and the Universal Plastid Amplicon (UPA) markers in a format that is ready-to-use in the QIIME pipeline (Caporaso et al., 2010) [2]. In addition to sequences previously available in GenBank, we included newly discovered endolithic algae lineages using both amplicon sequencing (Marcelino and Verbruggen, 2016) [1] and chloroplast genome data (Marcelino et al., 2016; Verbruggen et al., in press) [3,4]. We also provide a script to convert GenBank flatfiles into reference datasets that can be used with other markers. The *tuf*A and UPA reference datasets are made publicly available here to facilitate biodiversity assessments of microalgal communities.

Specifications Table

Subject area	*Biology*
More specific subject area	*Metabarcoding*
Type of data	*Text files (DNA sequence data, metadata and python script)*

* Corresponding author.
E-mail address: vrmarcelino@gmail.com (V. Rossetto Marcelino).

http://dx.doi.org/10.1016/j.dib.2017.02.013

How data was acquired	*GenBank data compilation, Amplicon sequencing and Chloroplast genome sequencing*
Data format	*Filtered*
Experimental factors	*Endolithic algae lineages were identified with metabarcoding and chloroplast genome sequencing*
Experimental features	*Genes were extracted from GenBank data, closely related organisms were filtered out and file was converted to a ready-to-use format.*
Data source location	*Melbourne, Australia*
Data accessibility	*The data are available with this article*

Value of the data

- The *tuf*A and UPA reference datasets facilitate biodiversity assessments of cyanobacterial and eukaryotic algal communities using high-throughput sequencing.
- When used with the Naive Bayesian Classifier (RDP classifier) implemented in QIIME [2,5], the taxonomic metadata of the reference datasets provided here allow classifying operational taxonomic units (OTUs) at higher taxonomic ranks when no match is found at lower ranks. For example, an OTU with no close relatives at species or genus level can be classified at the family level, facilitating the interpretation of the results.
- We incorporate in the datasets recently discovered endolithic (limestone-boring) algal lineages [1,3,4] to facilitate the identification of these algae in other studies.
- The script provided here facilitates the development of custom reference databases for non-standard metabarcoding markers.

1. Data

The datasets of this article provide reference sequences of the elongation factor Tu (*tuf*A) and the Universal Plastid Amplicon (UPA) loci and their corresponding taxonomic information. Supplementary File 1 is a set of identified *tuf*A reference sequences in fasta format. Supplementary File 2 is a tab-delimited file containing the taxonomic information of the *tuf*A reference sequences. The *tuf*A reference dataset contains bacterial and chloroplast *tuf*A sequences, including green algae, red algae, heterokonts, cryptophytes and haptophytes. Supplementary File 3 is a set of identified UPA reference sequences (a fragment of the 23S rDNA) in fasta format. Supplementary File 4 is a tab-delimited file containing the taxonomic information of the UPA reference sequences. This reference dataset contains bacterial and chloroplast 23S rDNA sequences, including cyanobacteria, green algae, red algae, heterokonts, cryptophytes and haptophytes. Supplementary File 5 is a python script that takes a GenBank (.gb) flatfile as input and produces the 2 files needed by the RDP classifier (QIIME version). This script requires Biopython [6].

2. Experimental design, materials and methods

We produced reference datasets that can be used with the Naive Bayesian Classifier (RDP classifier) implemented in the QIIME pipeline [2,5]. Each of these datasets consists of: 1) a fasta file containing the reference DNA sequences and short sequence identifiers and 2) a text file matching the sequence identifiers to their taxonomic metadata. To produce these datasets we first mined sequences from GenBank by querying the marker name and downloading all matching items as full GenBank records. We added endolithic (limestone-boring) green algal lineages discovered with the *tuf*A marker in our study “Multi-marker metabarcoding of coral skeletons reveals a rich microbiome and diverse evolutionary origins of endolithic algae” [1]. We identified these algal lineages in a phylogenetic context [see [1]] and included representatives of the main endolithic clades in the *tuf*A reference dataset. We also retrieved a large diversity of algae with the UPA marker but these lineages