Ecological Informatics 33 (2016) 45–50

Contents lists available at ScienceDirect

Ecological Informatics

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

Metaxa2 Diversity Tools: Easing microbial community analysis with Metaxa2

Johan Bengtsson-Palme [a,*], Kaisa Thorell [b], Christian Wurzbacher [c,d], Åsa Sjöling [b], R. Henrik Nilsson [d]

[a] Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, 413 46 Gothenburg, Sweden
[b] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280, 171 77 Stockholm, Sweden
[c] Leibniz-Institute of Freshwater Ecology and Inland Fisheries & Berlin Center for Genomics in Biodiversity Research, Königin-Luise-Str. 6-8, 14195 Berlin, Germany
[d] Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 24 January 2016
Received in revised form 9 April 2016
Accepted 10 April 2016
Available online 23 April 2016

Keywords:
Community similarity
Diversity assessment
Metagenomics
Microbial communities
rRNA
Rarefaction analysis

ABSTRACT

DNA sequencing has become an integrated part of microbial ecology, and taxonomic marker genes such as the SSU and LSU rRNA are frequently used to assess community structure. One solution for taxonomic community analysis based on shotgun metagenomic data is the Metaxa2 software, which can extract and classify sequence fragments belonging to the rRNA genes. This paper describes the Metaxa2 Diversity Tools, a set of new open-source software programs that extends the capabilities of the Metaxa2 software. These tools allow for better handling of data from multiple samples, improved species classifications, rarefaction analysis accounting for unclassified entries, and determination of significant differences in community composition of different samples. We demonstrate the performance of the software tools on rRNA data extracted from different shotgun metagenomes, and find the tools to streamline and improve the assessments of community diversity, particularly for samples from environments for which few reference genomes are available. Finally, we establish that our resampling algorithm for determining community dissimilarity is robust to differences in coverage depth, suggesting that it forms a complement to multidimensional visualization approaches for finding differences between communities. The Metaxa2 Diversity Tools are included in recent versions (2.1 and later) of Metaxa2 (http://microbiology.se/software/metaxa2/) and facilitate implementation of Metaxa2 within software pipelines for taxonomic analysis of environmental communities.

1. Introduction

With DNA sequencing steadily becoming a routine commodity in microbial ecology, and shotgun metagenomics turning into an economically viable approach to study organism diversity across a multitude of environments, the need for robust and ideally open-source software tools for community analysis based on fragmentary sequence data has become increasingly obvious. Taxonomic marker genes are often used to assess community structure, and although other marker genes more suitable for particular organism groups are known (Pawlowski et al., 2012; Schoch et al., 2012; Wang et al., 2015), the ribosomal small subunit (SSU; 16S/18S) and large subunit (LSU; 23S/28S) rRNA genes remain frequently used instruments for studying diversity in all organism domains (Glaeser and Kämpfer, 2015). To extract these sequences from shotgun metagenomic data can be cumbersome and prone to error, and it is furthermore important to keep archaeal, bacterial, mitochondrial, and chloroplast rRNA separately to avoid biasing results. To remedy this problem, the software tool Metaxa (Bengtsson et al., 2011) was developed, and it has found use in metagenomic analysis (Duguma et al., 2013; Sanli et al., 2015; Yau et al., 2013), quality control (Ghai et al., 2012; Lühring et al., 2015), and full-length rRNA reconstruction (Fan et al., 2012). Recently, Metaxa was updated with a new taxonomic classifier, focused on resolving taxonomic affiliations to the level of precision supported by the data, but not further. Thus, some sequences will be identified to species level, other to the genus level, and so on (Bengtsson-Palme et al., 2015b). That update – Metaxa2 – brought full-scale classification capabilities to the software, but still required the user to perform a range of tasks manually or in other software packages, such as collection of taxonomic data from several samples, rarefaction analysis, and assessment of community differences. In this paper, we introduce the Metaxa2 Diversity Tools, a set of additional tools bundled with Metaxa2 to facilitate such downstream data analyses.

2. Material and methods

The Metaxa2 Diversity Tools comprise four open-source programs written in Perl for automating and improving the analysis of taxonomy output from Metaxa2. They use the taxonomy output files from Metaxa2 or the output from the Metaxa2 Taxonomic Traversal Tool (metaxa2_ttt) as input, and are intended as an extension to the standard Metaxa2 analysis (Fig. 1). The four tools are the Data Collector (metaxa2_dc), the Species Inference tool (metaxa2_si), the Rarefaction

* Corresponding author. Tel.: +46 31 342 46 26.
E-mail address: johan.bengtsson-palme@microbiology.se (J. Bengtsson-Palme).

http://dx.doi.org/10.1016/j.ecoinf.2016.04.004
1574-9541/

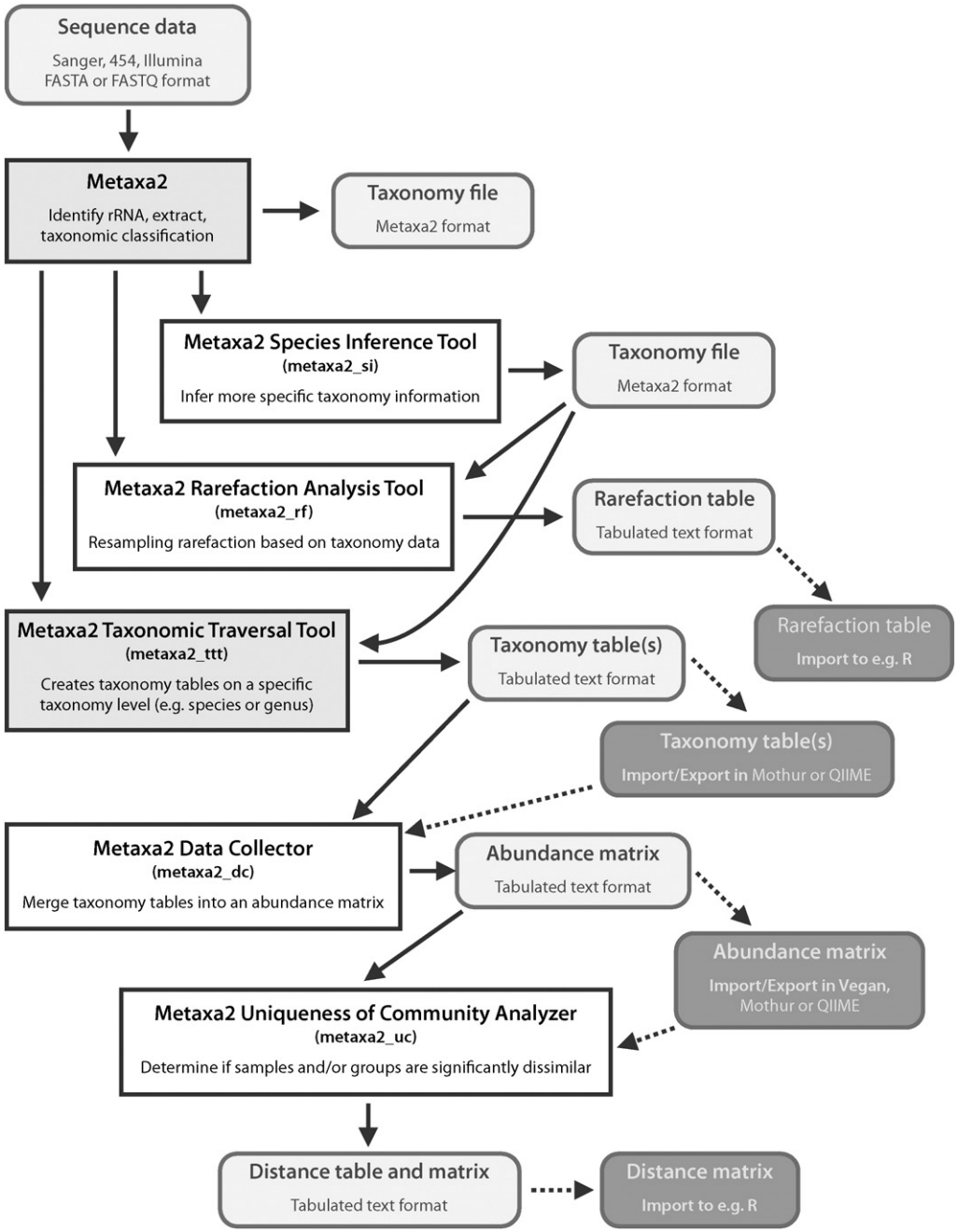

Fig. 1. Flowchart showing an example of how the Metaxa2 Diversity Tools can be used in the analysis of microbial community data. Rectangles with corners correspond to analysis steps and rounded rectangles represents input and output data. Analysis steps with white background are new in the Metaxa2 Diversity Tools. Dark gray rectangles represent data that can easily be imported or exported to/from other software packages.

analysis tool (metaxa2_rf), and the Uniqueness of Community analyzer (metaxa2_uc). The algorithms underlying these tools are outlined below.

2.1. Metaxa2 Data Collector

The Metaxa2 Data Collector merges the output of several samples processed using the Metaxa2 Taxonomic Traversal Tool into one large abundance matrix, suitable for further analysis in, for example, the R package Vegan (Oksanen et al., 2011). The Metaxa2 Data Collector tool is by far the simplest of the four, since it only converts several files containing taxon abundances of one sample each into one large table containing the same information but for all samples, making sure that taxa not present in one or more samples are represented by zeros in the final table.

2.2. Metaxa2 Species Inference tool

The Metaxa2 Species Inference tool uses the taxonomic information from other entries to better infer taxon information on, for example, the species level for entries whose taxonomic affiliation at the corresponding level could not be establish in a way satisfying the threshold cutoffs of the Metaxa2 classifier (Bengtsson-Palme et al., 2015b). It operates under the assumption that if only one species in the dataset has been detected in, e.g., the Flavobacteriaceae family (say *Ornithobacterium rhinotracheale*) and a particular entry subsequently is assigned to the Flavobacteriaceae family, but not classified to the species level, that sequence will also be