Reference-based error correction of amplicon sequencing data from synthetic communities

Pengfan Zhang[1], Stjin Spaepen[3], Yang Bai[4], Stephane Hacquard[1,2], Ruben Garrido-Oter[1,2]

1 Department of Plant-Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.

2 Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, Cologne, Germany.

3 CMPG Laboratory of Genetics and Genomics, Department M2S, KU Leuven, Gaston Geenslaan 1, 3001, Leuven, Belgium.

4 State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, 100101 Beijing, China.

Correspondence to: garridoo@mpipz.mpg.de

Abstract

Motivation: Synthetic microbial communities (SynComs) constitute an emergent and powerful tool in biological, biomedical, and biotechnological research. Despite recent advances in algorithms for analysis of culture-independent amplicon sequencing data from microbial communities, there is a lack of tools specifically designed for analysing SynCom data, where reference sequences for each strain are available.

Results: Here we present Rbec, a tool designed for analysing SynCom data that outperforms current methods by accurately correcting errors in amplicon sequences and identifying intra-strain polymorphic variation. Extensive evaluation using mock bacterial and fungal communities show that our tool performs robustly for samples of varying complexity, diversity, and sequencing depth. Further, Rbec also allows accurate detection of contaminations in SynCom experiments.

Availability: Rbec is freely available as an open-source R package and can be downloaded at: https://github.com/PengfanZhang/Microbiome.

Introduction

Amplicon sequencing is a powerful technique to characterize microbial composition from environmental samples. Recent advances in algorithms and tools for the analysis of marker gene amplicon data have driven a shift from clustering approaches (Van Leuven *et al.*, 2014), based on operational taxonomic units (OTUs) and arbitrary sequence similarity thresholds, to error correction methods (Callahan *et al.*, 2016; Edgar and Flyvbjerg, 2015; Amir *et al.*, 2017; Peng and Dorman, 2020) that seek to estimate abundances of individual amplicon sequence variants (ASVs). A new generation of integrated pipelines (Bolyen *et al.*, 2019) allows researchers from a variety of fields in the environmental, biological, and medical sciences to reproducibly analyse marker gene sequencing data.

Synthetic microbial communities (SynComs) constitute an emerging and powerful tool to build experimentally tractable, reproducible microbial systems in the laboratory that enable controlled perturbation experiments and testing of falsifiable hypotheses. These reductionist approaches are being increasingly employed in studies of microbial ecology and evolution (Cairns *et al.*, 2020), plant and animal microbiota (Bai *et al.*, 2015; Vorholt *et al.*, 2017; Zhang *et al.*, 2019; Vrancken *et al.*, 2019), and biotechnology (McCarty and Ledesma-Amaro, 2019). A factor limiting these new experimental approaches from developing to their full potential is the lack of bioinformatic tools specifically designed for the analysis of sequencing data obtained from gnotobiotic systems and SynComs. As a result, researchers typically employ standard clustering (Carlström *et al.*, 2019), error correcting (Pauvert *et al.*, 2019) or mapping approaches (Durán *et al.*, 2018) that do not take full advantage of these experimental systems (e.g. reduced complexity and the availability of reference sequences for classification), resulting in reduced resolution, accuracy or data loss. To address this limitation, we developed a **r**eference-based **e**rror **c**orrection **a**lgorithm, called Rbec, that is able to accurately and precisely correct errors in SynCom amplicon data, identify intra-strain polymorphism, and detect the presence of contaminants in gnotobiotic systems.

Methods

The Rbec algorithm identifies and corrects sequencing errors by implementing an adapted version of the quality-aware model implemented in the DADA2 tool (Callahan *et al.*, 2016) . Rbec is specifically designed to efficiently and accurately process data from SynComs, for which reference sequences of individual community members are available (Fig. S1). First, reads are de-replicated into unique tags and subsequently aligned to the reference database,

after which initial abundances are assigned to each strain according to the copy number of each exactly aligned tag. Next, tags that are not exactly matched to any sequence in the database are assigned a candidate error-producing reference based on *k*-mer distances. Sequencing reads are then subsampled and an error matrix is calculated using the mapping between subsampled reads and candidate error-producing sequences. The probability that a unique tag is erroneously produced by a given candidate error-producing sequence is calculated using a *Poisson* distribution. The probability and expectation values of this distribution are then used to determine whether a unique tag can be corrected from a reference sequence, and tags that cannot be corrected are subsequently removed. The parameters of the error model are recomputed iteratively until the number of re-assignments falls below a set threshold. Strain abundances are then estimated from the number of error-corrected reads mapped to each reference sequence. Finally, putatively contaminated samples are determined by a significant deviation from the expected proportion of corrected reads.

Results

To validate the performance of Rbec, we used mock samples generated using a taxonomically wide set of bacterial and fungal strains obtained from sequenced culture collections derived from the *Arabidopsis thaliana* microbiota (Bai *et al.*, 2015; Durán *et al.*, 2018). Mapping of amplicon reads to the reference sequences showed that only 31.8% of all reads per sample on average had a perfect match in the database, indicating the presence of extensive sequencing and PCR errors (Fig. 1a, Fig. S2). Our implementation of the Rbec algorithm successfully corrected most erroneous reads (89.2% on average), out-performing all other tested *de novo* correction methods (Fig. 1b). This improvement was most pronounced for reads generated from polymorphic copies of the marker sequences within a single strain, owing to the fact that Rbec is capable of correctly classifying paralogous sequences (Fig. S3). To evaluate the accuracy of Rbec in characterizing community composition, we simulated *in silico* bacterial and fungal mock samples by mixing reads generated from sequencing individual isolates separately. For these simulations, we varied community complexity, strain similarity and sequencing depth. Across these three parameters, Rbec consistently performed better than all other tested methods in characterizing microbial composition in terms of deviation from the ground truth (Figs. 1c, S4 and S5), as well as precision and recall (Fig. S6), and was capable of detecting comminated samples (Fig. S7).

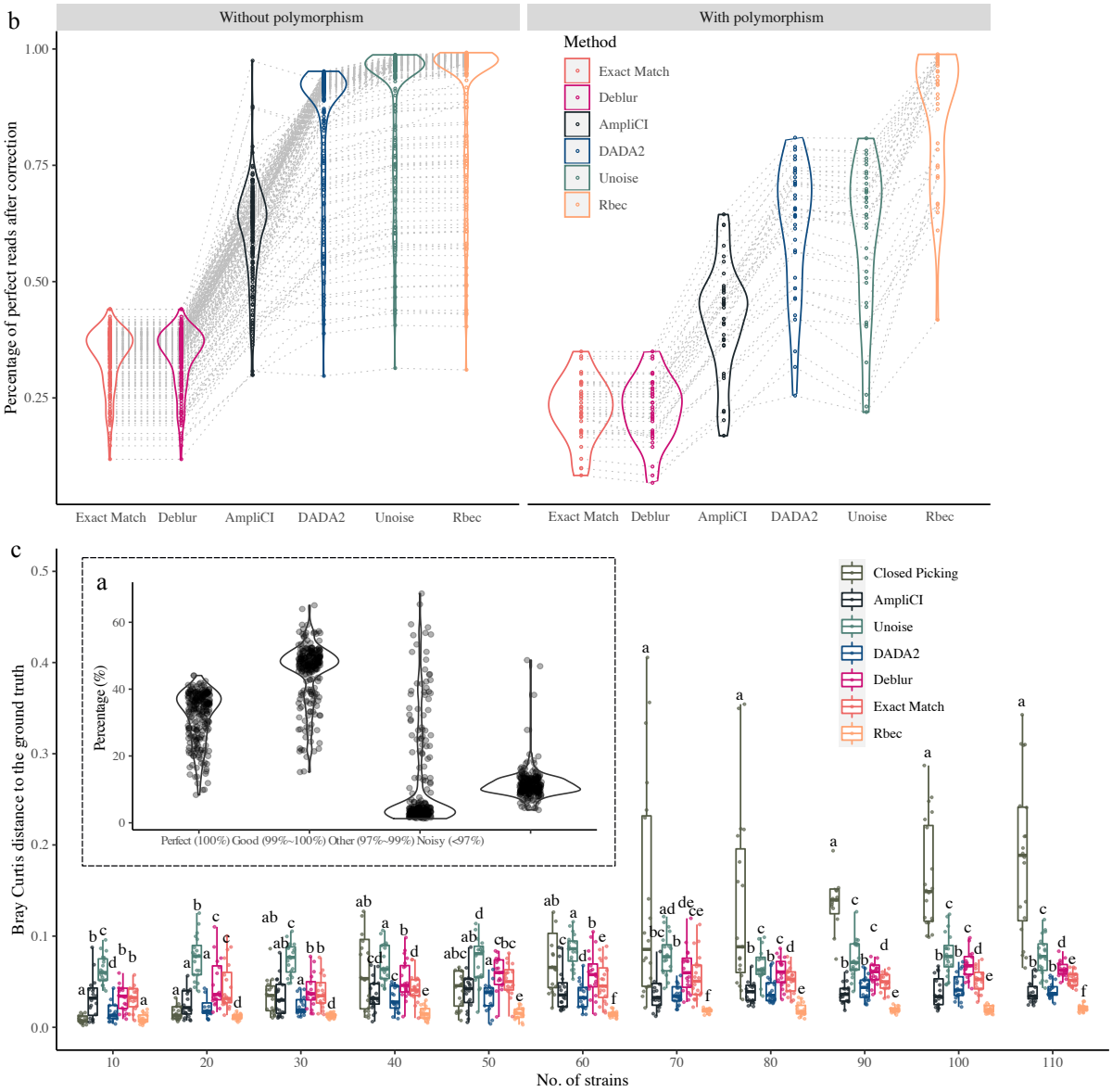

Fig 1. Evaluation of the Rbec algorithm. (**a**) Error profiles of amplicon sequencing data from 234 bacterial strains sequenced on Illumina platform. (**b**) Percentage of perfectly aligned reads after correction with different methods. (**c**) Evaluation of the influence of community complexity on the performance of different methods, measured as a deviation from the ground truth using Bray-Curtis dissimilarities. The x-axis represents the number of strains in *in silico* simulated mock samples. Letters indicate the significant groups (paired Wilcox test, $P < 0.05$) within mocks with the same complexity.

Acknowledgements

We would like to acknowledge the useful feedback provided by Prof. Alga Zuccaro, Prof. Eric Kemen, and Dr. Yulong Niu during the development of this algorithm. We also would like to thank Anna Lisa Roth for her help in generating amplicon data from individual strains.

Funding

This research was funded by the Max Planck Society and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-Nummer 2048/1– project 390686111 and the '2125 DECRyPT' Priority Programme through funds to R. G.-O. and P. Z.

References

Amir, A. *et al*. (2017) Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*, **2**.

Bai, Y. *et al*. (2015) Functional overlap of the Arabidopsis leaf and root microbiota. *Nature*, **528**, 364–369.

Bolyen, E. *et al*. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.

Cairns, J. *et al*. (2020) Repeatable ecological dynamics govern the response of experimental communities to antibiotic pulse perturbation. *Nat. Ecol. Evol.*, **4**, 1385–1394.

Callahan, B.J. *et al*. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583.

Carlström, C.I. *et al*. (2019) Synthetic microbiota reveal priority effects and keystone strains in the Arabidopsis phyllosphere. *Nat. Ecol. Evol.*, **3**, 1445–1454.

Durán, P. *et al*. (2018) Microbial Interkingdom Interactions in Roots Promote Arabidopsis Survival. *Cell*, **175**, 973-983.e14.

Edgar, R.C. and Flyvbjerg, H. (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, **31**, 3476–3482.

McCarty, N.S. and Ledesma-Amaro, R. (2019) Synthetic Biology Tools to Engineer Microbial Communities for Biotechnology. *Trends Biotechnol.*, **37**, 181–197.

Pauvert, C. *et al*. (2019) Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecol.*, **41**, 23–33.

Peng, X. and Dorman, K.S. (2020) AmpliCI: a high-resolution model-based approach for denoising Illumina amplicon data. *Bioinformatics*.

Van Leuven, J.T.T. *et al*. (2014) Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. *Cell*, **158**, 1270–1280.

Vorholt, J.A. *et al*. (2017) Establishing Causality: Opportunities of Synthetic Communities for Plant Microbiome Research. *Cell Host Microbe*, **22**, 142–155.

Vrancken, G. *et al*. (2019) Synthetic ecology of the human gut microbiota. *Nat. Rev.*

Microbiol., **17**, 754–763.

Zhang, J. *et al*. (2019) NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. *Nat. Biotechnol.*, **37**, 676–684.