Title: DISCo: Design of an identifiable synthetic community

Authors: Dana L. Carper1, Travis J. Lawrence1, Alyssa A. Carrell2, Dale A. Pelletier1, Dave J. Weston1

Affiliations:

1 Biosciences Division, Oak Ridge National Laboratory, Oak Ridge TN, USA

2 Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN, USA

**Abstract:**

Background:

Results:

We describe DISCo, a program for the design of an identifiable synthetic community. The program is broken up into three modules; (1) create, to generate a high diverse community from an input DNA sequence alignment, (2) correct, to correct sequencing errors using the designed community as a reference and (3) subset, to create a subset from the original diverse community either by the number of strains or by a taxonomic proportion.

Conclusions:

DISCo can create a highly diverse community of microbes that can be distinguished through 16S rRNA sequencing with the ability to correct sequencing errors to identify the original community members. Although designed for bacteria, the program allows for any alignment input from any organism group allowing for broad application. The program and data from this paper are freely available through github (https://github.com/dlcarper/DISCo-Design-of-an-identifiable-synthetic-community), conda or pip installation.

**Background:**

Multicellular eukaryotes are associated with complex communities of microorganisms. These host associated communities provide many beneficial functions for the host organism including, nutrient acquisition, pathogen defense, and moderate abiotic stresses. Although, DNA sequencing-based approaches have given insight into determining taxonomically who is present in the community and functional potential of that community but interpretation and reconciling of these results can be extremely difficult. For this reason, some researchers have shifted approaches to synthetic or constructed based microbial communities’ experiments to unravel host-microorganism interactions. This method involves adding known microbial strains to a host organism allowing for an easier interpretation of colonization, shifts in community structure, microbe-microbe interactions and with genomic information of the strain’s potential functions. However, there is no clear method for how to design a reproducible microbial community for these experiments.

Designing of a microbial community can be carried out by many methods. Some researchers, primarily in microbial engineering have focused on functional assets of the microbes in order to engineer specific microbial outputs. While this method works well for bio-engineering purposes it may not be as informative for microbiome studies. Host associated synthetic communities have primarily consisted a very few strains and confirmed through re-isolation, limiting the extrapolation and interpretation in more diverse communities. Designing a diverse community that is reproducible and can be distinguished through cheaper sequencing technologies has been extremely difficult without strong computational abilities.

In this paper, we describe an easy to use command-line software program, DISCo, for the creation of a reproducible diverse community of organisms that can be distinguished through DNA sequencing technology. Currently, this program operates in three modules; (1) create, (2) correct and (3) subset. The create module will take an input alignment of sequences and create the most diverse community possible at a specified sequence difference level. The program can either start with no previous knowledge or with an initial input community of organisms of interest. The correct module has sequence error correcting ability for DNA sequences for interpretation after synthetic community experimentation. The subset module will output a subset of the community either by user specification of the number of strains or by user specified proportions of taxonomic group. While this program was initial designed for bacterial community construction, any DNA alignment can be used and thus users are not restricted to specific organismal group. This program is implemented in python and is available through github, conda or pip installation.

Implementation:

FIGURE 1 (Diagram of workflow)

Results and Discussion:

As an example, the Ribosomal database project [1] file of aligned 16S rRNA was downloaded and uncultured strains were filtered using fasgrep [2]. The alignment was trimmed to the V4 region which is commonly used region for next generation sequencing of bacterial communities [3]. The alignment was used to create a highly diverse community at an edit distance of 3. The community construction took ### to construct. The resulting community consisted of ### strains that could be distinguished through 16S rRNA sequencing of the V4 region.

FIGURE 2 (Community diversity)

FIGURE 3 (Error correction accuracy)

References:

1. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42:633–42.

2. Lawrence TJ, Kauffman KT, Amrine KCH, Carper DL, Lee RS, Becich PJ, et al. FAST: FAST Analysis of Sequences Toolbox. Front Genet. 2015;6.

3. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature. 2017;551:457–63.

Availability and requirements

Project name: DISCo: Design of an identifiable synthetic community

Project home page: <https://github.com/dlcarper/DISCo-Design-of-an-identifiable-synthetic-community.git>

Operating system(s): Platform independent

Programming language: Python

Other requirements: None

License:

Abbreviations:

DNA: Deoxyribonucleic acid

rRNA: Ribosomal ribonucleic acid

FASTA: Fast-all (file format)