BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 4 2015, pages 602–603
doi:10.1093/bioinformatics/btu693

Genetics and population analysis

Advance Access publication October 20, 2014

Seed: a user-friendly tool for exploring and visualizing microbial community data

Daniel Beck*, Christopher Dennis and James A. Foster

Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: In this article we present Simple Exploration of Ecological Data (Seed), a data exploration tool for microbial communities. Seed is written in R using the Shiny library. This provides access to powerful R-based functions and libraries through a simple user interface. Seed allows users to explore ecological datasets using principal coordinate analyses, scatter plots, bar plots, hierarchal clustering and heatmaps.

Availability and implementation: Seed is open source and available at https://github.com/danlbek/Seed.

Contact: danlbek@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on August 6, 2014; revised on October 7, 2014; accepted on October 16, 2014

1 INTRODUCTION

The proliferation of microbial community profiling is allowing researchers to study microbial communities in new ways. Increasingly, researchers in diverse fields are asking questions relating to how microbial communities vary across samples. For example, researchers studying the human microbiome are interested in how microbial composition changes across body sites and through time (HMP Consortium *et al*., 2012). Researchers studying disease look at how microbial communities differ between samples from healthy and unhealthy individuals (Srinivasan and Fredricks, 2009). It is now standard practice to use cultivation independent high-throughput sequencing to identify the microbial composition of many samples. This produces a wealth of data about microbial composition in many different environments and conditions.

In conjunction with advances in sequencing resources, researchers have developed a number of powerful software tools to analyze and visualize this wealth of data. Packages such as mothur (Schloss *et al*., 2009) and Qiime (Caporaso *et al*., 2010) aggregate many tools to allow researchers to quickly and efficiently process large sequencing datasets. These currently available packages excel at performing robust, computationally intensive calculations that attempt to minimize the effects of noise and sequencing artifacts on downstream analyses. They often use a non-visual interface for analysis, even when they provide a graphical user interface for their own functions, requiring the user to know specific command and parameter combinations. While this setup is ideal for pipeline development, it is often a hindrance for data exploration.

Simple Exploration of Ecological Data (Seed) fills a currently unmet need for a tool that allows researchers to quickly and easily visualize and explore the data that results from these pipelines. This so-called exploratory data analysis has an 'important place in the toolbox of ecologists' (Borcard *et al*., 2011). Though there are texts that recommend specific exploratory techniques (Borcard *et al*., 2011; Legendre and Legendre, 2012), we know of no tool such as Seed that bundles appropriate tools into an easy-to-use system for non-programmers.

In this article, we present Seed, a software package that focuses on data exploration and visualization of microbial community data derived from high-throughput sequencing.

2 SEED SOFTWARE

Seed is an open-source application that allows researchers to visually explore microbial community data. It is designed to allow many different analyses and visualizations including principal component and coordinate analysis (PCA/PCoA), hierarchal clustering, scatter plots, bar plots and heatmaps. These plots allow users to visualize similarities and differences among samples and how environmental and microbial features vary across samples.

Seed is written in the R programming language (R Core Team, 2013) using the Shiny framework (RStudio Inc., 2013). R is open source and available for Linux, MacOS and Windows operating systems. The use of R allows us to take advantage of the wealth of R packages available for complex analyses and visualizations.

Seed is a web-based application, which may be installed locally or hosted on a remote server. When running Seed from a central server, users can access it through a web browser and are not required to install it locally. This means non-expert users can quickly and easily begin using Seed, even without local installations of R. Additionally, updates to R, Shiny, Seed and underlying packages can be done seamlessly and invisibly to the end user. The use of a web browser also provides a familiar interface to most users, allowing them to quickly and easily learn to use Seed. The user interface for seed can be seen in Figure 1.

Currently, Seed requires two types of data, microbial abundance data and sample metadata. The microbial abundance data contain counts or abundances of each microbial taxon in each sample. The sample metadata contain information about each sample, for example the sample pH or temperature. Seed allows the user to modify the abundance data using a number of common transformations including presence/absence, relative

*To whom correspondence should be addressed.

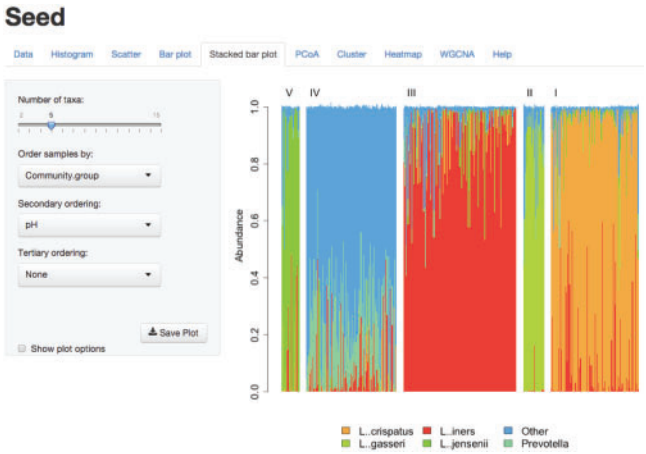

Fig. 1. This figure shows Seed's simple web-based interface. The stacked bar plot shown here is based on data originally published by Ravel *et al.*

abundance and Hellinger transformations. Seed is not limited to microbial data, though that was our primary research domain. It can be used to explore any data that include both feature counts and values for response variables.

Once the user has imported and verified their dataset, they may easily explore their data with many plot types. Examples of some of the plots generated by Seed are shown in the supplementary information. Many of the plots include options to incorporate sample information by coloring points or bars according to metadata values. This allows users to easily visualize the relationship between the sample metadata and the structure of the microbial communities present in the samples.

The design of Seed emphasizes simplicity over exhaustive inclusion of parameters. In many or most cases, researchers will use Seed to understand general trends in the data, which may then inform more specialized analyses. Seed is designed to quickly explore ecological datasets and to act as a hypothesis-generating tool. Publication quality figures and polished analyses are beyond the current scope of this project, though Seed can output all plots in pdf or png format. Additionally, large dataset analysis may be too slow for a comfortable user experience. Note, however, that we used published microbiome and patient data with nearly 400 samples and 250 taxa (Ravel *et al.*, 2011) on a standard laptop while preparing this publication. Seed is certainly capable of handling datasets with hundreds of samples and more than a thousand taxa.

As with any software package, not all analyses have been implemented in Seed. We encourage users to also consider other visualization tools including phyloseq (McMurdie and Holmes, 2013) for analyses incorporating phylogenetic relationships and EMPeror (Vázquez-Baeza *et al.*, 2013) for PCoA analyses of very large datasets. Additionally, while Seed provides some guidance for users, tool selection and result interpretation still relies on user expertise.

3 FUTURE DIRECTIONS

Seed is freely available at https://github.com/danlbek/Seed. Development of Seed is ongoing. We are continuing to add new visualizations and to improve existing ones. Future development will focus on adding phylogenetic and taxonomic data structures, which will allow for analyses that take microbial relationships into account. We welcome user contributions to the project and encourage labs to copy and modify the code to suit their own needs.

ACKNOWLEDGEMENTS

We thank Larry Forney, Roxana Hickey, Janet Williams and other users for helpful conversations, recommendations and bug reports and for the datasets used for the figures herein.

Funding: This work was supported by the National Institutes of Health (P20GM016454) and by the National Science Foundation (DBI0939454). Computational support provided by National Institutes of Health (P20GM16448).

Conflict of interest: none declared.

REFERENCES

Borcard,D. *et al.* (2011) *Numerical Ecology with R*. Springer, New York, NY.

Caporaso,J.G. *et al.* (2010) Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

HMP Consortium et al. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Legendre,P. and Legendre,L. (2012) *Numerical Ecology*. Elsevier, Amsterdam, The Netherlands.

McMurdie,P.J. and Holmes,S. (2013) phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, **8**, e61217.

R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ravel,J. *et al.* (2011) Vaginal microbiome of reproductive-age women. *Proc. Natl Acad. Sci. USA*, **108**(Suppl. 1), 4680–4687.

RStudio Inc. (2013) *shiny: Web Application Framework for R*. R package version 0.8.0.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Srinivasan,S. and Fredricks,D.N. (2009) The human vaginal bacterial biota and bacterial vaginosis. *Interdiscip. Persp. Infectious Dis.*, **2008**, doi:10.1155/2008/750479.

Vázquez-Baeza,Y. *et al.* (2013) Emperor: a tool for visualizing high-throughput microbial community data. *Structure*, **585**, 20.