Cao et al. BMC Bioinformatics (2016) 17:294
DOI 10.1186/s12859-016-1088-4

BMC Bioinformatics

SOFTWARE Open Access

RevEcoR: an R package for the reverse ecology analysis of microbiomes

Yang Cao[1,2], Yuanyuan Wang[3], Xiaofei Zheng[1], Fei Li[1*] and Xiaochen Bo[1*]

Abstract

Background: All species live in complex ecosystems. The structure and complexity of a microbial community reflects not only diversity and function, but also the environment in which it occurs. However, traditional ecological methods can only be applied on a small scale and for relatively well-understood biological systems. Recently, a graph-theory-based algorithm called the reverse ecology approach has been developed that can analyze the metabolic networks of all the species in a microbial community, and predict the metabolic interface between species and their environment.

Results: Here, we present RevEcoR, an R package and a Shiny Web application that implements the reverse ecology algorithm for determining microbe–microbe interactions in microbial communities. This software allows users to obtain large-scale ecological insights into species' ecology directly from high-throughput metagenomic data. The software has great potential for facilitating the study of microbiomes.

Conclusions: RevEcoR is open source software for the study of microbial community ecology. The RevEcoR R package is freely available under the GNU General Public License v. 2.0 at http://cran.r-project.org/web/packages/RevEcoR/ with the vignette and typical usage examples, and the interactive Shiny web application is available at http://yiluheihei.shinyapps.io/shiny-RevEcoR, or can be installed locally with the source code accessed from https://github.com/yiluheihei/shiny-RevEcoR.

Keywords: Metabolic network, Microbiome, Reverse ecology

Background

All of the species living within a particular environment comprise a complex biological community. Various and complicated interactions exist between all the species in the community. Ecology is the scientific analysis and study of the interactions between living organisms, including humans, and their physical surroundings. It is the branch of biology concerned with the relations between organisms and the environment. Traditionally, ecological research strategies have been limited to organisms with well-characterized habitats. This strategy, however, can only be applied on a small scale and to relatively well-understood biological systems. Unfortunately, very few (<1 %) microbial organisms can be cultured in a laboratory, and, therefore, most have not been adequately characterized, particularly from an ecological perspective [1, 2].

With the development of next-generation sequencing technologies and community genomics, the full genomic information for species whose ecology and habitat are largely uncharted can now be easily collected. Specifically, next-generation sequencing-based metagenomic sequencing now allows researchers to investigate the genomes of all species present in a given complex environment. This technology enables microbiologists to study unculturable microorganisms that have previously not been well studied. Metagenomics advances the study of microbial communities on the most fundamental genomic level, and has emerged as a powerful tool for research of the microorganisms living in microbial communities. The structure of complex biological systems in microbial communities reflects not only their diversity and function, but also the environments in which they live [3]. However, most recent metagenomics research has primarily focused on species diversity analysis and

* Correspondence: pittacus@gmail.com; boxc@bmi.ac.cn
[1]Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 100850, China
Full list of author information is available at the end of the article

BioMed Central

© 2016 Cao et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

the functional study and identification of marker genes correlated with host state detection, and neglects the interactions between various species and their natural environment. A systematic approach for describing microbiome ecologies and the interactions between microbiota is lacking. To address this challenge, a systems biology approach called reverse ecology has been developed to study the complex interactions and species composition of microbial communities [4]. Reverse ecology uses genomics to study community ecology with no *a priori* assumptions about the organisms under consideration. Researchers can use it to infer the ecology of a system directly from genomic information. The reverse ecology framework uses advances in systems biology and genomic metabolic modeling and the system-level analysis of complex biological networks to predict the ecological traits of poorly studied microorganisms, their interactions with other microorganisms, and the ecology of microbial communities. Several studies have applied this approach to investigate the interactions between microorganisms and their surroundings on a large scale [4, 5].

Several network-based reverse ecology tools, such as NetSeed [6] and NetCooperate [7], have been developed for studying the species interface with its environment and interactions with other species. Unfortunately, neither NetSeed nor NetCooperate support the metabolic network reconstruction of species, and both are limited to small-scale analyses. Here, we describe RevEcoR, a freely available, easy-to-use software for studying the interface between species and their environments on a large scale, and also for predicting the interactions between species in a given environment. RevEcoR implements the reverse ecology framework allowing users to reconstruct the metabolic networks and study the ecology of poorly characterized microbial species from their genomic information. It has substantial potential for microbial community ecological analysis. See the RevEcoR vignette included with the download for full function and application descriptions. An installation guide and various generic use case examples are also described in the vignette provided with the software.

Implementation

This package uses genome-scale metabolic network models to predict the species interactions. To this end, functional annotated genomic data is used to reconstruct the metabolic network for each species. Seed set algorithm [5] is then employed to identity the species exogenously acquired compounds from its surroundings. Based on the species seed set, two interaction indices, competition and complementarity index are calculated for pair of species. RevEcoR employs igraph [8] to store the metabolic networks for making network analysis efficiency, portability and ease-to-use. And data manipulation functions are built upon hadlyverse packages to make our package easy to extend, such as purrr [9], plyr [10] and stringr [11]. In the following we explain the three main features of the package mentioned above. Figure 1 shows the RevEcoR analysis flow chart.

Metabolic network reconstructions

A graph-based algorithm is used to reconstruct the genome-scale metabolic network. As shown in Fig. 1, there are several ways to obtain the metabolic data for metabolic network reconstruction. Both the Kyoto Encyclopedia of Genes and Genomes (KEGG) [12] and the Integrated Microbial Genomes database (IMG) [13] collect complete high-quality genome sequences and metagenome

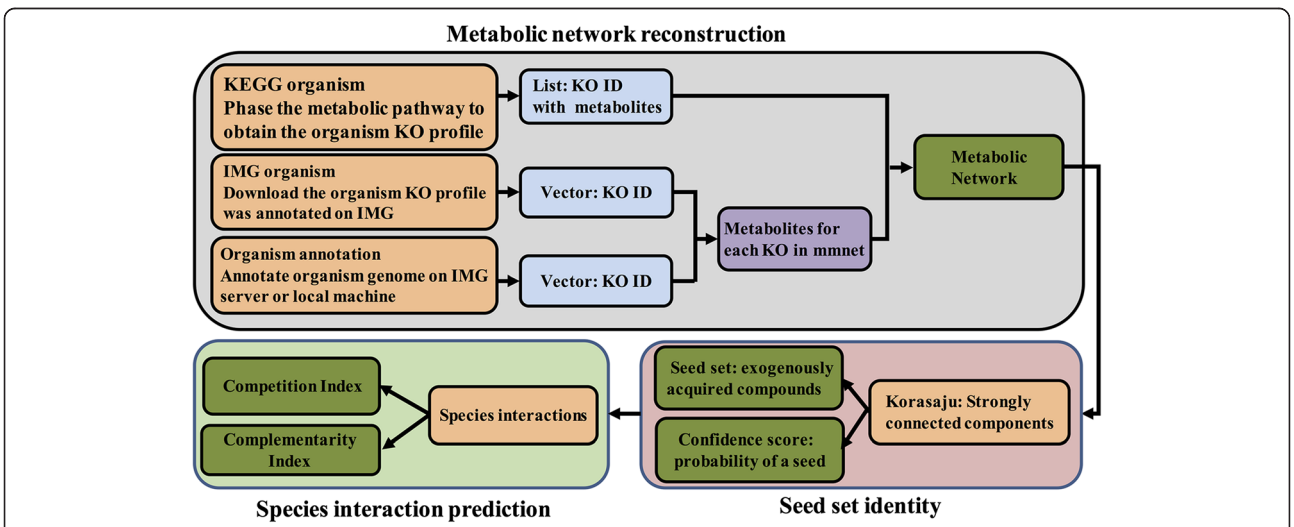

Fig. 1 RevEcoR analysis flow diagram

sequences offering a comprehensive set of publicly available bacterial, archaeal, eukaryotic, and phage genomes, as well as engineered, environmental, and host-associated metagenome samples. All the sequences and functional annotation profiles can be downloaded directly. In RevEcoR, function *getOrgMetablicData* utilizes KEGG REST API to obtain the metabolic data of specific species from KEGG database. In addition, users can annotate their private genomic data with KEGG Orthology terms on IMG systems or on their local machine to obtain annotation profiles.

After the microbial metabolic data are obtained, users can then reconstruct the metabolic network, in which nodes represent compounds involved in metabolism, and direct edges from node A to node B indicate that compound A is a substrate in some reaction that produces compound B.

Predicting exogenously acquired compounds

Exogenously acquired compounds are specified as the seed set [5]. The seed set is defined as the minimal subset of compounds involved in an organism's metabolism that cannot be synthesized from other compounds in the network, and all other compounds in the network can be produced by the presence of the seed set. The seed set of a given network is predicted based on the metabolic network topology. It represents the externally acquired compounds of the species, serves as the interface between species and their environments, and can be used to calculate interactions between various species. RevEcoR uses the Kosaraju algorithm [14] to decompose the metabolic network into strongly connected components. A strongly connected component (SCC) is a maximal set of vertex with no additional edges or vertices from the network can be included in the subgraph without breaking its property of being strongly connected. The seed set of the given network comprised SCCs without incoming edges and at least one outgoing edge. All SCCs form a collection of candidate exogenously acquired compounds. It should be noted that the seed set compounds must include exactly one compound from each SCC, since if one compound in a SCC can be produced then all others will be produced as well. All compounds in a given SCC has the same probability be identity as a seed compound. Here, we assigned a confidence score: 1/(SCC size) to each seed compound, to measure the probability of each seed compound. A versatile data structure, the R S4 class, was taken to manage the seed set.

Calculating species interactions

Species in complex biological systems are able to communicate with each other. Microbial ecology research has revealed that there are two primary routes of species interaction: competition and complementarity. Species interactions are reflected in the topology and calculated from the seed set of species metabolic networks, based on the reverse ecology method. Thus, two topology-based metrics, a competition index and a complementarity index, quantify the interspecific interactions between pairs of species [4]. The competition index is defined as the fraction of compounds in species A's seed set that are also included in species B's seed set, and provides a measure of competition between species A and B. Similarly, the complementarity index is defined as the fraction of compounds in species A's seed set appearing in the metabolic network, but not appearing in species B's seed set, and is used to measure the support capability from species B to A [15]. Notably, the two interaction indices are calculated as a normalized weighted sum for each seed compound is associated with a confidence score.

Results and discussion

RevEcoR software can be used for pairwise species interaction predictions from whole-genomic data. In this section, we briefly demonstrate the results from applying RevEcoR to two datasets, a simple dataset containing seven species, and a large-scale dataset consisting of a human gut microbiome.

Predicting species interactions

We applied RevEcoR to predict cooperation among several human oral microbiota species whose co-occurrence patterns have previously been described [16]. Human oral microorganisms have been extensively cultured and characterized, and oral species not amenable to growth in culture have been described using culture-independent molecular methods such as metagenomic sequencing [17]. Our seven sample oral species metabolic dataset was downloaded from the IMG server; it consists of the following Genomes Online Database (GOLD, http://www.genomesonline.org) [18] IMG identification numbers: Gc0016386, Gi07614, Gi00264, Gc00809, Gc00643, Gi03876, and Gi07289.

Function *getSeedSets* is used for seed set prediction. With a valid dataset, the seed sets (Additional file 1: Table S1) were identified using the following code:

```
## load the KO annotation dataset of seven oral species

library(RevEcoR)

data(anno.species)

##load the reference metabolic data

data(RefDbcache,package="mmnet")
## metabolic network reconstruction of these seven species
net  <- lapply(anno.species, reconstructGsMN)
## identify exogenously acquired compounds (seed set)
seed.set<- lapply(net, getSeedSets)
```

Interactions among various species is calculated using the function *caculateCooperationIndex*. Species interactions can affect the evolution of species and the development of an environment, as well as the species composition in an ecosystem. Both the competition and complementarity indices can be calculated using this function (Additional file 1: Table S2A–D).

```
## caculate interactions among vious species
interactions  <- calculateCooperationIndex(net)
## extract competetion and complementarity index
interactions  <- interactions[c("competition.index","complementarity.index")]
```

We found that *Streptococcus oralis* and *Streptococcus gordonii* had the lowest complementarity index (0.04 and 0.00, respectively) and the highest competition index (0.91 and 0.93, respectively) among all pairs. This indicates that these two species are antagonistic, which is consistent with previous findings [19, 20].

Comparing predicted interactions and co-occurrences

Subsequently, RevEcoR was used to investigate species interactions in a large-scale human microbiome dataset containing 124 samples. We focused on a list of 116 prevalent gut species, whose genome sequences are available in the IMG database and that possess sequence coverage of more than 1 % in at least one metagenomic sample of 124 individuals. Genome annotation profiles for these 116 species were collected from the IMG database and used to calculate the interactions (competition and complementarity indices) for all pairs of species (see Additional file 2 Dataset for details). Abundance of these species across 124 samples was obtained from metagenomic analysis [21]. Co-occurrence scores, which are calculated based on species abundances across all samples and measured by the widely-used Jaccard similarity index [22], were collected from Carr and Borenstein [6] (see Additional file 2 Dataset for details). A comparison of species interactions and co-occurrences allowed us to predict whether species competing with one another

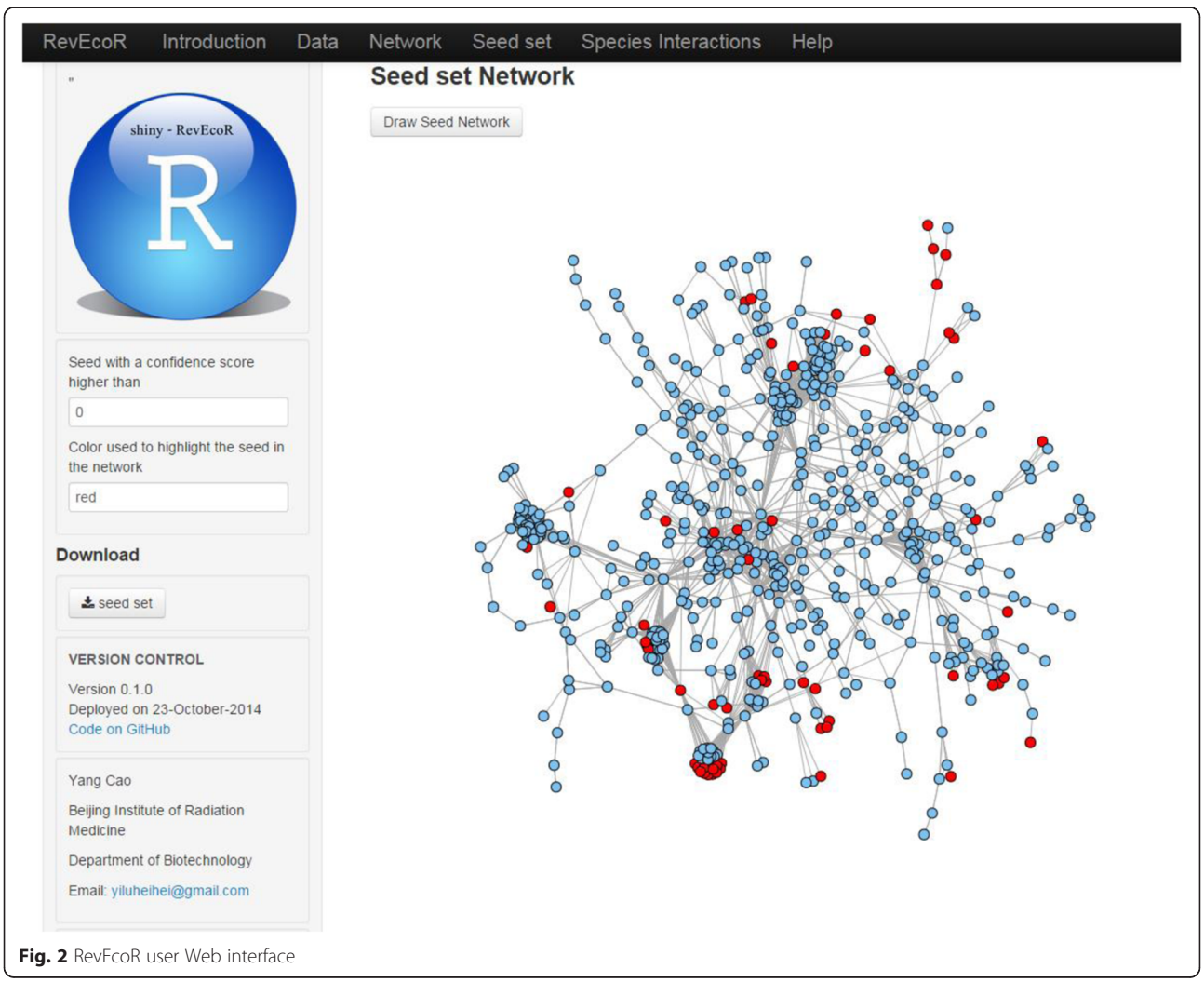

Fig. 2 RevEcoR user Web interface

tended to co-occur or be excluded by the competitor. We found that the competition index is significantly and positively correlated with co-occurrence (cor = 0.261, $P < 10^{-4}$, Mantel correlation test [23], a test commonly used in ecology where the data are usually estimates of the "distance" between objects such as species of organisms), whereas the complementarity index is significantly and negatively correlated with co-occurrence (cor =–0.259, $P < 10^{-4}$, Mantel correlation test).

Moreover, we developed a Shiny-based [24] application based on top of the RevEcoR R package for the rapid, reproducible, and interactive exploration of reverse ecology analysis. Shiny is an interactive web application framework for R, which is cross-platform compatible, and can be launched locally in any R environment or hosted by a remote web server. This means users without a computational background can quickly and easily perform reverse ecology analysis with Shiny RevEcoR. More details and an in-depth manual are available on the RevEcoR app webpage (http://yiluheihei.shinyapps.io/shiny-RevEcoR). Figure 2 shows the RevEcoR user Web interface.

Conclusions

RevEcoR delivers a graph-theory based approach to facilitate the analysis of microbial communities on a large scale. RevEcoR can deepen our understanding of microbial community ecology, enabling the prediction of a variety of interactions in complex systems. Moreover, the reverse ecology framework can also be applied to predict complex community-level differences between microbial communities (such as the human microbiome) and their environments. Therefore, RevEcoR will prove useful for the analysis of microorganisms, and specifically the human microbiome.

Note the current version of RevEcoR has two limitations that we will address in future. One limitation is the computation speed, seed set identity and the p value calculation of species interaction is the main cause of slow computation. We plan to take C++ API of R in a new version of our package that alleviates the computation efficiency. The second limitation is the initial input of the software must be the annotated metabolic data, which requires extra data processing procedures to prepare the data. We are currently developing modules that will achieving this process in R.

Availability and requirements

Project name: RevEcoR

Project home page: https://cran.r-project.org/web/packages/RevEcoR/index.html

Operating system(s): Platform independent

Programming language: R

Other requirements: R 2.14 or higher

License: GNU GPL v. 2.

Any restrictions to use by non-academics: none

Additional files

Additional file 1: Table S1: The seed sets of seven oral species. *seed* represents the exogenous required compound from its environment, and *confidence* represents the compounds probability of being a seed. **Table S2A-D**. The competition and complementarity index of seven oral species. (DOCX 23 kb)

Additional file 2: Additional information on 116 gut prevalent species, including species names, species interactions and co-occurrence scores. (XLSX 300 kb)

Abbreviations

GOLD, genomes online database; IMG, integrated microbial genomes database; KEGG, Kyoto encyclopedia of genes and genomes; SCC, strongly connected component.

Acknowledgements

We thank all members of the Department of Biotechnology, who helped us improve our software.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 81273488), the National Key Technologies R & D Program for New Drugs (Grant No.2012ZX09301003-002), and the Program of International S&T Cooperation (Grant No. 2014DFB30020).

Availability of data and materials

All materials are included as additional files.

Authors' contributions

YC implemented the package and wrote the user manual. YW tested the software and prepared vignettes of the package. FL designed the structure and interface of the software, and drafted the manuscript. XZ and XB participated in the design of the package, and helped to draft the manuscript. All authors read and approved the final manuscript.

Authors' information

YC is a PhD candidate in the Department of Biotechnology at the Beijing Institute of Radiation Medicine, and also an intern researcher in Tianjin Key Laboratory of Risk Assessment and Control Technology for Environment and Food Safety at Tianjin Institute of Health and Environmental Medicine (contact: yiluheihei@gmail.com). YW is a master candidate in the Department of Basic Courses at the Army Officer Academy (contact: yywang1217@gmail.com). XZ is a professor in the Department of Biochemistry and Molecular Biology at the Beijing Institute of Radiation Medicine (contact: xfzheng100@126.com). FL is an assistant professor in the Department of Biotechnology at the Beijing Institute of Radiation Medicine (contact: pittacus@gmail.com). XB is a professor in the Department of Biotechnology at the Beijing Institute of Radiation Medicine (contact: boxc@bmi.ac.cn).

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 100850, China. [2]Tianjin Key Laboratory of Risk Assessment and Control Technology for Environment and Food Safety, Tianjin Institute of Health and Environmental Medicine, Tianjin 300050, China. [3]Department of Basic Courses, Army Officer Academy, Hefei 230031, China.

Received: 9 September 2015 Accepted: 21 May 2016
Published online: 29 July 2016

References

1. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol. 1998; 180:4765–74.
2. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59:143–69.
3. Levy R, Borenstein E. Reverse Ecology: from systems to environments and back. Adv Exp Med Biol. 2012;751:329–345.
4. Levy R, Borenstein E. Metagenomic systems biology and metabolic modeling of the human microbiome. Gut Microbes. 2014;5:265–270.
5. Borenstein E, Kupiec M, Feldman MW, et al. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci. 2008; 105:14482–7.
6. Carr R, Borenstein E. NetSeed: a network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment. Bioinforma Oxf Engl. 2012;28:734–5.
7. Levy R, Carr R, Kreimer A, et al. NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. BMC Bioinf. 2015; 16:164.
8. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006;1695:1–9.
9. Hadley Wickham. purrr: Functional Programming Tools. R package version 0.2.1. 2016. https://CRAN.R-project.org/package=purrr. Accessed 12 Feb 2016.
10. Wickham H et al. The split-apply-combine strategy for data analysis. J Stat Softw. 2011;40:1–29.
11. Hadley Wickham. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.0.0.9000. https://github.com/hadley/stringr. Accessed 29 Apr 2015.
12. Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–462.
13. Markowitz VM, Chen I-MA, Chu K, et al. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res. 2014;42: D568–573.
14. Tarjan R. Depth first search and linear graph algorithms. SIAM J Comput. 1972;1:146–160.
15. Levy R, Borenstein E. Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc Natl Acad Sci U S A. 2013;110:12804–9.
16. Kolenbrander PE. Multispecies communities: interspecies interactions influence growth on saliva as sole nutritional source. Int J Oral Sci. 2011;3: 49–54.
17. Dewhirst FE, Chen T, Izard J, et al. The human oral microbiome. J Bacteriol. 2010;192:5002–17.
18. Reddy TBK, Thomas AD, Stamatis D, et al. The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 2015;43:D1099–D1106.
19. Palmer RJ, Kazmerzak K, Hansen MC, et al. Mutualism versus independence: strategies of mixed-species oral biofilms in vitro using saliva as the sole nutrient source. Infect Immun. 2001;69:5794–804.
20. Periasamy S, Kolenbrander PE. Mutualistic biofilm communities develop with Porphyromonas gingivalis and initial, early, and late colonizers of enamel. J Bacteriol. 2009;191:6804–11.
21. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
22. Chao A, Chazdon RL, Colwell RK, et al. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett. 2005;8:148–59.
23. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967;27:209–20.
24. Rstudio. Shiny, A web application framework for R. [http://shiny.rstudio.com/]. Accessed 5 Aug 2015

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit