

Chapter 1

Sampling of Riverine or Marine Bacterial Communities in Remote Locations: From Field to Publication

Katja Lehmann

Abstract

This protocol describes how to sample and preserve microbial water column samples from rivers that can be used for 16S or 18S metabarcoding studies or shotgun sequencing. It further describes how to extract the DNA for sequencing and how to prepare raw Illumina MiSeq amplicon data and analyze it in the R environment.

Key words Biodiversity, Riverine microbial communities, Community analysis, Illumina MiSeq

1 Introduction

High-throughput sequencing technologies have revolutionized biodiversity studies of prokaryotes and are increasingly used to study whole communities. In rivers, which are diverse environments with different sub-habitats that are closely linked to adjacent terrestrial biomes, the collection and analysis of eDNA offer unprecedented monitoring opportunities. Collection of viable riverine microbiological samples, however, is often confounded by the remoteness and inaccessibility of sampling sites. In addition to this, increasingly open access to research data creates a need for data interoperability and so a need for standardized procedures to ensure consistency between datasets. Mega-sequencing campaigns, such as the Earth Microbiome Project or Ocean Sampling Day (OSD) [1, 2], can help this goal by driving the development of cost-effective protocols and analysis methods.

Here, I describe the standardized procedures to collect and analyze DNA samples for River Sampling Day (RSD) (part of Ocean Sampling Day [2]), a simultaneous sequencing campaign collecting a time series of ocean and river microbial data from the June and December solstices each year in rivers worldwide on the same day. Low-cost, manual sampling tools are utilized to collect

riverine water column samples, which can subsequently be extracted for 16S, 18S, metabarcoding, or shotgun sequencing. I then proceed to describe the preparation of raw Illumina MiSeq amplicon data for analysis, followed by a description of an analysis pipeline for the open-source statistics software package R. The sampling protocol is based on methods used in the Freshwater Biological Sampling Manual [3] and the methods used at the Western Channel Observatory in the UK [4].

2 Materials

2.1 Sample Collection

1. Sterivex 0.22 μm filter cartridge SVGPL10RC (male luer-lock outlet, Millipore, UK) or SVGP010 (male nipple outlet, Millipore, UK).
2. RNeasy (Thermo Fisher Scientific, Loughborough, UK, or Ambion Inc., Austin, Texas).
3. Luer-lock syringes, 3 mL (e.g., Medisave, Weymouth, UK).
4. High-pressure sampling bottle fitted with bicycle valve and tube outlet, 10% acid washed (e.g., Nalgene™ Heavy-Duty PPCO Vacuum Bottle, Thermo Fisher Scientific, Loughborough, UK). See Fig. 1 for illustration.
5. PTFE tubing.
6. Two hose clamps to secure PTFE tubes.
7. Bicycle pump.
8. Nitrile gloves.
9. 70% ethanol to clean gloves/equipment.
10. Sterilized sticky tac (e.g., Blu-Tac).

2.2 Sample Preparation

1. DNA extraction solution, e.g., a commercial kit
2. Sterile consumables and shakers/bead beaters/centrifuges required for DNA extraction method of choice.

2.3 Raw Data Preparation and Analysis

1. QIIME software (qiime.org).
2. R environment (www.r-project.org).

3 Methods

3.1 Sample Collection

Collect three replicate bacterial samples at each location. If sampling multiple locations, take the samples at each location within the same time frame. For specifics on safety, clean sampling, and record taking, see **Notes 1–3**.

3.1.1 Option A: Sampling Midstream

1. With a 10% acid-washed sampling bottle, wade into the river downstream from the point at which you will collect the sample.

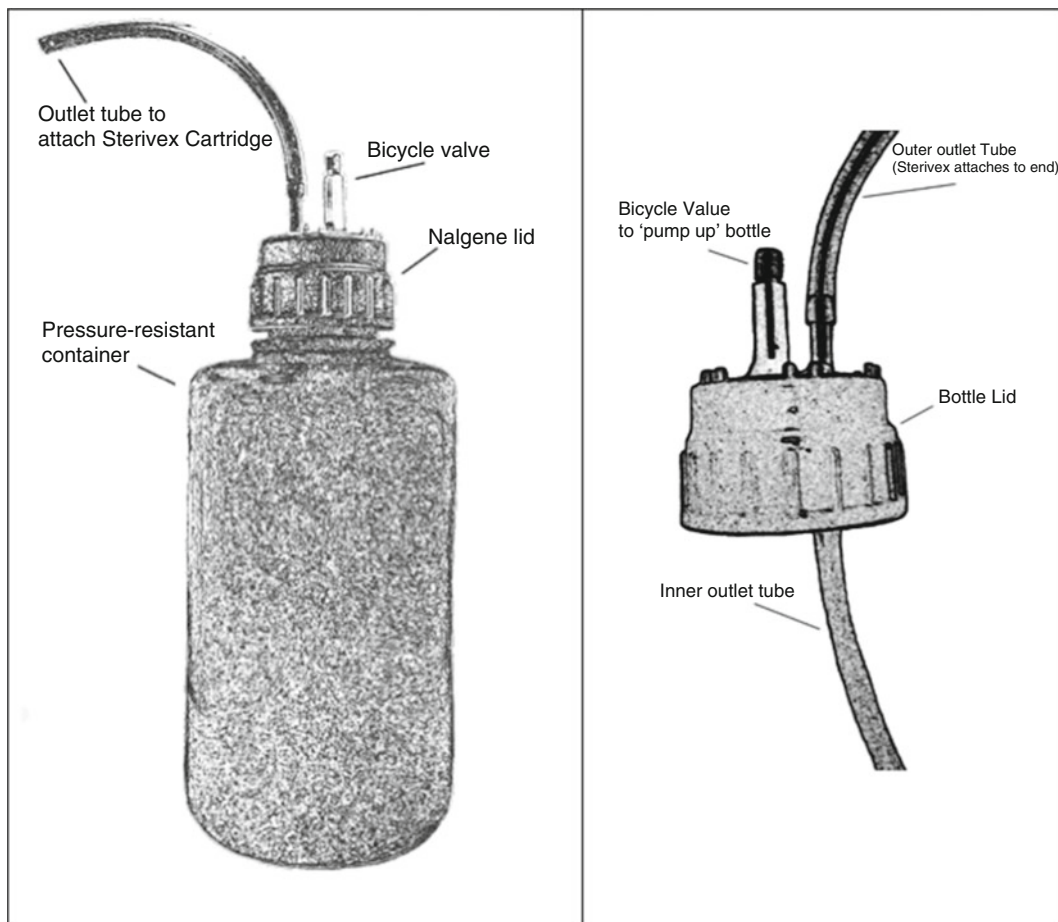


Fig. 1 Field sampling bottle with altered screw top to allow pressurized filtering

2. Wade upstream to the sample site. This ensures that you will not disturb sediments upstream from the sample point. Stand perpendicular to the flow and face upstream.
3. Remove the lid and hold it aside without allowing anything to touch the inner surface. With your other hand, grasp the bottle well below the neck.
4. Plunge it beneath the surface with the opening facing directly down; then immediately orient the bottle into the current.
5. Once the bottle is full, remove it from the water by forcing it forward (into the current) and upward.

3.1.2 Option B: Sampling from the Stream Bank

1. Secure yourself to a solid object on the shore.
2. Remove lid from a 10% acid-washed sampling bottle.

3. Hold the bottle well below the neck. Reach out (arm length only) and plunge the bottle under the water, and immediately orient it into the current.
4. When the bottle is full, pull it up through the water while forcing it into the current.

3.1.3 Filtering

Now collect the microbial community by passing 1–5 L of the sampled water through a 0.22 μm filter using Sterivex cartridges. Filtration through the Sterivex filter should be done using the field sampling bottle:

1. Attach the Sterivex filter to the tube that comes out of the cap of your bottle (*see* Fig. 1). Secure the tube to the cap outlet and the Sterivex filter with two hose clamps.
2. Attach the bicycle pump to the valve on the bottle cap and pump up the bottle. If the water is very clear, you might need to refill the bottle and filter up to 5 L through the same filter. If you have to collect more water, place the bottle cap, and filter in sterile bags while you refill the bottle as described in Subheading 3.1. The filtration is done when the filter begins to clog up.
3. When the filtering is done, the Sterivex should be pumped free of standing water.
4. Seal the nipple side of the filter using sterilized sticky tac or similar; then use a 3 mL luer-lock syringe to fill the filter with RNAlater.
5. Seal the Sterivex filter using sterilized sticky tac or similar. Note that parafilm will crumble at temperatures below $-45\text{ }^{\circ}\text{C}$ and therefore should not be used.

3.1.4 Transport and Storage

1. Label the sample and place it in a sterile bag or tube. For transport from the sampling location to the lab, samples can be stored in the sealed bag in a cooling container. Samples in RNAlater can be kept at room temperature up to a week.
2. On arrival at the lab, provided the samples are stored in RNAlater, freeze samples at $-20\text{ }^{\circ}\text{C}$ (not $-80\text{ }^{\circ}\text{C}$, or store in the refrigerator at $4\text{ }^{\circ}\text{C}$ for up to a month if no freezer is available.

3.1.5 Metadata

A subset or all of the following metadata will greatly help to make sense of the microbial data, especially when multiple locations are being sampled. At minimum metadata should consist of sample volume, depth, temperature, lat/lon, time of day, and pH. Depending on the study, these should be supplemented by alkalinity, suspended sediment, soluble reactive phosphorus (SRP), total dissolved phosphorus (TDP), total phosphorus (TP), Si, F, Cl, Br, SO_4 , total

dissolved nitrogen (TDN), NH_4 , NO_2 , NO_3 , dissolved organic matter (DOC), Na, K, Ca, Mg, B, Fe, Mn, Zn, Cu, and Al. The metadata measurements should be taken at or close to the date/time when the samples are collected.

3.2 Sample Preparation in the Lab

The DNA captured on the Sterivex filters needs to be extracted. For OSD/RSD, a commercial kit ensures consistency during extraction, but commercial kits are often not sold sterile, and some have been found to contaminate samples [5, 6]. This can confound sequencing results, especially when DNA concentration in the samples is expected to be low. Contamination of samples from a number of sources is a well-known danger at many stages of the extraction and sequencing process; it is therefore advisable to run a control alongside the samples during all steps, including the sequencing. Problems can also be caused by unexpected cross effects of preservative residues and kit reagents. It is advisable to do a test extraction first and verify DNA yield and quality on a gel. If necessary, an additional PEG precipitation prior to final cleaning steps on a kit spin filters might insure against DNA loss. Once the DNA is extracted, it can be amplified by PCR for 16S or any other amplicon analysis or transferred to the sequencing facility for shotgun sequencing as is. See Chapters 12 (Fonseca and Lallias), 13 (Bourlat et al.), and 14 (Leray et al.) for protocols detailing the preparation of amplicon libraries for Illumina sequencing. Here, we will focus on the analysis of 16S amplicon data derived from Illumina sequencing.

3.3 From Raw Sequencing Data to Operational Taxonomic Unit (OTU) Table

One of the most popular pipelines to process data derived from next-generation sequencing is QIIME [7], an open-source software wrapper that incorporates a great number of python scripts, including complete programs such as MOTHUR, in itself a comprehensive analysis pipeline [8], and USEARCH, a BLAST alternative [9]. For installation options, see **Note 4**. QIIME can process Illumina data, but also 454 and (with a bit of preprocessing) Ion Torrent data. We will focus on Illumina here, which has emerged as the predominant sequencing method in the last few years. To process any raw sequence data, QIIME requires a mapping file with metadata such as sample ID and primer- and barcode information (see Chapter 15 by Leray and Knowlton for further details on QIIME mapping file format). Depending on the format of the raw sequences, four scripts are required to process a set of Illumina-derived sequences in QIIME to obtain data that can be used for statistical analysis:

1. *join_paired_ends.py*, a script that joins paired end reads.
2. *validate_mapping_file.py*, a script which checks the soundness of the mapping file.

3. *split_libraries_fastq.py*, a script that divides the raw sequence library by barcode.
4. *pick_de_novo_otus.py*, a workflow that produces an OTU mapping file, a representative set of sequences, a sequence alignment file, a taxonomy assignment file, a filtered sequence alignment, a phylogenetic tree, and a biom-formatted OTU table.

Both the phylogenetic tree and the OTU table can then be exported into other programs for further analysis. In QIIME itself, further scripts allow for exploration of alpha diversity and beta diversity, notably:

5. *summarize_taxa_through_plots.py*, which creates taxonomy summary plots.
6. *alpha_rarefaction.py*, which calculates rarefaction curves.
7. *beta_diversity_through_plots.py*, which performs principle coordinates (PCoA) analysis on the samples.

For detailed instructions on performing diversity analyses with QIIME, see chapter 15 by Leray and Knowlton. QIIME also offers network analysis, which can be visualized in Cytoscape. QIIME has comprehensive help pages for each script (<http://qiime.org/scripts/>) and a number of tutorials (<http://qiime.org/tutorials/index.html>).

3.4 Basic Statistics in R

The open-source software package R [10] is a well-known statistics and scripting environment. It is available for Linux, Mac OS X, and Windows (www.r-project.org). Before starting the analysis, the following libraries need to be installed in R: biom [11], RColorBrewer [12], vegan [13], gplots [14], calibrate [15], ape [16], picante [17], Hmisc [18], BiodiversityR [19], psych [20], ggplot2 [21], grid [10], and biocLite.R [22].

3.4.1 Data Preparation

As a first step, the OTU table has to be imported into R either as text or as biom-formatted file with the following commands:

- Untransposed:

```
otu.table=read.table("your_otu_table.txt ",
  sep="\t",header=T,row.names=1)
```

transposed:

```
otu.table=t(read.table("your_otu_table.txt", sep="\t",header=T,row.names=1))
```

- Or in biom format:

```
1. otu_biom="your_otu_table.biom"
```

```
2.. Which needs to be transformed into a matrix:
```

```
otu_biom=as(biom_data(otu_biom), "matrix")
```

Secondly, read in a mapping file:

```
otu.map=read.csv("your_map.csv",header=T)
```

Make sure to match the row order in your data matrix to that in your mapping file:

```
otu.map=otu.map[match(rownames(otu.table),otu.map$OtuID),]
```

Save your original import as backup, in case you mess up your newly created dataframes at some point during the process:

```
otu.raw=otu.table
otu.map.raw=otu.map
```

Do a random rarefaction with the `rrarefy` function from the `vegan` package to reduce the number of sequences in each sample to that of the sample with the lowest number of sequences in the set (check your file or QIIME demultiplex log):

```
otu.table.rar<-rrarefy(otu.table, sample=x)
```

Now create factors—e.g., make the experimental treatment a factor (look at your mapping file to assess which factors you should create to make your analysis interesting or viable):

```
otu.map$Treatment=as.factor(otu.map$Treatment)
```

The data can now be analyzed statistically.

3.4.2 Exploring Alpha Diversity

In datasets, where abundance is represented in an unbiased way, it is easy to determine the most abundant OTU:

```
mostAbundantOtu<-which(colSums(otu.table)==max(colSums(otu.table)))
```

It is then possible to calculate a rank-abundance curve for the data:

```
RankAbun.otus<-rankabundance(otu.table)
```

which is followed by plotting the results proportionally on a log scale. Set the plot panel (e.g., one row, two columns):

```
par(mfrow=c(1,2))
```

Assign x and y axes:

```
x<-RankAbun.otus[,1]
y<-RankAbun.otus[,2]/colSums(RankAbun.otus)[2]
```

Create labels manually:

```
l<-c("label 1", "label 2", "label 3")
```

Plot the data with labels:

```
plot(x, y, log="y", type="o", pch=16, xlab="Species rank", ylab="Proportion", main="Rank Abundance, OTUs", axes=FALSE)
textxy(x, y, labs=l, cex=0.8)
```

Now, traditional diversity indices can be calculated via the diversity function of the vegan package. Shannon-Wiener is set as default:

```
H<- diversity(otu.table)
```

Simpson and others can be calculated by specifying them especially:

```
S<- diversity(otu.table, index="simpson")
```

whereas related indices such as Pielou's J can be calculated with a simple function:

```
J<- H/log(specnumber(otu.table))
```

3.4.3 Exploring Beta Diversity

When working with a number of sites or treatments, the diversity indices can be presented next to each other in a boxplot for easy comparison.

First, create new objects for each treatment group or site:

```
H_resultsGroup1<-c(Result1, Result2, Result3, etc)
```

```
H_resultsGroup2<-c(Result1, Result2, Result3, etc)
```

The next step is to create a boxplot (with outliers represented as dots):

```
lmts<- range(H_resultsGroup1, H_resultsGroup2)
boxplot(H_resultsGroup1,H_resultsGroup2, ylim=lmts,
names=c("Group_1","Group_2"), xlab="myExperiment",
main="Shannon diversity with SD")
```

It is also possible to test for significant differences between the calculated diversity indices with an ANOVA. Read in a matrix that lists samples/replicates in rows and diversity index results and treatment assignments per replicates in columns. To start the analysis, the linear model needs to be created first:

```
div.an<- lm(Shannon~factor(Treatment)+factor
(myFactor1)*factor(myFactor2), data=div.matrix)
```

The results can be printed in the console:

```
summary(div.an)
```

The ANOVA is then produced as follows:

```
anova(div.an)
```

A convenient way to explore differences between samples visually is by nonmetric multidimensional scaling (NMDS [23]). In an initial step, R has to produce a dissimilarity matrix (e.g., Bray-Curtis) as the basis:

```
otu.table.nmnds=metaMDS(otu.table,distance="bray",tr
ymax=49)
sites=scores(otu.table.nmnds, display="sites")
```



```
taxa=scores(otu.table.nmds, display="species")
```

If there are missing data points in the OTU table, it is good to remove them:

```
x1=range(sites[,1], taxa[,1],na.rm=T)y1=range(sites[,2],
taxa[,2],na.rm=T)
```

The NMDS results can be plotted in the following way:

```
plot(otu.table.nmds)
```

This has created a plot without labels and with crosses for species.

To create a plot with adjusted x and y axis ranges in which the dots are labeled by treatment, R can be instructed as follows:

```
plot(otu.table.nmds, type="n", xlim=x1, ylim=y1,
main="NMDS of bacterial samples, plotted by treatment")
points(sites,col=c("red","blue")[as.numeric(otumap$Treatment)], pch=16, cex=1.0)
```

To create site labels, the following command can be used:

```
text(otu.table.nmds, display="sites", pos=4, cex=0.7,
offset=0.3)
```

R can also create species labels—if there are many species, this can make a plot far too busy:

```
text(otu.table.nmds, display="species", pos=4,
cex=0.7, offset=0.3)
```

The *vegan* package includes several multivariate statistical tests to test for differences between your treatments or locations.

Adonis (aka PERMANOVA [24]) is a multivariate equivalent to ANOVA. Any data needs to be checked for multivariate normality to make sure that the PERMANOVA is applicable:

```
adonis.otu=adonis(otu.table~otu.map$Treatment,method="
bray");adonis.otu
```

If a PERMANOVA is not permissible, there is another multivariate equivalent to ANOVA, called ANOSIM [25]. ANOSIM is a randomization-based method to analyze differences by comparing dissimilarity matrices of ranked data. It is less robust than PERMANOVA/Adonis, as it does not compare the distances directly. ANOSIM produces an R-statistic, which shows increasing differences in community composition on a scale of 0–1. The accompanying P-value shows if the R-value is significant or not. This is how the ANOSIM is called in R:

```
anosim.otu=anosim(vegdist(otu.table),grouping=otu.
map$Treatment,permutations=999)
```

This will produce a summary:

```
summary(anosim.otu)
```

And this will produce a boxplot of the result:

```
plot(anosim.otu)
```

Lastly, a *simper* analysis [25] can yield information about which OTUs contributed most to the observed dissimilarity between the treatments/locations. It is called as follows:

```
sim<- with(otu.map, simper(otu.table, Treatment))
```

The output can be assessed with the *summary* command:

```
summary(sim, ordered=TRUE, digits=3)
```

3.4.4 Further Analysis Options

This is only a small selection of analysis methods that can be performed in R and exploration is encouraged. Additionally, there are software packages such as MEGAN [26] or PICRUSt [27], which can offer additional workflows for data in biom format to explore metagenome data further.

4 Notes

1. If you collect from more than one location without being able to acid wash the equipment in between, please pump sterilized, deionized water through the equipment (bar filter) between locations. If that isn't possible, pump river water from the next location through your bottle before you start to filter (not recommended unless there is no better option).
2. Wherever practical, samples should be collected at midstream/offshore rather than nearshore. Samples collected from midstream/offshore reduce the possibilities of contamination (e.g., back eddies or seepage from nearshore soils). The most important issue to consider when deciding where the sample should be collected from is safety. If the flow is sufficiently slow and shallow for the collector to wade, e.g., into a stream without risk, then the sample can be collected at a depth where there is no risk that water might flow into the waders from above.
3. Record:
 - How much water you filtered
 - The time taken to filter the sample
 - Your observations about the color of the filter
 - If you collected from the stream bank or mid-river
4. QIIME can be installed natively on Linux (qiime.org) and Mac OS X (www.wernerlab.org/software/macqiime/) or can be run on Windows via VirtualBox (www.virtualbox.org). QIIME comes pre-installed and pre-configured on Bio-Linux, an open-source curated Linux distribution for bioinformaticians (environmentalomics.org/bio-linux/).

References

- Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome Project: successes and aspirations. *BMC Biol* 12(1):69
- Kopf A et al (2015) The ocean sampling day consortium. *GigaScience* 4(1):1–5
- Clark MJR (2003) British Columbia field sampling manual. Water, Air, and Climate Change Branch, Ministry of Water, Land, and Air Protection, Victoria, BC
- Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K et al (2010) The taxonomic and functional diversity of microbes at a temperate coastal site: a ‘multi-omic’ study of seasonal and diel temporal variation. *PLoS One* 5(11), e15545
- Evans GE, Murdoch DR, Anderson TP, Potter HC, George PM, Chambers ST (2003) Contamination of Qiagen DNA extraction kits with *Legionella* DNA. *J Clin Microbiol* 41(7):3452–3453
- Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 12(1):87
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Gonzalez Pena A, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunencko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537–7541
- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461
- R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, <http://www.R-project.org/>
- McMurdie PJ, The Biom-Format Team (2014) biom: an interface package (beta) for the BIOM file format. <http://biom-format.org/>. R package version 0.3.12. <http://CRAN.R-project.org/package=biom>
- Neuwirth E (2014) RColorBrewer: ColorBrewer palettes R package version 1.1-2. <http://CRAN.R-project.org/package=RColorBrewer>
- Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2015) vegan: community ecology package. R package version 2.3-0. <http://CRAN.R-project.org/package=vegan>
- Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2015) “ggplots: various R programming tools for plotting data. R package version 2.17.0. <http://CRAN.R-project.org/package=ggplots>
- Graffelman J (2013) calibrate: calibration of Scatterplot and biplot axes. R package version 1.7.2. <http://CRAN.R-project.org/package=calibrate>
- Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290
- Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26:1463–1464
- Harrell FE Jr., with Contributions from Charles Dupont and Many Others (2015) Hmisc: Harrell miscellaneous. R package version 3.16-0. <http://CRAN.R-project.org/package=Hmisc>
- Kindt R, Coe R (2005) Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies. World Agroforestry Centre (ICRAF), Nairobi. ISBN 92-9059-179-X
- Revelle W (2015) psych: procedures for personality and psychological research. Version = 1.5.8. Northwestern University, Evanston, IL, <http://CRAN.R-project.org/package=psych>
- Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer, New York, NY
- Huber W, Carey VJ, Gentleman R, Morgan M (2015) Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods* 12:115
- Kruskal J (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29:115–129
- Anderson M (2005) “PERMANOVA: a FORTRAN computer program for permutational multivariate analysis of variance, vol 24. Department of Statistics, University of Auckland, Auckland
- Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. *Aust J Ecol* 18:117–143
- Huson DH, Auch AF, Qi J et al (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386
- Langille MGI et al (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31(9):814–821