Briefings in Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 01/07/2018
Manuscript Category

OXFORD

Genome analysis

Spatial Data Visualization of Microbiome Megasequencing Initiatives

Miguel Faria [1,*], Catarina Magalhães [2] and Rita P. Ribeiro [1,*]

[1]Department of Computer Science, Faculty of Sciences of the University of Porto, Porto, Portugal and
[2]Department of Biology, Faculty of Sciences of the University of Porto, Porto, Portugal

Abstract

Motivation: Marine microbial communities constitute the majority of living biomass of oceans and seas but it still remains poorly described, relatively to their constitution, structure and functionality. The Ocean Sampling Day was a one-day global initiative to gather samples to analyze and characterize marine microbial communities diversity and its genes' functionality, resorting to pipelines based on next-generation sequencing and metagenomic approaches.
Results: The present work aims to take advantage of the metagenomic data that resulted from this event to build an interactive web application for spatial data visualization using the Shiny package from the R programming language. Thus, our effort is to supply a publicly available platform to explore, analyze and acquire deeper insights on the marine microbial communities distribution at a global scale.
Availability: The code for the application and the all the files are publicly available on a GitHub repository [4]. The application is also available online [7]
Contact: up201302745@fc.up.pt

1 Introduction

Microorganisms play an important role in all ecosystems due to their abundance and diverse metabolic pathways [13]. In fact, microbial communities represent 90% of all living biomass [24]. The marine microbiome could be defined as a numerous and complex mixture of microorganisms inhabiting the oceans. As they have high turnover rates this populations have a great genetic diversity [29]. Thus, different microbial communities may influence functionally the oceans and seas as they are the base of the marine food-web and are also involved in relevant biogeochemical processes like the marine carbon, nitrogen and sulphur cycles [13, 24, 27]. Characterizing their diversity as well as their structure and organization becomes a relevant but demanding task [29].

The marine microbiome has billion of genes who are capable of expressing an immense diversity of molecules such as proteins, lipids and other small metabolites with ecological functions, but they are also extremely relevant in the discovery of new drugs, enzymes or biomaterials or in the improvement of those already existing [12]. Only 1% of these microorganisms can be cultured in the laboratory [12], so it is essential to have tools to characterize these microorganisms without the need to grow them in the laboratory environment. With the relatively recent advances in sequence technology and bioinformatics, it is now possible to further understand the dynamics, structure and genomic profile of the marine microbiome. Metagenomic techniques, including next-generation gene sequencing, can be used to surpass the experimental limitations and help investigators by creating new approaches for the study of microbiomes [12].

Metagenomics is described by Bragg *et al.* as "shotgun sequencing of the genomic DNA of a sample taken directly from the environment", representative of the collective genome of the microbial community, providing insights into its structure and function [9]. Next-generation sequencing (NGS) techniques can be used to sequence thousands of sequences from any sample without cultivating it, generating data that provides information on microbial populations and their effects on the environment and health [30, 21]. Marker genes metagenomics, also designated amplicon sequencing, are methods to obtain taxonomic distribution profiles of microbial communities associated with environmental data derived from sampling, using PCR amplification and sequencing of evolutionarily conserved marker genes, such as the 16S rRNA gene [21]. Thus, amplicon sequencing is a widely used method to survey the relative abundance of individual taxa. Amplicon sequencing may fail to characterize diversity in a community due to biases associated with PCR and is limited to the identification and analysis of taxa for which taxonomically informative genetic markers are previously described [25, 28]. Alternatively, whole genome shotgun sequencing, a method in which random fragments of genome are sequenced, allows not only to determine the constitution of a microbiome but also to delve into the functional potential of the microorganisms present [25].

Major next-generation sequencing platform types that have been used for microbiome studies include 454, Illumina, SOLiD and Ion Torrent. Illumina's sequencing instruments, based on reversible dye terminators, are the most widely used on the market due to their superior per-base cost efficiency and high sequencing accuracy [15, 22].

These techniques play an important role to further characterize the marine microbiome, granting crucial advances in marine metagenomic studies. Sampling initiatives such as the Global Ocean Sampling expedition and the Tara Oceans project aim to provide open access data for metagenomic studies and analysis [27, 16]. The Ocean Sampling Day (OSD), via the Micro B3 project ([6]), was a worldwide spatially and chronologically coordinated event to collect ocean samples that aspires to describe the structure, diversity and function of marine microbial communities [19]. The OSD project was a simultaneous global mega-sequencing event that happened on the solstice (June 21st 2014) that resulted in large publicly available data being: i) 155 16S/18S rRNA amplicon data sets, ii) 150 metagenomes and iii) environmental metadata [19]. Oceanographic, environmental and biodiversity data was stored at the PANGAEA, [8], and molecular data was archived at the European Nucleotide Archive (ENA) at the EMBL-EBI, [2, 20]. The OSD sampling sites cover all the major oceans (Pacific, Atlantic, Indian, Antarctic and Arctic Ocean) and continents, being mostly located in coastal regions within exclusive economic zones (EEZ) [19, 20].

The goal of this project is to create a data visualization tool using the open access metagenomic datasets from OSD initiative. By providing researchers a new, interactive and user-friendly spatial platform to visualize microbial communities composition at global geographic scale, this work aims to help them to intuitively explore and analyze the data, identify patterns and causalities more easily. The platform was developed using the package Shiny [10] from the R programming language [23], as it is a way to build interactive web applications while still being able to take advantage of the statistical power of R programming language.

2 Methods

The OSD data was archived and made publicly accessible according to the Fort Lauderdale rules for sharing data [19]. The sequence and contextual data retrieved in this event is only available via the International Nucleotide Sequence Database Collaboration (INSDC) umbrella study PRJEB5129 ([3]) and at PANGAEA. The pipeline used for the taxonomic analysis was based on 16S rRNA amplicon using the QIIME software and is fully described on EBI Metagenomics website ([1]). Note that the pipeline also describes the steps for the functional analysis, a topic not addressed in this work.

Indeed, "env.table.xls" and "Species.xls" were the two datasets used for this application. While "env.table.xls" provided environmental data related to the 150 sampling sites, "Species.xls" incorporated metagenomic data represented by normalized relative prevalence of each taxa in every sampling site. The data in "Species.xls" was the result of the pipeline previously mentioned.

Variables present in the "env.table.xls" dataset include "SampleName", "Latitude", "Longitude", "DepthWater", "Temp" and other environmental variables as can be verified in the "Ocean Sampling Day Handbook" by Micro B3 [20]. The other dataset, "Species.xls", contains variables that refer to the lower taxonomic level OUT table with information about specific taxa that was possible to identify using Greenens database, meaning not all taxa could be described at the genus level.

Minor pre-processing steps were carried at this stage. Namely: i) remove rows in "env.table.xls" that correspond to the same sample site but in which the water depth value (variable "DepthWater") was superior, ii) remove rows in "env.table.xls" that corresponded to sampling sites whose coordinates did not refer to an ocean or sea location, iii) save that file as csv with commas as column separators and dots as decimal separators, changing its name to "NoDepth.table.csv", iv) remove the second row from the "Species.xls" dataset as it is duplicated and v) save that file as csv with commas as column separators and dots as decimal separators. After this process, the "NoDepth.table.csv" dataset ended up with 137 rows each corresponding to a different sampling site identifiable by the first column called "SampleName".

Afterwards, some extra data manipulation and pre-processing techniques were executed. Namely, i) remove the second column in the "NoDepth.table.csv" dataset as it redundant and unnecessary, ii) create a variable called "speciesNames" that is a vector with all the taxa names available in the "Species.csv" dataset, iii) create a new dataset called "allData" that is the result of the merging of the two initial datasets and iv) create a variable named pop that is a vector of all the sample site codes.

3 Shiny App: Ocean Sampling Day

The application shows a world map with a set of initial markers. These markers constitute the observations (sampling sites) contained in the environmental data set. This data set contains the exact coordinates, "Latitude" and "Longitude", for each observation (sampling site) (cf. Figure 1). When clicked by the user, each marker exhibits the sampling code on a tag for reference.

On a see-trough top layer over the world map there is a form that allows the user to specify what should be visualized. This form is composed by two main input boxes and five action buttons. The first input box is for the selection of the taxon is a scroll-down window and enables the selection of up to five different taxa simultaneously. The other input box allows the selection of a single sampling site (Figure 2). The five action buttons enable different visualization options as follows.

Four main visualization options were considered to increase the functionality, interactivity and the ability to visually explore the data: i) "Simultaneous taxa prevalence sampling sites"; ii) "Pie charts of taxa prevalence sampling sites"; iii) "Pie charts of simultaneous taxa prevalence sampling sites" and iv) "Pie charts of all taxa prevalent".

The first visualization option, labeled in a button as "Simultaneous taxa prevalence sampling sites" on a button in the application, allows the visualization of the sampling sites in which the user-selected taxa occur simultaneously (cf. Figure 2a). These sampling sites are represented as markers.

The second visualization option is labeled in a button as "Pie charts of taxa prevalence sampling sites" and outputs in the map pie charts on the location of every sampling site in which at least one of the taxa selected by the user occurs (cf. Figure 2c). In case of a single taxon selection by the user, this option shows in the map the relative occurrence in each sampling site in which that taxon exists. This information is given by the size of the pie circle, meaning that larger circles correspond to higher prevalence sampling sites for that taxon (cf. Figure 2b).

The third visualization alternative is labeled in a button as "Pie charts of simultaneous taxa prevalence sampling sites" and permits the visualization of pie charts located on the sampling sites in which every taxa selected by the user occurs simultaneously (cf. Figures 2d and 2e).

The fourth, and last visualization option, labeled in a button as "Pie charts of all taxa prevalent", permits the visualization, through pie charts, of every taxa occuring on a single selected sampling site chosen by the user, allowing the visualization of the total composition of that specific sampling site (cf. Figure 2f). When the occurring pie charts are clicked by the user, a tag appears showing the concentration levels of each taxon selected as can be seen in Figures 2b, 2c and 2d.

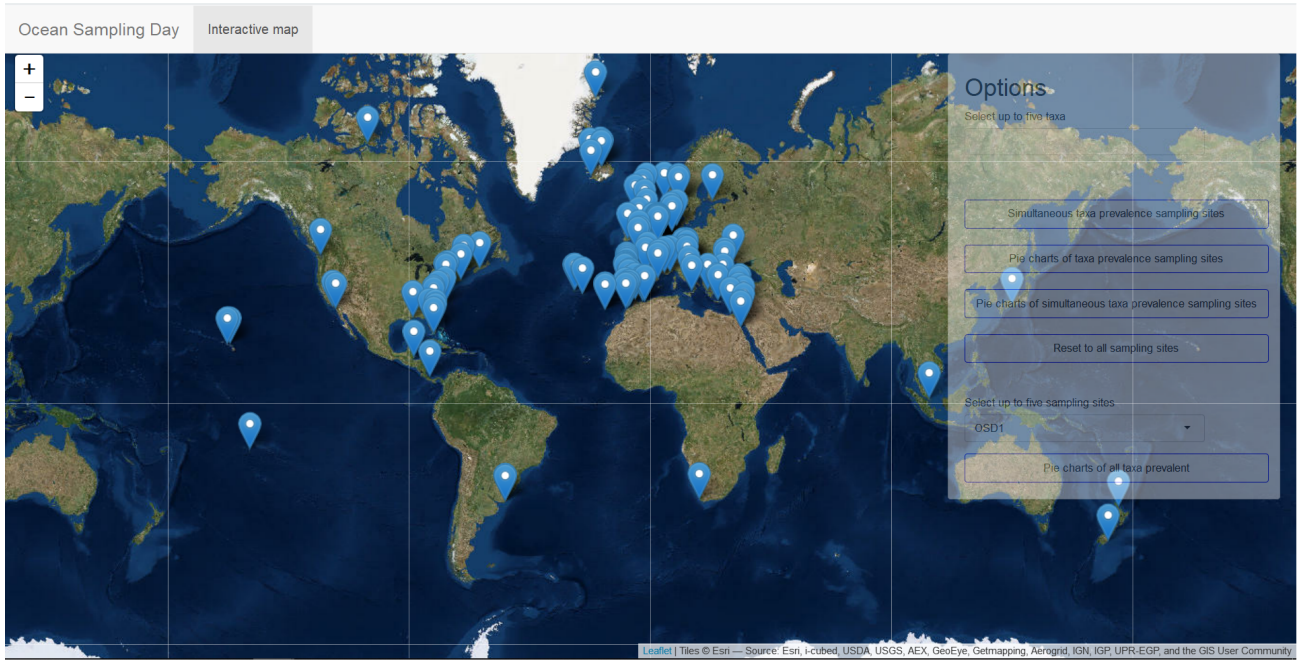


Fig. 1: Initial application interface showing the world map view with markers located in each OSD sampling site.

The button labeled as "Reset to all sampling sites", resets the application to its initial state where all the markers are shown in the world map.

4 Implementation Issues

As above mentioned, the Ocean Sampling Day App is a Shiny application developed in R [23] with shiny package [10].

The Shiny application is in a R script file and has two main functions: an `ui` function for layout and appearance customization and a `server` function for the processing of data and input/output instructions.

In order to build the spatial visualization web application, another two R packages called `leaflet` [11] and `leaflet.minicharts` [33] were used to generate an interactive world map with pie charts visualization at specific map points. For that purpose, a function named `leafletOutput` is used on the `ui` side of the script, while `renderLeaflet` is called inside the server function with defined parameters for the first rendering of the world when the app is initiated. The idea, at this point, is to render the map with markers located on the sampling sites.

The four visualization options were developed based on three Shiny functions: i) `selectizeInput` on the `ui` side of the script to enable multiple choices by the user, ii) `actionButton` on the `ui` side of the script to generate a button that triggers an observable action when clicked by the user and iii) `observeEvent` on the `server` side of the script that contains the code to perform the intended action only when the `actionButton` is activated i.e. when the specific button is clicked by the user. Basically, each of the four visualization methods consist of one `actionButton` function (clickable button by the user) and one `observeEvent` function (event to run the specific code to generate an action after the button is clicked by the user), while one of the `selectizeInput` function serves for three of the visualization methods and another `selectizeInput` for the remaining method. The reset option is implemented in a similar fashion, relying on the same type of functions described previously.

All the illustrated functions resort to another Leaflet function called `leafletProxy` that enables the leaflet map modification without having to render it again. The `leafletProxy` functions are contained in the `observeEvent` function of each different visualization technique and in the reset option.

In addition to the packages `shiny` [10], `leaflet` [11] and `leaflet.minicharts` [33], the application development included the package `dplyr` [31] for data manipulation and the package `colorRamps` [18] to access color palettes for the graphs.

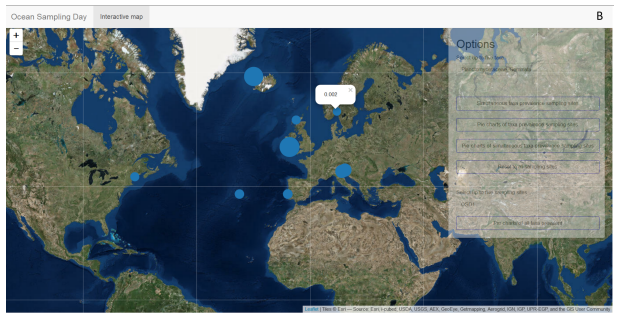
Additionally, a *css* file was used for interface customization: a file named "styles.css" retrieved and adapted from a publicly available Shiny application [5]. The function `includeCSS` was used on the `ui` side of the R script to incorporate the "styles.css" file. This file was used to control and modify the layout of the application, altering for example the opacity of the boxes containing the options of taxa or sampling selection and the five buttons previously described allowing a see-trough visualization of the map.

The R code and all the referred files are publicly available in a GitHub repository [4] and the web application is available online, deployed using shinyapps.io [7].

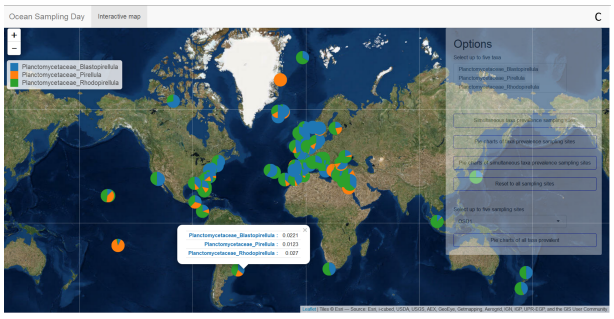
5 Conclusion

Sharing and communicating information is a problem often encountered when dealing with the ever-increasing volumes of data [32]. Data can be viewed as a potential source of valuable information and consequently knowledge [17]. However, not being able to explore and analyze the data leads to the waste of potentially relevant information [17]. So, understanding data may help draw better conclusions resulting, directly or indirectly, in innovation, business advantage, progress and profit [26]. Thus, the development of new tools and platforms to address this issue is a progressively essential task [26]. The `shiny` package of the R

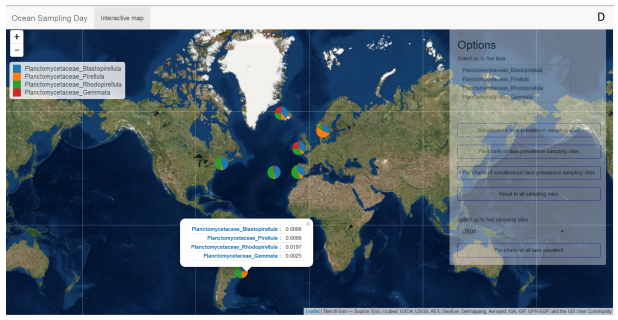
(a) Visualization of all sampling sites in which all the selected taxa occur simultaneously



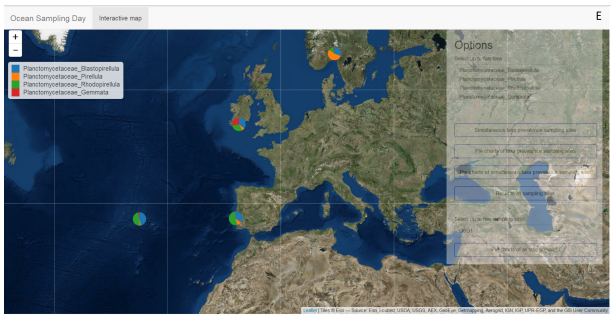
(b) Visualization of all sampling sites in which a single selected taxon occurs showing its relative abundance



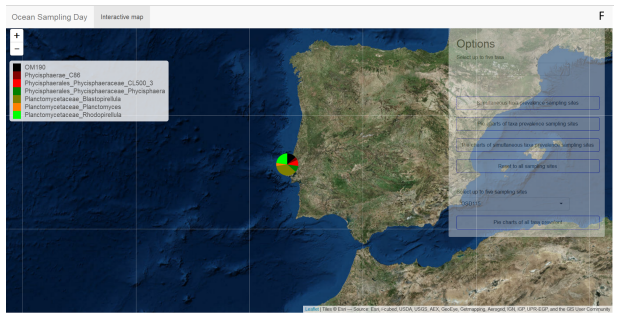
(c) Visualization of pie charts in all the sampling sites in which at least one of the taxa selected occurs



(d) Visualization of pie charts in all sampling sites in which every taxa selected occurs simultaneously



(e) Visualization of pie charts in all sampling sites in which every taxa selected occurs simultaneously, with zoom for a better visualization



(f) Visualization of the total composition of a selected sampling site with zoom for a better visualization

Fig. 2: Ocean Sampling Day Shiny application screenshots depicting the visualization functions.

language is a tool that facilitates this process, allowing the creation of simple yet informative web applications [32]. The present work aims to provide a solid, straightforward, easy-to-use environment for exploration, analysis and intuitive visualization of the OSD microbiome (Prokaryotes) taxonomic data. Hence, the visualization options proposed in this application ease the exploration and interpretation of data in a way that otherwise would not be possible with the same efficiency. This project can be seen as a first step contribution for the visualization and study of the marine microbiome at a global geographic scale. As the vast majority of the oceans and seas microbial communities are still very undocumented [12], it is of the most relevance to create novel tools that can in some way be useful to interpret and visualize valuable datasets available in public archives. Considering this, to further explore this work, new visualization methods could be added to enrich the application purpose. Furthermore, microorganism's ability to survive in extreme environmental conditions indicate a genetic adaptation and as a result the development of unique cellular biochemistry and metabolic pathways [14]. Therefore, enzymes produced by these microorganisms have the potential capacity to be used to many industrial, biotechnological and pharmaceutical purposes [12, 14]. As such, the visualization and exploration of microbiome communities gene functionality might be of utmost relevance. Our understanding is that the continued and increased development of new visualization platforms is essential to the further comprehension and analysis of the ocean's microbiome.

Acknowledgements

The authors would like to acknowledge all authors of the Open Sampling Day event that provided the valuable metagenomic and environmental data.

References

[1]Ebi metagenomics, pipeline version 3.0. https://www.ebi.ac.uk/metagenomics/pipelines/3.0. (acessed on June, 2018).

[2]Ena, european nucleotide archive. https://www.ebi.ac.uk/ena/ . (acessed on June, 2018).

[3]Ena, study: Prjeb5129. https://www.ebi.ac.uk/ena/data/view/PRJEB5129. (acessed on June, 2018).

[4]Github, osd application. https://github.com/miguelccfaria47/osdshinyapp.

[5]Github, superzip shiny r example. https://github.com/rstudio/shiny-examples/tree/master/063-superzip-example. (acessed on June, 2018).

[6]The micro b3 project. http://www.microb3.eu/osd. (acessed on June, 2018).

[7]Ocean sampling day web application. https://osdshinyapp.shinyapps.io/OceanSamplingDay/.

[8]Pangaea, data publisher for earth environmental science. http://www.pangaea.de/. (acessed on June, 2018).

[9]Lauren Bragg and Gene W. Tyson. *Metagenomics Using Next-Generation Sequencing*, pp. 183–201. Humana Press, Totowa, NJ, 2014.

[10]Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. R package version 1.0.5.

[11]Joe Cheng, Bhaskar Karambelkar, and Yihui Xie. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*, 2017. R package version 1.1.0.

[12]Calle Fernando. Marine microbiome as source of natural products. *Microbial Biotechnology*, 10(6):1293–1296.

[13]J. A. Fuhrman. Microbial community structure and its functional implications. *Nature*, 459(7244):193–199, May 2009.

[14]Kennedy J., O'Leary N.D., Kiran G.S., Morrissey J.P., O'Gara F., Selvin J., and Dobson A.D.W. Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. *Journal of Applied Microbiology*, 111(4):787–799.

[15]Sebastian Junemann, Nils Kleinbolting, Sebastian Jaenicke, Christian Henke, Julia Hassa, Johanna Nelkner, Yvonne Stolze, Stefan P. Albaum, Andreas Schlüter, Alexander Goesmann, Alexander Sczyrba, and Jens Stoye. Bioinformatics for ngs-based metagenomics and the application to biogas research. *Journal of Biotechnology*, 261:10 – 23, 2017.

[16]Eric Karsenti, Silvia G. Acinas, Peer Bork, Chris Bowler, Colomban De Vargas, Jeroen Raes, Matthew Sullivan, Detlev Arendt, Francesca Benzoni, Jean-Michel Claverie, Mick Follows, Gaby Gorsky, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Stefanie Kandels-Lewis, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Emmanuel Georges Reynaud, Christian Sardet, Michael E. Sieracki, Sabrina Speich, Didier Velayoudon, Jean Weissenbach, Patrick Wincker, and the Tara Oceans Consortium. A holistic approach to marine eco-systems biology. *PLOS Biology*, 9(10):1–5, 10 2011.

[17]Daniel Keim. Information visualization and visual data mining. 8:1–8, 01 2002.

[18]Tim Keitt. *colorRamps: Builds color tables*, 2012. R package version 2.3.

[19]Anna Kopf, Mesude Bicak, Renzo Kottmann, Julia Schnetzer, Ivaylo Kostadinov, Katja Lehmann, Antonio Fernandez-Guerra, Christian Jeanthon, Eyal Rahav, Matthias Ullrich, Antje Wichels, Gunnar Gerdts, Paraskevi Polymenakou, Georgios Kotoulas, Rania Siam, Rehab Abdallah, Eva Sonnenschein, Thierry Cariou, Fergal O'Gara, and Frank Glöckner. The ocean sampling day consortium. 4, 06 2015.

[20]Participants Ocean Sampling Day Consortium. Registry of samples and environmental context from the Ocean Sampling Day 2014, 2015.

[21]Anastasis Oulas, Christina Pavloudi, Paraskevi Polymenakou, Georgios Pavlopoulos, Nikolaos Papanikolaou, Georgios Kotoulas, C Arvanitidis, and Ioannis Iliopoulos. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. 9:75–88, 05 2015.

[22]Brendan P Hodkinson and Elizabeth A Grice. Next-generation sequencing: A review of technologies and tools for wound microbiome research. 4:50–58, 01 2015.

[23]R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[24]Justin R Seymour. A sea of microbes: the diversity and activity of marine microorganisms. 35:183, 01 2014.

[25]Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor Mcgee, and David L. Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. 469, 12 2015.

[26]José Fernando Rodrigues, Agma J. M. Traina, Maria Cristina Ferreira de Oliveira, and Caetano Traina. Reviewing data visualization: an analytical taxonomical study. *Tenth International Conference on Information Visualisation (IV'06)*, pp. 713–720, 2006.

[27]Douglas B Rusch, Aaron L Halpern, Granger Sutton, Karla B Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, Jonathan A Eisen, Jeff M Hoffman, Karin Remington, Karen Beeson, Bao Tran, Hamilton Smith, Holly Baden-Tillson, Clare Stewart, Joyce Thorpe, Jason Freeman, Cynthia Andrews-Pfannkoch, Joseph E Venter, Kelvin Li, Saul Kravitz, John F Heidelberg, Terry Utterback, Yu-Hui Rogers, Luisa I Falcon, Valeria Souza, Germán Bonilla-Rosso, Luis E Eguiarte, David M Karl, Shubha Sathyendranath, Trevor Platt, Eldredge Bermingham, Victor Gallardo, Giselle Tamayo-Castillo, Michael R Ferrari, Robert L Strausberg, Kenneth Nealson, Robert Friedman, Marvin Frazier, and J. Craig Venter. The sorcerer ii global ocean sampling expedition: Northwest atlantic through eastern tropical pacific. *PLOS Biology*, 5(3):1–34, 03 2007.

[28]Thomas Sharpton. An introduction to the analysis of shotgun metagenomic data. 5:209, 06 2014.

[29]S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, S. G. Acinas, P. Bork, E. Boss, C. Bowler, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, U. Krzic, F. Not, H. Ogata, S. Pesant, J. Raes, E. G. Reynaud, C. Sardet, M. Sieracki, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, D. Velayoudon, J. Weissenbach, and P. Wincker. Ocean plankton. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, May 2015.

[30]Jalpa R. Thakkar, Pritesh H. Sabara, and Prakash G. Koringa. *Exploring Metagenomes Using Next-Generation Sequencing*, pp. 29–40. Springer Singapore, Singapore, 2017.

[31]Hadley Wickham, Romain Francois, Lionel Henry, and Kirill Muller. *dplyr: A Grammar of Data Manipulation*, 2017. R package version 0.7.4.

[32]Jacek Wojciechowski, A Hopkins, and Richard N Upton. Interactive pharmacometric applications using r and the shiny package. In *CPT: pharmacometrics systems pharmacology*, 2015.

[33]Jalal-Edine ZAWAM, Benoit Thieurmel, and Francois Guillem. *leaflet.minicharts: Mini Charts for Interactive Maps*, 2018. R package version 0.5.4.