# Tools and a web server for data analysis and visualization in microbial ecology

Sergey Feranchuk [a b] (*), Natalia Belkova [a c], Ulyana Potapova [a] , Igor Ochirov [d] , Dmitry Kuzmin [e f], Sergei Belikov [a]

a) Limnological Institute, Siberian Branch of Russian Academy of Sciences, 664033 Irkutsk, Russia

(b) Department of Informatics, National Research Technical University, 664074 Irkutsk, Russia

(c) Scientific Centre for Family Health and Human Reproduction Problems, 664033 Irkutsk, Russia

(d) Regional Medical and Sports Clinic “Zdorovie”, 664003, Irkutsk

(e) Laboratory of Forest Genomics, Genome Research and Education Center, Siberian Federal University, 660036 Krasnoyarsk, Russia

(f) Department of High Performance Computing, Institute of Space and Information Technologies, Siberian Federal University, 660074 Krasnoyarsk, Russia

(*) Correspondce and reprints. Mail address: Limnological Institute SB RAS, Ulan-Batorskaya str., 3, 664033 Irkutsk, Russia, Tel: +7-3952-511874, Fax +7-3952-425405, feranchuk@gmail.com

Keywords: Microbiology, Bioinformatics, Biodiversity, 2D Graphics, Statistical Methods, Dissimilarities, Internet Service.

**Abstract**

The methods for data presentation are important in bioinformatics just as an data processing algorithms. Here, we describe the software package for the extensive analysis of tables with estimates of bacteria abundance levels in environmental samples. The package was designed to be executed in a distributed hardware environment, with powerful packages in Python in a back-end and interactive front-end form. Most the microbial ecology-specific functionality is implemented by the scikit-bio Python package, together with the other Python packages intended for big data analysis. The interactive visualisation tools are implemented by the D3.js software library, therefore, the software project is named D3b. The package is a suite of tools for the analysis of microbial ecology data implemented as a web-service and as a desktop application with freely available source codes. It supports a substantial part of the graphical and analytical descriptions of microbial communities used in scientific publications. An on-line version of the system is available at d3b-charts.bri-shur.com.

**Abbreviations**

OTU - Operional Taxonomic Unit, CA - Correspondence Analysis, PCA - Principal Component Analysis, PCoA - Principal Coordinates Analysis

Introduction

The contribution of microbiology to biological research has been highlighted with advances in sequencing technologies, as shown on Fig. 1. Newly developed tools for data analysis in microbiology need to be able to adapt to a growing number of possible interpretations of the data; the rich variety of tasks requires tools to address an almost continuous spectrum of questions.

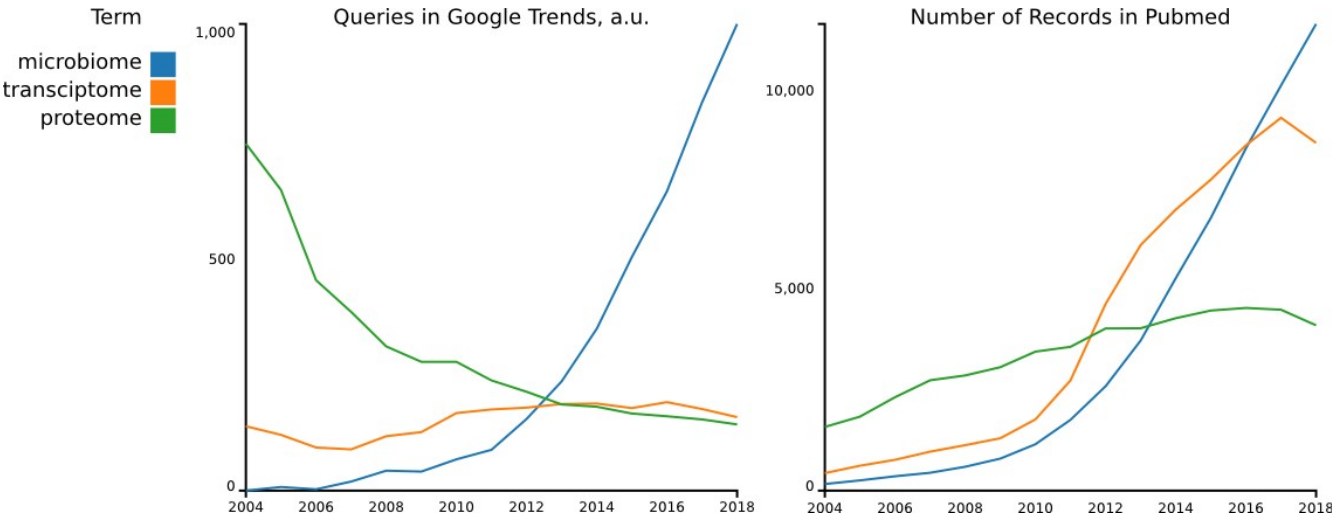


*Fig. 1*

*The growth in interest related to analysis of the microbiome, demonstrated using a chart provided by Google Trends (left) and the annual number of publications in PubMed (right).*

Amplicon gene sequencing remains one of the most important methods to study microbial communities (Segata et al., 2013; Boughner and Singh, 2016; Hugerth and Andersson, 2017), and the conventional bioinformatics pipelines used to process sequencing data in these studies can be roughly divided into two stages, a resource-consuming stage of sequence clustering and taxonomic annotation, and different types of analysis and interpretation, including diversity calculations, hypothesis testing,

and data visualisation. The first stage, considered in details, is subdivided into picking the amplified fragments of the reference gene into operational taxonomic units (OTU), and the assignment of the taxonomic annotation to each OTU. In the conventional systems of microbial taxonomy, seven levels are used to annotate the bacteria: phylum, class, order, family, genus, and species. The annotation of OTUs might be not precise enough and identifcation might be limited to a genus or even a higher level of the hierarchy (Almeida et al., 2018).

The most informative experiments include the simultaneous processing of several samples. Therefore, the results from the first stage of pipeline can usually be presented in a relatively compact form of abundance tables for taxonomic units, i.e. the number of reads included in each OTU for each of the samples. This tabular data is supported by standardised formats, such as the “Biology observation matrix” (BIOM) format (McDonald et al., 2012). In addition, for whole metagenome and metatranscriptome sequencing, software packages such as sortmerna (Kopylova et al., 2016) enable the representation of the microbiome community composition under study in the form of a table of abundance counts.

The intergated downstream pipeline with all the conventional types of data analysis and visualisation (such as diversity estimates, heatmaps, and PCA charts) could be useful to present a wide range of experiments. However, in some situations, the scientist should find a clue to the interpretation of the experiment, and the interactive flexible tools for data presentation could help in this task.

The unification of mathematical concepts from different areas of science could be implemented within a single software infrastructure such as python, so that the software interface for an Euclidean measure in geometry and Hamming distance in informatics could be almost the same as the Bray-Curtis and Jackard dissimilarity measures in ecology, by using the 'scipy.spatial.distance' python

library. This opens a possibility to satisfy the declared requirements of a wide flexibility of data analysis tools for microbiologists. The expressive and clear graphics contribute valuably to clarifying the interpretations of the data, however, software tools developed to manage graphics objects at a sufficiently high level of abstraction also allow the results to be represented in the most convenient ways.

The other side to the observed development and unification of software libraries is, firstly, the care taken to ensure the mutual consistency of heterogeneous packages which are composed together to support the wide spectrum of representations on another end of a software system. A signs that the problem of inconsistencies is growing is the development of packaging systems, such as Anaconda, at an increasing rate. Furthermore, the scarcity of resources directed to software development and to the teaching of newly developed tools leads to problems such competition and increased inequality in the allocation of that resource, revealing the problems typical of the most of societies, even with an increased contrast.

The list of software packages developed for microbiologists, presented below as an overwiew of related projects, is definitely incomplete. However, several important features could be stressed which could in some way characterise all of the listed packages. Firstly, many of the packages are designed to be incorporated into a wide software infrastructure, such as the C++, R, or python environment. Secondly, the most important and most costly effort which accomplishes the development of a package with a sufficient usability is the care taken to ensure the mutual consistency of component libraries. It is achieved either when software tools are provided as a web-services (RDP: Cole et al., 2014), or are incorporated into a single binary, with source codes which are mostly independent from auxiliary libraries (Mothur: Schloss et al., 2009). In the QIIME project (Caporaso et al., 2010), the balance is achieved when, several robust core pipelines are implemented, supported by specialised python

libraries and uclust software where elaborated algorithms are incorporated into the pipelines. This software project allows the easy incorporation of additional expansions, such as sortmerna/sumaclust. The consistency of the project is supported by a distribution within the anaconda packaging system. Just as the QIIME project effectively uses the advantages of a development within a python environment, the vegan project (Oksanen et al., 2007) uses the advantages of a development within an R environment, with an ability to use general-purpose tools from data science. And, aside from a microbiology software, two long-lasting projects for visualisation in structural bioinformatics, UCSF Chimera (Pettersen et al., 2004), and Pymol (Schrödinger, LLC), also use the advantages of a development within a python environment.

Since the volume of expected traffic and the computational costs in the downstream analysis are relatively low, web-based tools are well-suited for the implementation of the user interface in the interactive data presentation system. JavaScript and node.js packaging systems could be considered as an infrastructure to implement efficient online visualisation tools. The Biojs project (Yachdav et al, 2015) is an example of the general-purpose bioinformatics environment within a JavaScript infrastructure, similarly to projects such as biopython or bioruby in python and ruby, respectively.

The JavaScript infrastructure attracts the attention of developers aside from bioinformaticians, and the D3.js project (Bostock et al., 2011) should be considered as a universal framework for the development of online applications, such as those used for semantic analysis and text mining (Borke and Härdle, 2015) and in healthcare (Schroeder et al., 2017). The tools provided by D3.js are comparable in efficiency with the universal graphics libraries used in Python and R environments (matplotlib, ggplot2), and these libraries are often used to prepare publication-quality images in microbiology research projects.

## 2. Methods

### *2.1 Overview*

The design of the D3b system was composed from several parts: the JavaScript-based tools for data presentation on a client side, the back-end tools for data processing on the server side, and a web framework on the side of the hub server which manages user queries between the client and server sides. The instruments from the D3.js library were used to generate expressive images based of the input data and the user queries defined as html-based input forms. The python environment was chosen as an infrastructure for the development of tools for data processing on the server side, and this allows us to use a functionality of both the scikit-bio library with the function specific to microbial ecology, and general-purpose libraries such as scipy and sklearn.

Since the resources which could be directed to the development of the D3b package were knowingly limited, the project was not aimed to compete with packages where a lot of effort had been directed to support the consistency of the project with other packages within the infrastructure. The project is deposited in github in source codes as an installable package (sferanchuk/d3b_charts), however, the installation of the package might require additional efforts due to inconsistencies with updates and modifications introduced in the new releases of the dependencies used in the package.

Instead, the aim of this project was to satisfy the requirements from a local community of microbiologists from the Irkutsk scientific center and related organisations, therefore, the on-line version of the project is continuously supported at the bri-shur.com site. The bri-shur project has a long enough history, and, by design, the web-interface in that project is separated from the servers rendering

data processing (Feranchuk et al., 2012). Therefore, in the present version, the infrastructure used in a hub server is implemented within the django framework in python, however, it is relatively independent from both sides of the system and could be easily substituted to another type of web framework in a further development or adaptation of the project.

The rendering of graphics in a JavaScript framework is implemented in the user browser, however, the stable and consistent functionality for export of graphics in raster and vector formats is obligatory for that kind of system. Therefore, the phantomjs command-line tool was adopted to run on the server side, as a converter of dynamic html pages to PNG and SVG formats. An export of the tables generated on-line is possible using a conventional functionality of the browser and, in addition, export to tab-delimited format is supported in some cases.

The input data for the analysis, in the form of abundance tables, can be supplied in a biom format and tab-delimited format. When the input file is submitted, access becomes available in the menu of several tools which could be used to analyse the submitted data from different viewpoints. The privacy of the data is supported by assigning an unique url with a secure 32-byte key to access a page with any of data tables. To be able to specify any subset of rows and columns of the submitted table, two kinds of descriptors could be assigned on-line, and stored together with an input table. The first kind of descriptor, in the form of tag-value pairs, allows the selection of traits associated with each of the samples in the survey. The second kind of descriptor, in the form of a list of taxonomic identifiers, allows the analysis to be focused on specific taxonomic units, rather than on the whole content of the microbiome.

The figures below which illustrate the services described are based on the surveys described in (Feranchuk et al., 2018). Namely, one of the surveys which presents bacterial symbionts of marine

sponges in coral reefs near an Indonesian shore, as described in (Cleary et al., 2017) . The second survey presents the gut microbiome of A. indicus geese with different breeding patterns, as described in (Wang et al., 2016). The third survey presents the microbiome of jaw bone osteomyelitis in patents with two types of disease, as described in (Goda et al., 2014).

## 2.1 Specifications

The estimates of alpha-diversity adopted in the system are based on implementations from the scikit-bio package (Shannon, Simpson, Chao1, Ace, Fisher α, Gini). In addition, an alternative estimator of Gini measure and two parametric diversity measures, as specified in (Feranchuk et al., 2018), are included in the set of estimators. The rarefaction analysis uses a python code adapted from the scikit-bio package, and the ability to estimate the Michaelis-Menthen fit to rarefaction curves is included as an alpha-diversity estimator.

The set of metrics for distances between samples include weighted and unweighted Unifrac measures, as these are implemented in the scikit-bio package, and a set of metrics composed of Bray-Curtis similarity, Jackard similarity, and Euclidean distance, as these are implemented in the scipy package. The measures of Pearson, Spearman, and Kendall correlations are transformed into distances just by subtracting the values of the correlation from its maximal value: d = 2 - c. In addition, a Morisita-Horn measure is included in the set of metrics, using a python code incorporated in the system.

These distances could be used to construct dendograms of proximity between samples, to run Permanova and Anosim tests which are implemented in the scikit-bio package, and to run a principal coordinate analysis, correspondence analysis (implemented in scikit-bio package), principal component

analysis or multi-dimensional scaling (implemented in the sklearn package).

## 3. Results

### *3.1 Overview*

The tabular presentations of the data include a table of abundances, just as it is loaded into the system, and with various options for sorting the rows, merging the columns, reducing the level of taxonomic hierarchy, and others. Similar options are available for most of the services within the interactive system. Namely, most of the input forms include the choice of level of taxonomic hierarchy, the possibility to restrict the analysis or data presentation to certain taxonomic groups, and the possibility to merge samples into pre-defined groups.

The tabular presentations also include:

1) the values of alpha-diversity calculated using several of the most informative estimators,

2) the significance of differences for alpha-diversity values between several groups, calculated following the methods described in (Feranchuk et al., 2018),

3) the significance of differences between groups of samples calculated using the distances between samples, with several alternative measures of distance.

The graphical presentations, implemented with the use of d3.js library, include following charts:

1) A bubble chart and heatmap, to represent absolute/relative abundances.

2) 2D scatter charts, to represent the results of several data ordination methods, such as principal component analysis (PCA) or multi-dimensional scaling (MDS). The choice of several measures is available here to calculate distances between the samples.

3) A dendrogram (tree) to represent the degree of proximity between samples.

4) A Venn diagram to represent the unique and shared taxonomic units for the samples, implemented with the use of the jVenn plugin (Bardou et al., 2014) and venn.js library.

5) Two kind of diagrams to present distributions which describe a sample or a group of samples: a rank-abundance chart (Whittaker plot) to represent the distribution of relative species abundance, and a rarefaction curve to estimate the effect of insufficient coverage and sample size.

6) A ternary chart, to represent the relative abundances of bacterial phylotypes for the three samples or groups of samples.

7) Two combined 2D charts, to represent the results of PCA decomposition applied directly to a non-square matrix of abundances. One chart is for the samples in the survey and the second adjacent chart is for bacterial species in the rows of the submitted matrix.

*3.2 Case studies*

Figure 2 illustrates the variations in microbial communities for different families of marine sponges, at the level of the classes. The colors in Figure 2 are assigned at the level of phyla, the most general category of taxonomy. The presented result confirms both the original study (Cleary et al., 2017) and wider research where the composition of sponge microbiomes was compared (Thomas et al., 2016). In the second paper, the relative composition of the microbiomes was also demonstrated in Fig. 3 at the level the phyla, using a heatmap chart. In particular, the Proteobacteria phylum is the most abundant in all sponge families, which is demonstrated in from Fig. 2 and in the cited paper.

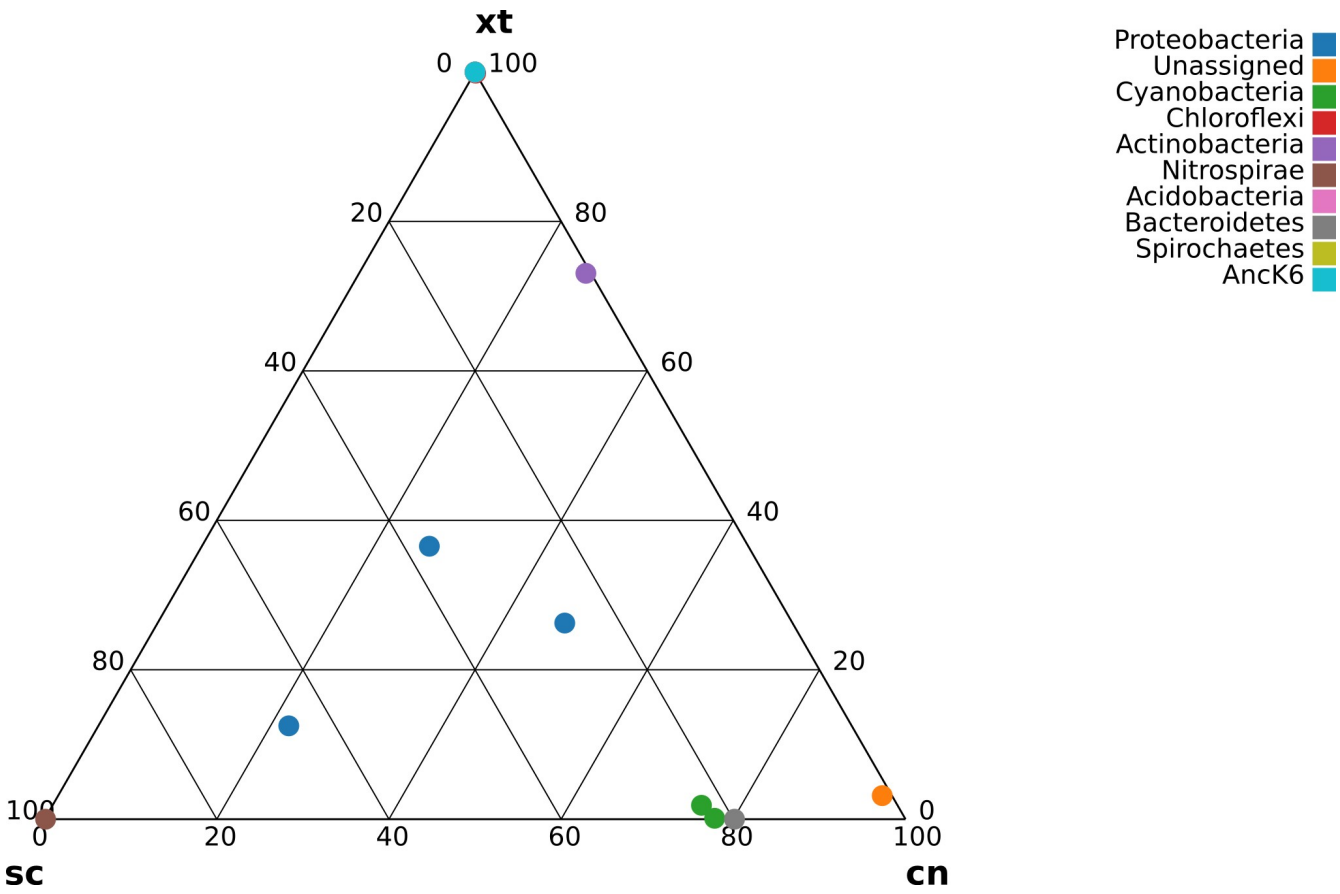


*Fig. 2*

*Example of a ternary chart. Quantitative composition of symbionts for three types of marine sponges at the level of bacterial classes, for 10 most abundant classes. Species of sponges are abberviated as follows: Xt, Xestospongia testudinaria (Haplosclerida order); Sc, Stylissa carteri (Halichondrida order); Cn, Cinachyrella (Spirophorida order). Colors on the chart are assigned at the level of phyla.*

It can be seen in Fig. 2 that several classes are specific to Xestospongia sponge, and this is confirmed in Table 1, where the values of biodiversity are presented at the class level. This table was also generated using online tools included in the described system. The significance for the separation of the samples is also shown in Table 1, using three statistical tests applied to lists of diversity values.

*Table 1*

*Estimates of alpha-diversity and the significance for the separation of microbial communties for the sponge samples.*

For the T-test and Mann-Whitney rank-based test, the minimal p-value, from six combinations of traits is shown. Species of sponges are abbreviated as follows: ap, *Aaptos suberitoides* (Suberitida order); xt, *Xestospongia testudinaria* (Haplosclerida order); sc, *Stylissa carteri* (Halichondrida order); cn, *Cinachyrella* (Spirophorida order).

| Samle ID | Species | Shannon | Simpson | Fisher Alpha | Otu Number | Chao1 | Ace | Gini |
|---|---|---|---|---|---|---|---|---|
| PapBSAp1Mer1 | Ap | 2.87 | 0.83 | 2.91 | 21 | 21.2 | 22.27 | 0.77 |
| PapBSAp1Mer2 | Ap | 2.8 | 0.81 | 2.57 | 15 | 15.5 | 17.61 | 0.7 |
| PapBSAp1Mer5 | Ap | 2.64 | 0.8 | 1.9 | 12 | 12 | 13.11 | 0.66 |
| PapBSCn1Kr02 | Cn | 2.22 | 0.73 | 1.77 | 12 | 12 | 12 | 0.76 |
| PapBSCn1Ms17 | Cn | 2.49 | 0.79 | 1.68 | 12 | 12 | 12.64 | 0.69 |
| PapBSCn2Kr02 | Cn | 1.96 | 0.69 | 1.14 | 8 | 8 | 8 | 0.71 |
| PapBSCn2Ms17 | Cn | 2.39 | 0.78 | 2.47 | 19 | 19 | 19.38 | 0.82 |
| PapBSCn3Kr02 | Cn | 2.95 | 0.84 | 3.35 | 23 | 23.1 | 24.19 | 0.77 |
| PapBSCn3Ms17 | Cn | 2.5 | 0.79 | 2.75 | 20 | 20.33 | 21.09 | 0.82 |
| PapBSSc1Mer1 | Sc | 1.67 | 0.52 | 1.74 | 11 | 11 | 11.59 | 0.84 |
| PapBSSc1Mer2 | Sc | 1.4 | 0.42 | 1.38 | 10 | 11 | 13.9 | 0.86 |
| PapBSSc1Mer5 | Sc | 1.7 | 0.62 | 1.32 | 9 | 9.5 | 12 | 0.81 |
| PapBSXt1Mer2 | Xt | 2.94 | 0.81 | 2.48 | 18 | 18 | 18.68 | 0.72 |
| PapBSXt1Mer5 | Xt | 2.89 | 0.81 | 3.12 | 15 | 15 | 15 | 0.67 |
| PapBSXt2Mer1 | Xt | 3.05 | 0.83 | 2.55 | 17 | 17 | 17.93 | 0.67 |
| *Student's T-test (min. value)* | | *<0.001* | *0.005* | *0.008* | *0.003* | *0.004* | *0.024* | *0.002* |
| *Mann-Whithney test (min. value)* | *p-value* | *0.014* | *0.014* | *0.04* | *0.04* | *0.038* | *0.04* | *0.04* |
| *Anova test* | | *<0.001* | *<0.001* | *0.137* | *0.272* | *0.334* | *0.553* | *0.012* |

Figure 3 illustrate significant differences in the composition of gut microbiome of A. indicus geese, depending on the breeding pattern. In the original study (Wang et al., 2016), a result similar to the one shown in Fig. 3 was presented in Fig. 4B using bar charts.

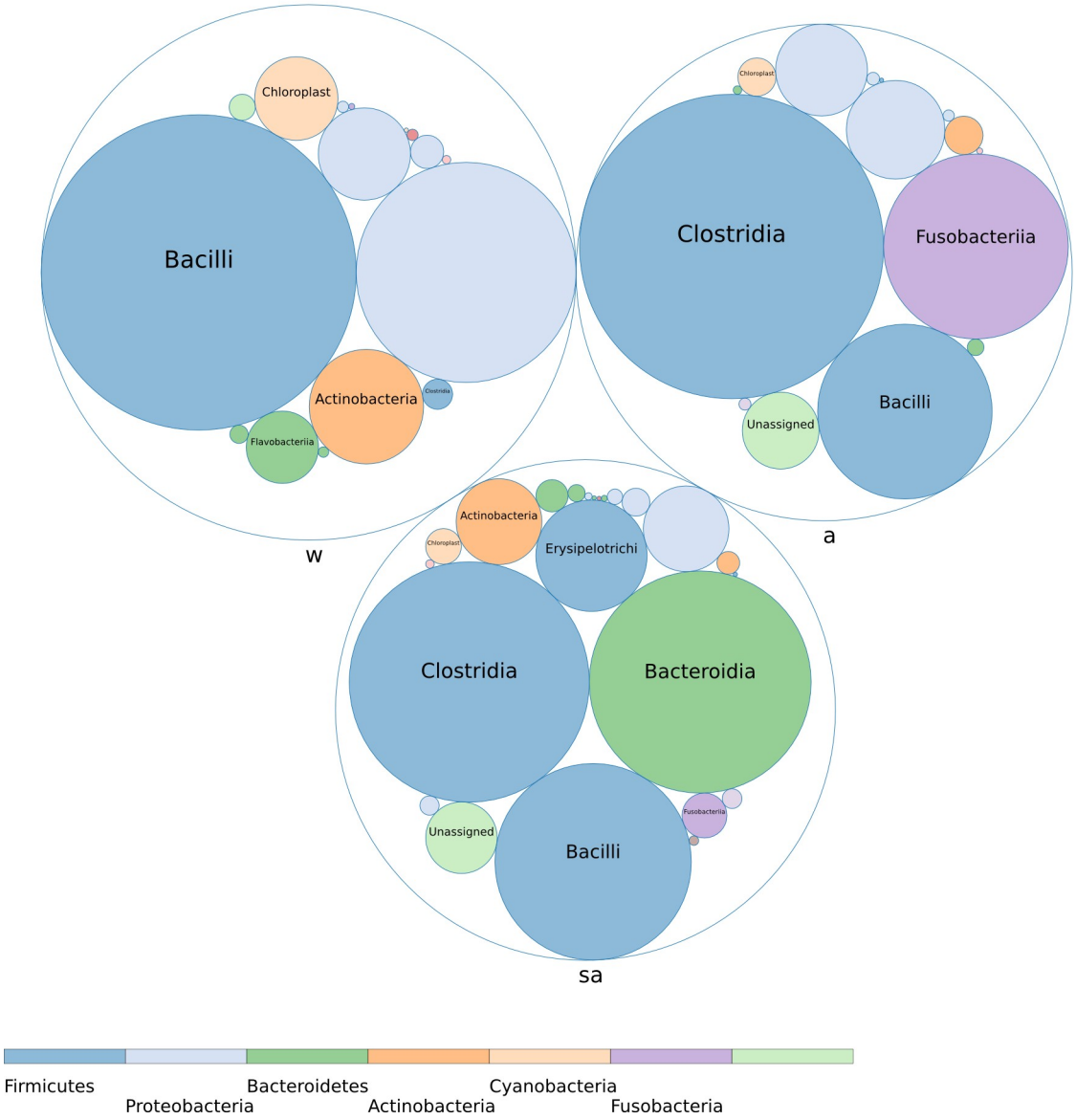


*Fig. 3*

*Example of a bubble chart showing the composition of the gut microbiome of A. indicus at the level of bacterial classes. Traits are assigned as follows: a, artificial breeding; w, wild type breeding; sa, semi-artificial breeding.*

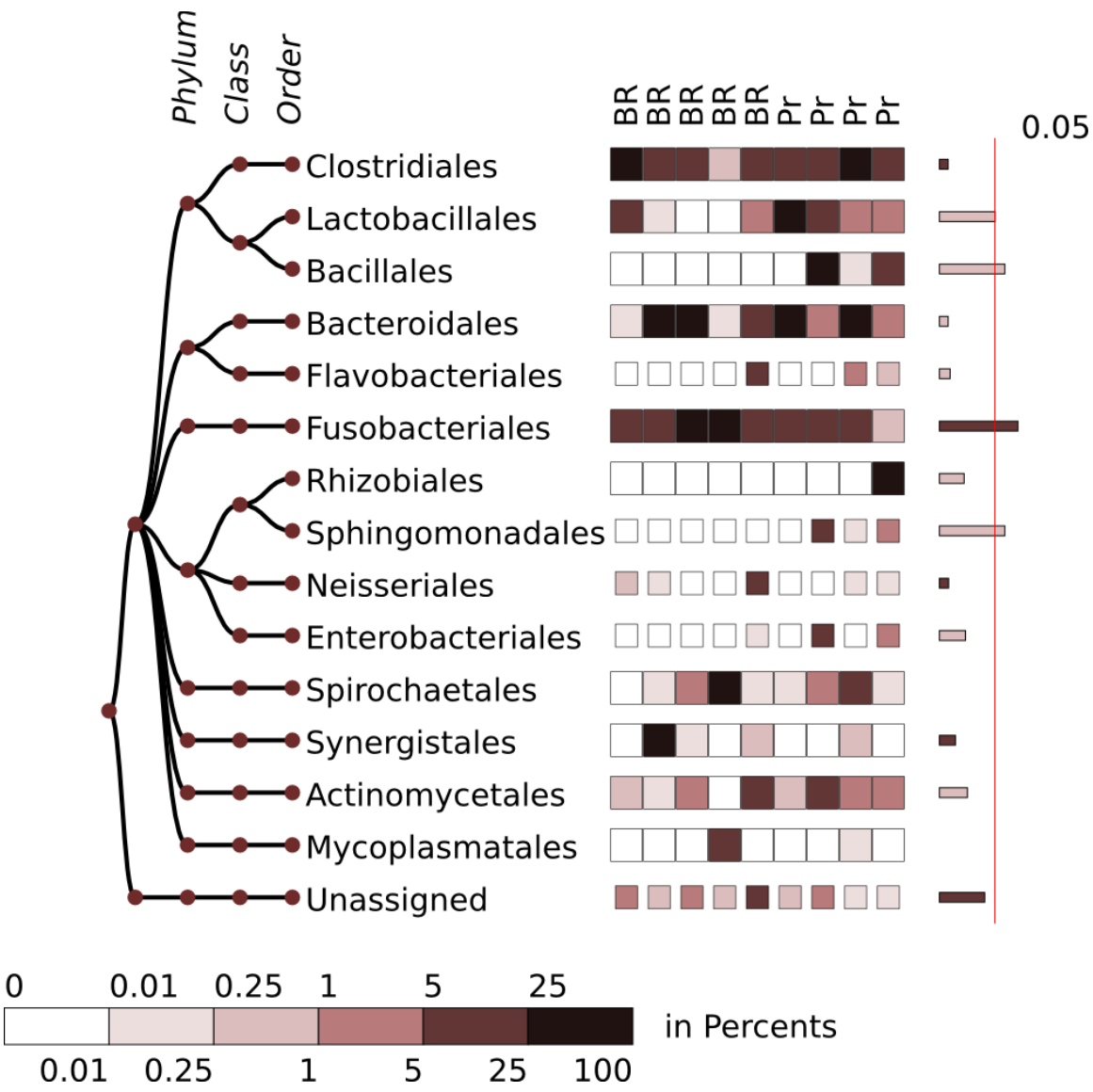


*Fig. 4*

*Example of a heatmap chart showing the composition of the microbiome in jaw bone osteomyelitis. Pr; primary osteomyelitis, BR; bisphosphonate-related osteonecrosis of the jaw. The 10 most abundant bacterial orders are shown. The orders present with an average abundance >2 % are highlighted. The scale of the heatmap indicates the abundance values as a percentage. The bars on the right show the degree of separation between two groups for each of the presented phylotypes. The width of the bars is proportional to the -log(p-value) for the Mann-Whitney test.*

Figure 4 illustrates the composition of the microbiome in inflamed jaw bones, in an agreement with Fig. 1 in the original study (Goda et al., 2014), where a similar result was presented using a bar chart. In the cited study the emphasis was placed to a wide diversity of microbiomes in the inflamed bones, with a prevalence of anaerobic bacteria for most samples from patients with different diagnoses. The presentation in Fig. 4 might also confirm this conclusion. Although the relative presence of abundant phylotypes might vary in the different samples, no specific bacterial phylotype was specific to each of the sub-diagnoses with a sufficient confidence.

## 4. Discussion

With the richness of material and a wide spectrum of possible applied results, scientific research in microbiology becomes, to a great extent, a creative work rather than routine investigation or a description of the observed phenomenon within pre-defined rigorous templates. In addition, as is common for any genre of creative work, a niche could easily be found to motivate the development of any concept or idea, even if it has an unlikely chance to generate a direct profit of any kind. The same is true for the development of a software for microbiologists.

The system described above features several kinds of thoroughly designed graphic presentations, and it is intended to provide the flexibility to choose a precise query and the best way describe an applied problem under study. Although it might be not sufficient or satisfactory to fit some templates which would allow it to be classified as a mature and consistent software project, it can be described as a completed kind of creative work, and might be of interest in some way to an extended audience of microbiologists.

The problem of understanding and separating the responsibility of different scientists, either

working in the same team or just distantly related, occurs in many situations. However, the development and testing of the system occurred in close collaboration with software developers and biologists with different areas and levels of competence, which contributed to its usability and flexibility.

## Conflict of Interest

The authors declare that there is no conflict of interest.

## Acknowledgements

The authors appreciate the contributions from the specialists from NIPCHI, LIN, SIFIBR, and SFU to the development of the system, and are personally grateful to L. Chernogor, L. Mironova, A. Ponomaryova, A. Gladkikh, A. Krasnopeev, Y. Putintseva, Y. Markova, and I. Petrushin. S.F. and U.P. thank C.H. Brown for a long-term support of their research activity.
This study was supported by budget projects of Federal Agency of Scientific Organizations number 497 0345-2019-0002 and Russian Foundation for Basic research (RFBR) grant numbers: 16-04-498 00065; 16-54-150007; 18-04-00224.

## References

Almeida, A., Mitchell, A.L., Tarkowska, A. and Finn, R.D. 2018. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience* 7:1-10.

Bardou, P., Mariette, J., Escudié, F., Djemiel, C., Klopp, C. 2014. jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 15(1):293.

Borke, L. and Härdle, W. 2016. Q3-D3-LSA. *SFB 649 Discussion Paper* 2016-049.

Bostock, M., Ogievetsky, V. and Heer, J. D3: data-driven documents. 2011. *IEEE Trans. Vis. Comput. Graph.* 17(12):2301-2309.

Boughner, L.A. and Singh, P. 2016. Microbial ecology: where are we now? *Postdoc J.* 4(11):3-17.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7:335-336.

Cleary, D.F.R., Polonia, A.R.M., Becking, L.E., de Voogd, N.J., Purwanto, Gomes, H. and Gomes, N.C.M. 2017. Compositional analysis of bacterial communities in seawater, sediment, and sponges in the Misool coral reef system, Indonesia. *Mar. Biodiv.* Epub ahead of print 23 April 2017;

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R. and Tiedje, J.M. 2014. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucl. Acids. Res.* 42:D633-D642

Feranchuk, S., Belkova, N., Potapova, U., Kuzmin, D. and Belikov S. 2018. Evaluating the use of

diversity indices to distinguish between microbial communities with different traits. *Res. Microbiol.* 169:254-261.

Feranchuk, S., Potapova, U., Potapov, V., Mukha, D., Nikolaev, V. and Belikov, S. 2012. Tools for protein structure prediction at the bri-shur.com web portal. *J. Life Sci.* 6:1074-1079.

Goda, A., Maruyama, F., Michi, Y., Nakagawa, I. and Harada, K. 2014. Analysis of the factors affecting the formation of the microbiome associated with chronic osteomyelitis of the jaw. *Clin. Microbiol. Infect.* 20:O309-0317.

Hugerth, L.W. and Andersson, A.F. 2017. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Front. Microbiol.* 8:1561.

Kopylova, E., Navas-Molina, J.A., Mercier, C., Xu, Z.Z., Mahé, F., He, Y., Zhou, H.W., Rognes, T., Caporaso, J.G. and Knight R. 2016. Open-source sequence clustering methods improve the state of the art. *mSystems* 1:e00003-15

McDonald, D., and Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R. and Caporaso, J.G. 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* 1:7.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H. and Oksanen, M.J. 2007. The vegan package. *Community ecology package.* 10:631-637.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E.

2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput Chem.* 25(13):1605-1612.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hatmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J. and Weber, C.F. 2009. Introducing Mothur: open-source, platform-independent community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75:7537-7541.

Schroeder, J., Hoffswell, J., Chung, C.F., Fogarty, J., Munson, S. and Zia, J. 2017. Supporting patient-provider collaboration to identify individual triggers using food and symptom journals. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* 1726-1739.

Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S. and Huttenhower, C. 2013. Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* 9:666.

Thomas, T., Moitinho-Silva, L., Lurgi, M., Björk, J.R., Easson, C., Astudillo-García, C., Olson, J.B., Erwin, P.M., López-Legentil, S., Luter, H., Chaves-Fonnegra, A., Costa, R., Schupp, P.J., Steindler, L., Erpenbeck, D., Gilbert, J., Knight R., Ackermann, G., Victor Lopez, J., Taylor, M.W., Thacker, R.W., Montoya, J.M., Hentschel, U. and Webster, N.S. 2016. Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* 7:11870.

Wang, W., Cao, J., Li, J.R., Yang, F., Li, Z. and Li, L.X.2016. Comparative analysis of the gastrointestinal microbial communities of bar-headed goose (Anser indicus) in different breeding

patterns by high-throughput sequencing. *Microbiol. Res.* 182:59-67.

Yachdav, G., Goldberg, T., Wilzbach, S., Dao, D., Shih, I., Choudhary, S., Crouch, S., Franz, M., García, A., García, L.J., Grüning, B.A., Inupakutika, D., Sillitoe, I., Thanki A.S., Vieira, B., Villaveces, J.M., Schneider, M.V., Lewis, S., Pettifer, S., Rost, B. and Corpas, M. 2015. Anatomy of BioJS, an open source community for the life sciences. *Elife* 8:4.