Published online 29 April 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W509–W513***
doi: 10.1093/nar/gkx304

Interactive microbial distribution analysis using BioAtlas

Jesper Beltoft Lund[1], Markus List[2] and Jan Baumbach[1,2,*]

[1]Department of Mathematics and Computer Science (IMADA), University of Southern Denmark, 5000 Odense, Denmark and [2]Max Planck Institute for Informatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany

Received February 12, 2017; Revised April 07, 2017; Editorial Decision April 08, 2017; Accepted April 12, 2017

ABSTRACT

Massive amounts of 16S rRNA sequencing data have been stored in publicly accessible databases, such as GOLD, SILVA, GreenGenes (GG), and the Ribosomal Database Project (RDP). Many of these sequences are tagged with geo-locations. Nevertheless, researchers currently lack a user-friendly tool to analyze microbial distribution in a location-specific context. BioAtlas is an interactive web application that closes this gap between sequence databases, taxonomy profiling and geo/body-location information. It enables users to browse taxonomically annotated sequences across (i) the world map, (ii) human body maps and (iii) user-defined maps. It further allows for (iv) uploading of own sample data, which can be placed on existing maps to (v) browse the distribution of the associated taxonomies. Finally, BioAtlas enables users to (vi) contribute custom maps (e.g. for plants or animals) and to map taxonomies to pre-defined map locations. In summary, BioAtlas facilitates map-supported browsing of public 16S rRNA sequence data and analyses of user-provided sequences without requiring manual mapping to taxonomies and existing databases. Availability: http://bioatlas.compbio.sdu.dk/

INTRODUCTION

Profiling of 16S ribosomal RNA (16S rRNA) has proven useful for deciphering the microbial composition of environmental samples (1). The 16S rRNA gene contains both, regions of high evolutionary conservation as well as hypervariable regions. 16S rRNA sequencing data is thus ideal for phylogenetic analyses and for the differentiation of bacterial kingdoms and species (2). Another important factor in microbiome research, besides the composition of an individual sample, is the sample origin and the spatial distribution of the associated prokaryotic species. For instance, some prokaryotes are only found on a specific continent or in a particular latitude (3). Moreover, considering the mircobiome of an individual organism, we expect a different composition of the skin microbiome in comparison to the one in the gut. Despite the importance of this information, software tools that enable users to study the spatial distribution of published microbiomes are currently missing. Moreover, while methods for assigning taxa to 16S rRNA samples exist, users currently lack a straight-forward method to determine where samples with the same taxa have previously been sampled.

We present BioAtlas (http://bioatlas.compbio.sdu.dk), a web application that utilizes 16S rRNA sequencing data for geo-profiling. BioAtlas supports two types of maps, namely geographical world maps and host-attached maps of, for instance, the human skin. BioAtlas includes two rich public data sets that demonstrate the utility of the system for microbiome geo-profiling.

Global geo-profiling with Google Maps

The first data set stems from two integrated databases: the Integrated Microbial Genomes (IMG) database (4) hosts genomic sequences, while the Genomes OnLine Database (5) (GOLD) hosts associated meta data. IMG and GOLD hold the largest collection of community-contributed sequences of 16S rRNA, making it an essential resource for microbiome research. Interestingly, GOLD provides information on the sampling site of 16S rRNA sequences but not on their taxa. To exploit this underutilized information for geo-profiling, BioAtlas places these samples on an interactive Google Maps powered world map. This aids researchers in tracing data suitable for their study, in understanding global distribution of individual species and in directing future efforts for global microbiome analysis, for instance, by allowing researchers to choose sampling sites that were not previously covered. However, to facilitate geo-profiling of 16S rRNA sequences from IMG on the species or genus level, corresponding taxa first need to be determined.

Exploring microbiome data with host-attached maps

The second data set published by Bouslimani *et al.* (6) provides a microbiome cartography of the human skin. The

*To whom correspondence should be addressed. Tel: +45 6550 2309; Email: jan.baumbach@imada.sdu.dk

study involved one subject of each gender. For each subject, the authors took ∼400 samples from 24 larger locations and analyzed their microbial composition using 16S rRNA amplicon sequencing. Sequences were subsequently mapped to 851 distinct prokaryotic taxa. We mapped the samples from these study to images of gender-specific body maps using the map editor functionality of BioAtlas.

User-provided data

Data sets available in BioAtlas can also be explored in the context of additional, user-provided 16S rRNA sequences. Suitable taxa are automatically assigned using the Mothur classifier (see below). As a novelty, BioAtlas provides a simple and user-friendly work-flow for creating custom maps based on simple image files. User-uploaded 16S rRNA sequences can subsequently be placed freely on the user-provided maps and thus allow users to extend BioAtlas for new application scenarios.

DATA PROCESSING

To deal with non-annotated 16S rRNA sequences in BioAtlas, a robust 16S rRNA classification method as well as a reference database of taxonomically annotated sequences is required. We thus evaluated existing reference databases, including SILVA(7), GreenGenes(8) (GG) and the Ribosomal Database Project(9) (RDP). Table 1 summarizes our findings. SILVA appeared best suited since it contains the largest number of sequences and since it is regularly updated. We used SILVA SSU Ref NR 99, a non-redundant version of the high quality control database storing 645 151 sequences.

Several methods are widely used in the microbiome research community for assigning taxa based on 16S rRNA sequencing data, including Mothur (10), QIIME (11), 16S Classifier (12) and RDP Classifier (13). Resource and time efficiency are crucial factors in a web application environment. Consequently, we chose Mothur, which features a fast implementation of the RDP algorithm in C++.

GOLD and IMG data of global microbiome diversity

As the largest community-driven database for microbial research, GOLD contains a total of 21 786 analysis projects, of which 2627 contain both geographical and 16S rRNA sequence data from IMG. Of these, we extracted a total of 459 304 16S rRNA sequences for taxonomical classification in Mothur.

We executed Mothur with the following run parameters: k-mer size 8, confidence cutoff 80 and 100 iterations. Mothur was able to classify 423 543 sequences (92.2%). 126 431 (29.8%) were classified down to genus level, yielding 1805 distinct taxons that can be selected in BioAtlas (Figure 1).

Microbiome diversity map of the human skin

We downloaded a data set by Bouslimani *et al.* (6), which contains mappings of more than 800 16S rRNA sequences to 24 distinct regions of the human body for both genders. These data were imported into BioAtlas using the map editor. We used pictures of the human body obtained from

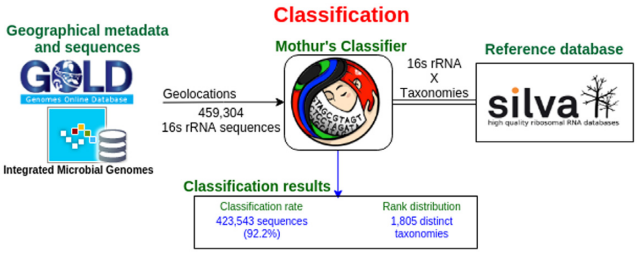

Figure 1. Work-flow for classification of 16S rRNA sequence from IMG and GOLD using Mothur and the Silva reference database.

http://opencliparts.com to generate two interactive maps (one for each gender) with different viewing angles. Taxa assigned by Bouslimani *et al.* were automatically mapped to SILVA IDs to construct phylogenetic trees for browsing available taxa.

IMPLEMENTATION

BioAtlas is implemented in Python 3 using the bare bone web framework Flask (https://github.com/ctsit/barebones-flask-app). Data are stored in a MySQL database. For geo-profiling, we build upon the GoogleMaps API (https://developers.google.com/maps/). User-uploaded 16S rRNA sequences are processed and mapped to the world- and user maps using a dedicated job scheduler that follows the pipeline shown in Figure 6.

User authentication

To facilitate anonymous access, BioAtlas creates a unique token ID for each new browser session. Uploaded data are not disclosed to other users and deleted frequently. Optionally, users can create a user account, which will allow them to store their data long-term. Secure user management and authentication is implemented using 32-bit UUID objects according to RFC 4122 (https://www.ietf.org/rfc/rfc4122.txt). These UUIDs also facilitate sharing results with other users through a corresponding link that is shown on the respective result page and that embeds the UUID.

INTERFACE AND FUNCTIONALITY

The BioAtlas web interface features three types of interactive analyses:

Global geo-profiling with Google Maps

BioAtlas enables users to browse projects with geolocations imported from the GOLD and IMG databases on an interactive world map (http://bioatlas.compbio.sdu.dk/map/). Here, each marker depicts an individual sample site. All sequences were mapped to taxonomies and can be filtered accordingly. To this end, BioAtlas features a phylogenetic tree that shows the hierarchy of taxons available on the corresponding map (Figure 2). Users can select one or several taxa in the tree and a search bar further allows for finding prokaryotes of interest quickly (Figure 3).

If many markers overlap, the map view may quickly appear cluttered. In this case, users can switch to a heat map

Table 1. Number of sequences, with domains and date of last release for the SILVA, RDP and GG databases. [1]16S rRNA sequences for prokaryotes

	SILVA	RDP	GG
# sequences[1]	5 616 941	3 356 809	1 262 986
Domains	Prokaryota, eukaryota	Prokaryota, fungal	Prokaryota
Last release	September 2016	September 2016	May 2013

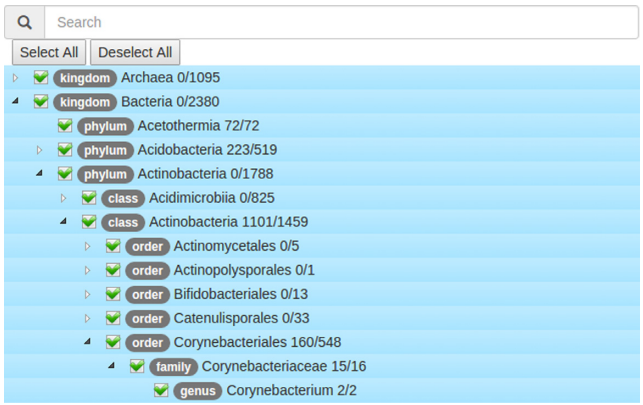

Figure 2. Hierarchical phylogenetic tree featured on the front end, allowing users to interact with a given map based on specific prokaryota selected.

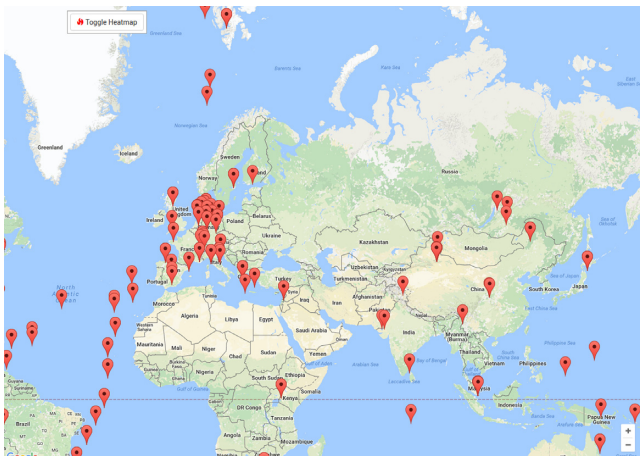

Figure 3. On the world map, each marker represents an individual analysis project imported from the GOLD database.

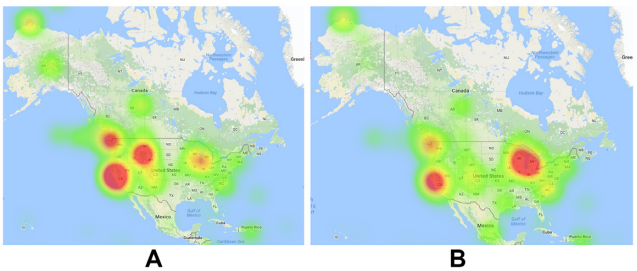

Figure 4. Heat map comparison for prokaryota kingdoms in USA. (**A**) The archaeal heatmap. (**B**) The bacterial heatmap. Red: high abundance, green: low abundance

representation where large concentrations of markers can easily be spotted (Figure 4). To better distinguish the heat map from the terrain, the map can be switched to a gray scale version.

Browsing the distribution of microbial taxonomies across user-defined maps

The map view features a list of the currently shown markers to the right. The selection of entries is synchronized with the map to allow users to focus on specific information. To gain more information, users can either click on markers on the map or in the list. A popup-view shows a color gradient indicating the percentage of taxonomies in this particular location in comparison to the overall taxonomies on this map (Figure 5).

User-provided data

BioAtlas encourages users to upload additional 16S rRNA sequences through the classification interface (http://bioatlas.compbio.sdu.dk/classifier). Uploaded sequences are processed as jobs and submitted to the Mothur classifier previously described for the GOLD and IMG sequences. After job completion, users can browse their sequences on the associated map or identify the location of previously published sequences with the same taxa as their own sequences. For convenience, jobs can be named and customized run parameters for Mothur can be selected. Sequences are accepted in FASTA format. However, when uploading several sequence files, these need to be compressed as a single file in either the zip (**.zip**) or tarball (**.tar.gz**) format.. Jobs started by the scheduler use the pipeline depicted in Figure 6 for processing.

After a job is completed, results can be accessed on a job page consisting of the tabs 'Results and statistics', 'Geo-map' and 'User-map'.

On the **Results and statistics** page, a list of sampling sites is shown along with pie charts depicting classification results such as species-, kingdom- and rank- distributions, along with a list of taxa for each sequence. Figure 7 shows an example for a species distribution chart for 2000 submitted prokaryote sequences.

On the **Geo-map** page, the world map is displayed. Markers are filtered based on the intersection of the resulting taxons from the current Mothur run and the full map data based on GOLD and IMG.

On the **User-map** page, we list all publicly available user maps as well as maps uploaded by the current user. As in the global map, user maps show only taxa that were both assigned in the current Mothur run and that are already part of the respective map.

Building own user maps

BioAtlas features an online editor, where users can generate own maps based on image files. Subsequently, locations

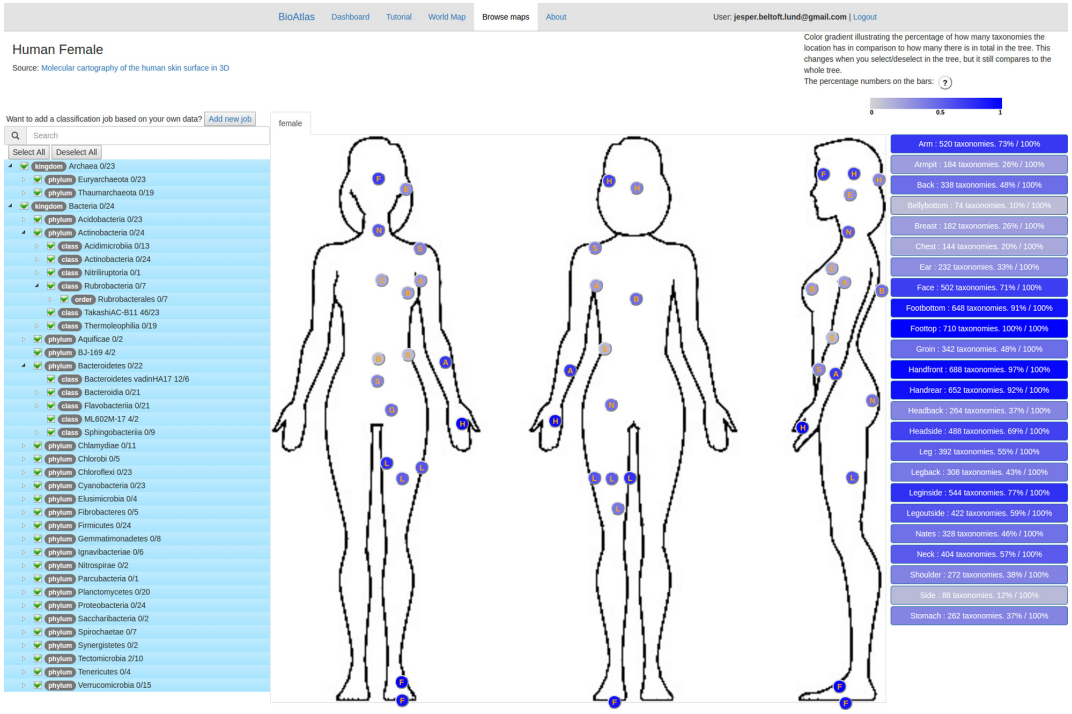

Figure 5. The view for the Human Female map. A phylogenetic tree (left) and a list sample locations (right) can be used to filter results.

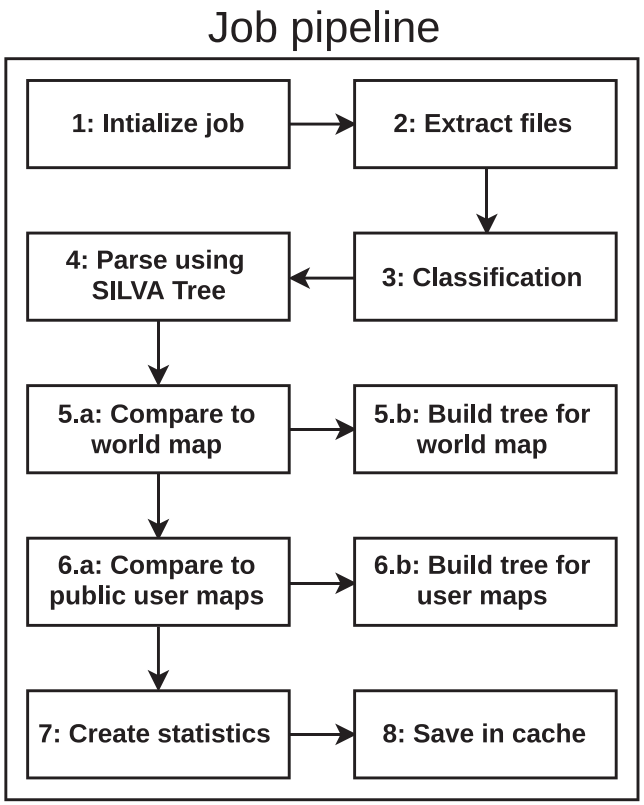

Figure 6. An overview of the processing pipeline to process user-submitted 16S rRNA sequences.

can be defined as prokaryote sampling sites. In principle, any image file can serve as a map, giving users freedom to consider, for instance, any host of interest e.g. plants, animals, a representation of a complex soil sample, etc. User-generated maps are private by default but can be released to the public on request. To demonstrate the features of the map editor, we used this system to create maps depicting a human male and female from varying angles (e.g. http://bioatlas.compbio.sdu.dk/hostmap/26). These custom maps can be used to place user-uploaded 16S rRNA sequences or simply to explore the map with respect to already uploaded samples.

CONCLUSION

BioAtlas is a user-friendly web application that enables scientists to study microbial 16S rRNA data in a location-specific context. As an online service, it does not require a complex setup or time-demanding processing of public data. With its Google Maps powered world map, users can, for the first time, utilize location information stored along with 16S rRNA sequences to study the global distribution of samples collected in microbiome studies. Results can conveniently be filtered by one or several taxa. BioAtlas enables scientists to identify under-represented areas on the world map and could thus be instrumental for determining optimal routes for future studies with the long-term goal to obtain a dense map of global microbial diversity. Similar to existing tools, BioAtlas facilitates assignment of genus-level taxa to unclassified sequences using Mothur. We note that BioAtlas is limited to the features offered by Mothur and can, for instance, not offer species-level annotations. As a distinct feature, however, BioAtlas enables researchers to

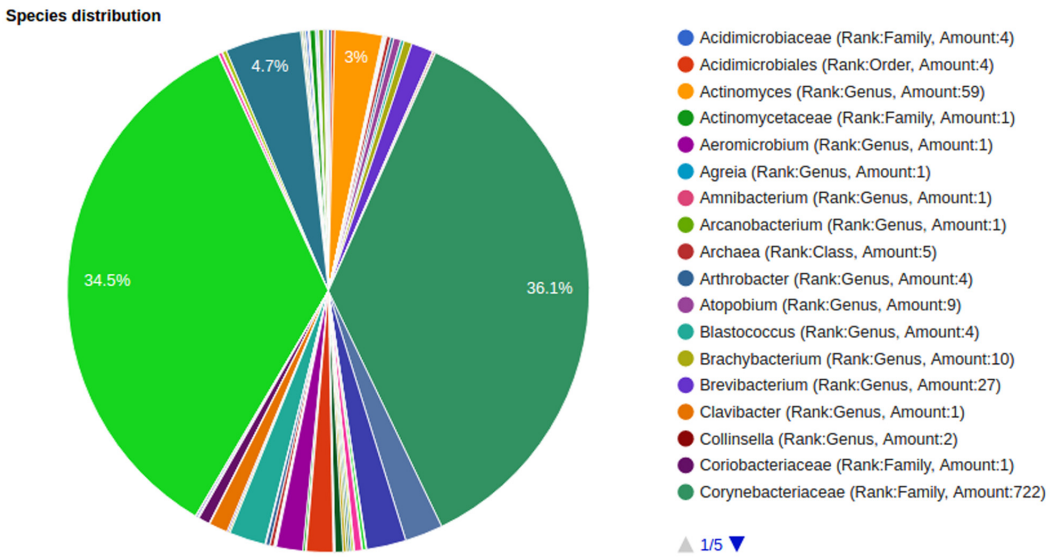

Figure 7. Species distribution for 2,000 bacteria sequences with the phylum rank of Actinobacteria.

place samples on arbitrary maps as we demonstrate by example of a host-attached maps of a human male and female. This facilitates better accessibility and visual representation of 16S rRNA profiling data, which may be useful, for example, as interactive supplemental material in future publications. We will also evaluate additional potential marker genes (14) in the future. For now BioAtlas is a prokaryote-specific system, but we will evaluate marker genes for eukaryotic profiling in the future and prepare a new version with corresponding support.

ACKNOWLEDGEMENTS

JB is grateful for financial support from his VILLUM Young Investigator Grant. Funding for open access charge: VILLUM Young Investigator Grant of Jan Baumbach.

FUNDING

VILLUM Young Investigator Grant (to J.B.). Funding for open access charge: VILLUM Young Investigator Grant nr. 13154 of Jan Baumbach.
Conflict of interest statement. None declared.

REFERENCES

1. McCaig,A.E., Glover,L.A. and Prosser,J.I. (1999) Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Appl. Environ. Microbiol.*, **65**, 1721–1730.
2. Chakravorty,S., Helb,D., Burday,M., Connell,N. and Alland,D. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods*, **69**, 330–339.
3. Green,J. and Bohannan,B.J. (2006) Spatial scaling of microbial biodiversity. *Trends Ecol. Evol.*, **21**, 501–507.
4. Chen,I.M.A., Markowitz,V.M., Chu,K., Palaniappan,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Andersen,E., Huntemann,M. *et al.* (2016) IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.*, **45**, D507–D516.
5. Mukherjee,S., Stamatis,D., Bertsch,J., Ovchinnikova,G., Verezemska,O., Isbandi,M., Thomas,A.D., Ali,R., Sharma,K., Kyrpides,N.C. *et al.* (2017) Genomes OnLine Database (GOLD) v. 6: data updates and feature enhancements. *Nucleic Acids Res.*, **45**, D446–D456.
6. Bouslimani,A., Porto,C., Rath,C.M., Wang,M., Guo,Y., Gonzalez,A., Berg-Lyon,D., Ackermann,G., Christensen,G.J.M., Nakatsuji,T. *et al.* (2015) Molecular cartography of the human skin surface in 3D. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E2120–E2129.
7. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
8. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
9. 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. (2013) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, **42**, D633–D642.
10. Schloss,P.D., Westcott,S.L., Ryabin,T., Hall,J.R., Hartmann,M., Hollister,E.B., Lesniewski,R.A., Oakley,B.B., Parks,D.H., Robinson,C.J. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.
11. 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. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.
12. Chaudhary,N., Sharma,A.K., Agarwal,P., Gupta,A. and Sharma,V.K. (2015) 16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS One*, **10**, e0116106.
13. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.
14. Baumbach,J. (2010) On the power and limits of evolutionary conservation–unraveling bacterial gene regulatory networks. *Nucleic Acids Res.*, **38**, 7877–7884.