Databases of the marine metagenomics

Katsuhiko Mineta, Takashi Gojobori

PII: S0378-1119(15)01243-3
DOI: doi: 10.1016/j.gene.2015.10.035
Reference: GENE 40941

To appear in: *Gene*

Please cite this article as: Mineta, Katsuhiko, Gojobori, Takashi, Databases of the marine metagenomics, *Gene* (2015), doi: 10.1016/j.gene.2015.10.035

Title

[Mini Review] Databases of the marine metagenomics

Authors

Katsuhiko Mineta, Takashi Gojobori

Affiliation

Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

Corresponding author

Takashi Gojobori, Ph.D.

Distinguished Professor,

Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

E-mail: takashi.gojobori@kaust.edu.sa

Abstract

The metagenomic data obtained from marine environments is significantly useful for understanding marine microbial communities. In comparison with the conventional amplicon-based approach of metagenomics, the recent shotgun sequencing-based approach has become a powerful tool that provides an efficient way of grasping a diversity of the entire microbial community at a sampling point in the sea. However, this approach accelerates accumulation of the metagenome data as well as increase of data complexity. Moreover, when metagenomic approach is used for monitoring a time change of marine environments at multiple locations of the seawater, accumulation of metagenomics data will become tremendous with an enormous speed. Because this kind of situation has started becoming of reality at many marine research institutions and stations all over the world, it looks obvious that the data management and analysis will be confronted by the so-called Big Data issues such as how the database can be constructed in an efficient way and how useful knowledge should be extracted from a vast amount of the data. In this review, we summarize the outline of all the major databases of marine metagenome that are currently publically available, noting that database exclusively on marine metagenome is none but the number of metagenome databases including marine metagenome data are six, unexpectedly still small. We also extend our explanation to the databases, as reference database we call, that will be useful for constructing a marine metagenome database as well as complementing important information with the database. Then, we would point out a number of challenges to be conquered in constructing the marine metagenome database.

Keywords

Metagenomics, database, bioinformatics, biodiversity, comparative metagenomics, marine science, microbe, environment

Highlights

1. An overview of available databases for marine metagenome is described
2. Reliable and comprehensive reference databases are crucial for metagenomic analysis
3. The comparative study of the metagenomics is an inevitable approach to utilizing massive metagenome data in the databases

1. Introduction

1.1 Microbial diversity and metagenome

Microbes are found in everywhere, particularly in a natural environment such as in soil, water and air. Moreover, microbes thrive in an amazing diversity of environmental conditions such as different degrees of temperature, radiation, pressure, gravity, vacuum, desiccation, salinity, pH, oxygen tension and chemical extremes (Rothschild and Mancinelli, 2001). In these diverse environments, microbes compose a wide variety of communities that are often adapted for given environmental conditions (Cowan et al., 2015).

The studies of microbial community will help us to understand the repertoire of microbes adapted in specialized niches, leading to eventually understanding of the mechanisms in microbial dynamics by which they interact with each other in the biosphere. Dynamic changes in the diversity of microorganisms can be utilized for monitoring the environmental conditions to predict disastrous and harmful changes in the environments. It is also useful for conducting effective exploration of novel and useful proteins and metabolites for industrial application. In fact, the huge repertoire of microbes can be considered as valuable resources for potential drugs and materials.

A term "metagenome" was used by Handelsman et al. in 1998 as “the genomes of the total microbiota found in nature”, refers to sequence data directly sampled from

the environments (Handelsman et al., 1998). In other words, metagenome is an efficient method to examine a diversity of the microbial community. Because of its broad application, metagenome has become a very popular method particularly when it is used together with the next-generation sequencing (NGS) technologies.

In these situations, a huge amount of data has been produced in the metagenomic studies. It is no doubt that without a proper management of such huge data, any significant outcome should not be obtained from any metagenomic studies. Thus, it is obvious that construction of the database is an important key to ensure successful developments of the metagenomic studies.

1.2 Marine metagenomics

Approximately 3.67×10^{30} microorganisms are considered to be living in the marine environments (Whitman et al., 1998), noting that approximately 71% of the Earth surface is covered by the ocean (Kennedy et al., 2008). A huge diversity of marine microbes is reasonably conceivable, which should be an important target for the studies of marine science as well as exploitable biotechnologies. Metagenomics is surely a powerful tool for surveying a diversity of marine microbes.

One of the milestones in marine metagenomics is an expedition that was conducted by Venter et al. at the Sargasso Sea (Venter et al., 2004). More than a

million of genes previously undiscovered were found in sequenced DNA fragments, leading to a potential discovery of new biochemical functions.

The *Sorcerer II* expeditions (2003–2010) (Rusch et al., 2007; Yooseph et al., 2007; Gross, 2007) and the Malaspina expedition (2010–2011) (Laursen, 2011) conducted global surveys of prokaryotic metagenomes from the surface of the ocean and bathypelagic layer of more than 1,000 m, respectively. Moreover, it is noteworthy that the most recent topic on marine metagenomics was brought by a TARA ocean expedition (Bork et al., 2015; Sunagawa et al., 2015). This expedition was done by an international effort from 2009 to 2013. Their findings show a surprisingly high level of biodiversity in the oceans, unveiling hidden interactions between these microorganisms. They also showed how serious impact planktons give impact to the biodiversity of marine microbes, identifying several million novel genes.

These studies are typical examples of how metagenomic sequence data can be translated into understanding of the impact of microbes on their local environment and the influence of the environment on microbial communities. In practice, from the metagenomic sequence data, functional genes were inferred from the related databases, as references, using sophisticated bioinformatics tools. In order to make this practice

possible, construction of the marine metagenome database is crucial with proper functional annotations of the sequencing data.

1.3 Growth of marine metagenomics data

Marine metagenomic studies are producing a huge amount of sequence data from which an increasing number of new species of plankton, bacteria, and viruses were discovered. The DNA Data Bank of Japan (DDBJ) (Nakamura et al., 2013; Kosuge et al., 2014), which is a collaborating member of the International Nucleotide Sequence Databases (INSDs: DDBJ/EMBL/GenBank), collects all nucleotide sequence data worldwide. According to the statistics reports for DDBJ release 101 (June, 2015; http://www.ddbj.nig.ac.jp/documents-e.html), a total of 3,196,890 entries were found for the entry "marine metagenome", corresponding to 2,486,893,637 nucleotides. Because the most extensively deposited data in DDBJ is of *Homo sapiens* with 20,946,173 entries for a total of 17,738,676,173 nucleotides, the marine metagenomic data in the DDBJ accounts for almost 15% of the data for human. Taking into account the fact that only less than 20 years passed since Handelsman et al. (1998) proposed the definition of metagenome, the marine metagenome data has accumulated very rapidly.

1.4 Data production by two approaches in marine metagenomics

A rapid increase of the marine metagenome data is mostly due to the recent progress in sequencing capabilities of the NGS technology. Two different approaches are used in NGS-based metagenomic studies; an amplicon-based approach and a shotgun sequencing-based approach.

The amplicon-based approach using rRNA genes as target is the most extensively used method in marine metagenomic studies. PCR amplifies conserved regions in the

16S rRNA gene (for bacteria) that contains enough resolution of the sequence divergence to distinguish between different bacterial species (Woese and Fox, 1977; Pace, 1997). This approach generates a large number of 16S rRNA gene fragments from diverse communities of microbes in a cost effective and speedy way. Similarly, 18S rRNA genes are used for identification of eukaryotic microbes.

On the other hand, the shotgun-based approach is more time consuming and expensive. However, this approach produces a large number of short sequences (200-1000 bp) derived from different regions of the genomes, not just the rRNA gene. After assembly of the fragmented sequences and homology search against the reference database were conducted, specific genes and species can be identified. As reviewed (Kunin et al., 2008; Teeling and Glockner, 2012; Thomas et al., 2012; Kim et al., 2013; Sharpton, 2014; Behzad et al., 2015), a large numbers of sequence fragments generated by this approach require extensive bioinformatics analyses to ensure proper interpretation of the sequence data. One of the main advantages of the shotgun-based approach over the amplicon-based approach is an ability of examining the entire genome of microbes. In addition to detection of biodiversity, shotgun-based approach is also used routinely to identify characteristic sequences and novel genes.

In short, the amplicon-based and shotgun-based approaches are complementary, being used either or both for answering different questions particularly in marine metagenomic research.

2. Reference database for marine metagenome

Marine metagenome databases reviewed here can be divided into two types by their usages. One is the database that is used for construction of the metagenome database as a reference of functional annotation, for example. The other is the database for collecting marine metagenome data. In this review, we call the former

type of database as "reference database" and the latter simply as a "marine metagenome database."

As described earlier, there are two different approaches in metagenomics; amplicon-based and shotgun-based. The metagenome data generated from the amplicon-based method is distinctly different than those generated by shotgun-based approach. As a result, different reference databases should be used for database construction of the marine metagenome.

2.1 Reference database for shotgun-based marine metagenome

In the shotgun-based approach, the obtained data is an output of random sequences that are derived from various regions of the genomes of different species in the samples examined. Therefore, for functional annotation, the reference database should contain a universal set of data representing all types of genes/proteins as well as intergenic regions.

The most representative primary databases for shotgun marine metagenome data is INSD (DDBJ/ENA/GenBank) (Nakamura et al., 2013) because INSD contains all the nucleotide sequences.

RefSeq (Tatusova et al., 2015) and UniProtKB/Swiss-Prot (Boutet et al., 2007; UniProt, 2015) are also useful resources because the data is intensively curated for

users' utilities; *i.e.*, reduced noise through removal of duplicates and insufficient annotations. Since these reference databases contain sequence data with annotation, they can provide a platform for searching for homologous sequences to user's query sequence against these reference databases, making appropriate inferences on composition of genes as well as species in the user data.

Of course, other various types of databases are utilized for construction of the marine metagenome database, but their usage may depend heavily upon the scope of the database to be constructed.

2.2 Reference database for amplicon-based metagenome

The reference database for amplicon-based metagenome contains all the rRNA sequence data (16S rRNA genes for bacteria and archaea and 18S rRNA genes for eukaryotes). There are millions of known rRNA genes in the primary databases (*i.e.* INSD): However, only specific sequences are required for taxonomic identification. In addition, databases such as INSD contain a large set of uncertain rRNA sequences that are derived from unknown and/or uncultured organisms. These rRNA sequence data often cause the noise in taxonomic assignments and phylogenetic classification. It is, therefore, better to use reference databases that contain only taxonomically relevant sequences. In other words, the reference databases should be a comprehensive collection of taxonomically relevant sequence data that are derived from various metagenomes.

Table 1 provides a list of the major reference databases available for 16S rRNA

marine metagenomic data. Each of these databases is outlined as follows:

The Ribosomal Database Project (RDP) provides the phylogenetic classification of prokaryotes and eukaryotes (Cole et al., 2014). RDP collects the rRNA sequences from the INSD and classifies them into appropriate categories for use in phylogenetic analysis. RDP provides not only rRNA gene data but also various bioinformatics tools for phylogenetic classification such as an aligner, a comparison tool and a hierarchical browser. RDP Release 11 update 4 (as of May 26, 2015) contains 3,224,600 entries of 16S rRNA genes.

Greengenes also provides phylogenetic classifications for the 16S rRNA sequences in the INSD as well as the web application (DeSantis et al., 2006). Unfortunately, the most recent update of this database is as of May 2013, which is a bit obsolete. The data can be obtained only by the download at Greengenes website, but their phylogenetic classification is still useful as a reference.

SILVA provides phylogenetic classifications for the small and large rRNA subunit sequences (SSU and LSU, respectively) of prokaryotes and eukaryotes in the INSD (Quast et al., 2013). SILVA contains two different data sets: One is Parc that is intended to make overview of a broad diversity of the organisms, and the other is SSU/LSU Ref, a subset of Parc that contains only the high quality and full-length sequences for phylogenetic analysis and probe design. SILVA release 123 (as of July 23, 2015) contains 4,985,791 entries in SSU-Parc, 1,757,783 entries in SSU-Ref, 563,332 entries in LSU-Parc and 96,642 entries in LSU-Ref categories. As we mentioned earlier, these reference databases are useful for conducting sequence filtration for maintaining the quality enough to make the phylogenetic classification. In the case of SILVA, the original data is retrieved from INSD. The SSU-Parc category of SILVA release 123 is derived from 7,168,241 SSU of EMBL-EBI/ENA release 123. SILVA is useful for performing removal of the short sequences, ambiguous sequences, homopolymers,

vector contamination, and other low-quality sequences. SILVA is an official database of ARB, a program package for sequence analyses (Ludwig et al., 2004).

EzTaxon and EzTaxon-e provide phylogenetic classifications for the 16S rRNA sequences in INSD (Chun et al., 2007; Kim et al., 2012). EzTaxon-e is an extension of the original EzTaxon. Different from other reference databases mentioned above, EzTaxon-e contains 16S rRNA gene data from uncultured microbes that were obtained from the previous metagenomic studies. It can have potential expansion of the availability of taxonomical classification. EzTaxon/EzTaxon-e (as of August 1, 2015) contains 64,329 species/types of the data.

3. Marine metagenome databases

Table 2 shows the major databases that contain the marine metagenome data. In spite of a large amount of the marine metagenome data that are currently produced with an enormous speed, it is surprising that there is no database exclusively devoted for marine metagenome data. In fact, only six databases contain the marine metagenome data, which are currently available for deposition of the data and their further analyses. This suggests that the marine metagenome data are deposited in only a limited number of databases. Here, we provide an overview of the databases that contain marine metagenome data (Table 2) as follow:

iMicrobe is a collection of microbe data, not only for the metagenomic data but also for other microbial-related sequencing data such as transcriptomics and genomics (http://imicrobe.us). In collaboration with iPlant (Goff et al., 2011), iMicrobe provides a web-based computational environment for supporting metagenome data analysis. As of October 2, 2015, iMicrobe contains 128 projects including 3,338 different environmental omics samples.

VIROME is a collection of viral data derived from various environmental metagenome data (Wommack et al., 2012). VIROME predicts the open-reading frames from the viral sequence data, making their proper classification. Uploading the data form the web site, users can subsequently conduct the data analysis by use of VIROME's analytical environments. VIROME currently contains the data that were derived from 466 libraries.

MetaGenomics Rapid Annotation using Subsystem Technology (MG-RAST) is one of the most popular web-based systems that provide an extensive pipeline for analysis of the metagenomic data. In practice, MG-RAST offers a combination of analytical tools to visualize the metagenomic data (Aziz et al., 2008). Once the user submits the metagenome data to MG-RAST, the data can be analyzed by their system, providing annotations, taxonomic classifications, and comparison of the user data with the metagenome data contained in this database. MG-RAST contains 211,068 metagenomes that correspond to 29,927 publically available data sets, as of October 2, 2015.

EBI Metagenomics offers an automated pipeline for collection and analysis of metagenomic data (Hunter et al., 2014). The users can submit their metagenome data to this database, so that their data are subsequently analyzed by the analytical pipeline. Once the user agrees to share the data with other people, their data will become freely available to public, being deposited to INSD. EBI metagenomics represents 138 projects that contain 6,381 metagenome samples, as of October 1, 2015.

IMG/M is an integrated system for microbial data collection and analysis (Markowitz et al., 2014a). IMG is a component for data collection in the system (Markowitz et al., 2014b). IMG/M ER is an analytical pipeline for the metagenome data (Markowitz et al., 2014a). The users can submit their own data to the system and obtain annotations of their data. The uploaded data is kept only private, not shared

with other people for a limited period of time. IMG and IMG/M pipelines is tightly linked to The Genomes OnLine Database (GOLD), which is a web-based comprehensive online resource that catalogs and monitors genomic and metagenomic projects worldwide (Reddy et al., 2015). GOLD does not provide any actual raw data of metagenome, but it is useful to grasp the progress of the metagenome projects. IMG/M contains 4,210 metagenome data as of October 1, 2015.

As shown in Table 2, these three databases, MG-RAST, EBI Metagenomics, and IMG/M, are the databases that also provide analytical pipelines of metagenome data to users with the intuitively understandable web interface. Because of the limitation of bioinformatics resources as discussed below, the pipeline service is of great help to the metagenome community.

MEtaGenome ANalyzer DataBase (MeganDB) is a comprehensive database of pre-calculated metagenomic datasets, which is particularly designed for the analysis tool, MEGAN (MEtaGenome Analyzer) (Huson et al., 2007). MEGAN is one of the most popular tools that are used to examine the taxonomic and functional contents in the metagenome data. In MeganDB, all the public metagenome data are collected and annotated by MEGAN. 235 metagenome data is stored in MeganDB as of October 2, 2015. .

4. Challenges of metagenome databases for the marine sciences

Since an amount of the metagenome data is continuously increasing, the metagenome database should well represent the projects of massive data production of metagenomes, playing a essential role of storage for future analysis. A large amount of metagenome data sometimes causes a problem of data management, since it is so hard to deal with those data in a proper way. To analyze the data, one must invoke a high-

performance computer as well as an expert of bioinformatics or special software. Thus, construction of a metagenome database is essential.

As we discussed earlier, construction of the metagenome database require proactive usage of the reference database particularly when functional annotation is conducted. During this step of the process, we expect that the reference database contains comprehensive information on taxonomic species as well as the genes/proteins. Unfortunately, this is not the case often. Therefore, well-developments of the reference database is prerequisite for constructing the metagenome database.

It is of immediate need to construct a metagenome database that contains exclusively the marine metagenome data, because there is no such a database at present. Our collaborators in Japan are now constructing the marine metagenome database under the CREST project of JST (Japan Science and Technology Agency), which will be publically available very soon.

Another challenge in the current metagenome database is a lack of unified format and nomenclature that make comparison of data among different databases extremely difficult. This problem is essentially due to the fact that the metagenome data from various projects are produced by different experimental and analytical protocols and conditions. This is a serious problem because accumulation of the metagenome data cannot be utilized effectively for the data analysis particularly when the so-called meta analysis is needed. To solve this, we need to start establishing a standard experimental protocol. In addition, formation of the world-wide consortium for marine metagenome like a genomic standards consortium (http://gensc.org/) may help coordinated arrangements of the data format and meta analyses from the viewpoint of bioinformatics developments.

5. Conclusion and perspective

In this review, we made an overview of the current databases for metagenomics that contain marine metagenome, because marine metagenome is of particular concern since 71% of the Earth is covered with the ocean and 80% of species are living there. In fact, the marine metagenome data is acutely increasing, leading to accumulation of enormous amount of the data. In the present situation, construction of the marine metagenome database is crucial for further developments of marine metagenomics by extracting biologically significant knowledge from the big data. For this reason, we summarized the outline of the currently available the database that contain marine metagenome data. In particular, we emphasized importance of the reference databases in constructing the marine metagenome database. We also described challenges for construction of the marine metagenome database; formation of the common framework for the metagenome data.

Acknowledgements

We thank Ms. Asuka Kutsuma for assisting in preparing the database collection. We are also grateful to Mr. Kosuke Goto for the discussion at the initial stage of this manuscript. This work was supported by the research fund from King Abdullah University of Science and Technology (KAUST).

References

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D.,

Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. and Zagnitko, O., 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9, 75.

Behzad, H., Gojobori, T. and Mineta, K., 2015. Challenges and opportunities of airborne metagenomics. Genome Biol Evol 7, 1216-26.

Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E. and Wincker, P., 2015. Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. Science 348, 873.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A., 2007. UniProtKB/Swiss-Prot. Methods Mol Biol 406, 89-112.

Chun, J., Lee, J.H., Jung, Y., Kim, M., Kim, S., Kim, B.K. and Lim, Y.W., 2007. EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. Int J Syst Evol Microbiol 57, 2259-61.

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. Nucleic Acids Res 42, D633-42.

Cowan, D.A., Ramond, J.B., Makhalanyane, T.P. and De Maayer, P., 2015. Metagenomics of extreme environments. Curr Opin Microbiol 25, 97-102.

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, 5069-72.

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W.H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang, C., Lent, M., Kim,

S.J., Kvilekval, K., Manjunath, B.S., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S.M., Cranston, K.A., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D.J., White, J.W., Leebens-Mack, J., Donoghue, M.J., Spalding, E.P., Vision, T.J., Myers, C.R., Lowenthal, D., Enquist, B.J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L. and Stanzione, D., 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Front Plant Sci 2, 34.

Gross, L., 2007. Untapped bounty: sampling the seas to survey microbial biodiversity. *PLoS Biol.* **5**, e85.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J. and Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5, R245-9.

Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., Maslen, J., Mitchell, A., Nuka, G., Oisel, A., Pesseat, S., Radhakrishnan, R., Rocca-Serra, P., Scheremetjew, M., Sterk, P., Vaughan, D., Cochrane, G., Field, D. and Sansone, S.A., 2014. EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res 42, D600-6.

Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Res 17, 377-86.

Kennedy, J., Marchesi, J.R. and Dobson, A.D., 2008. Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. Microb Cell Fact 7, 27.

Kim, M., Lee, K.H., Yoon, S.W., Kim, B.S., Chun, J. and Yi, H., 2013. Analytical tools and databases for metagenomics in the next-generation sequencing era. Genomics Inform 11, 102-13.

Kim, O.S., Cho, Y.J., Lee, K., Yoon, S.H., Kim, M., Na, H., Park, S.C., Jeon, Y.S., Lee, J.H.,

Yi, H., Won, S. and Chun, J., 2012. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. Int J Syst Evol Microbiol 62, 716-21.

Kosuge, T., Mashima, J., Kodama, Y., Fujisawa, T., Kaminuma, E., Ogasawara, O., Okubo, K., Takagi, T. and Nakamura, Y., 2014. DDBJ progress report: a new submission system for leading to a correct annotation. Nucleic Acids Res 42, D44-9.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. and Hugenholtz, P., 2008. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72, 557-78, Table of Contents.

Laursen, L. 2011 Spain's ship comes. *Nature* 475, 16–17.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A. and Schleifer, K.H., 2004. ARB: a software environment for sequence data. Nucleic Acids Res 32, 1363-71.

Markowitz, V.M., Chen, I.M., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Pagani, I., Tringe, S., Huntemann, M., Billis, K., Varghese, N., Tennessen, K., Mavromatis, K., Pati, A., Ivanova, N.N. and Kyrpides, N.C., 2014a. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res 42, D568-73.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., Huntemann, M., Anderson, I., Billis, K., Varghese, N., Mavromatis, K., Pati, A., Ivanova, N.N. and Kyrpides, N.C., 2014b. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 42, D560-7.

Nakamura, Y., Cochrane, G., Karsch-Mizrachi, I. and International Nucleotide Sequence

Database, C., 2013. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 41, D21-4.

Pace, N.R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734-40.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glockner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41, D590-6.

Reddy, T.B., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A. and Kyrpides, N.C., 2015. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res 43, D1099-106.

Rothschild, L.J. and Mancinelli, R.L., 2001. Life in extreme environments. Nature 409, 1092-101.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.H., Falcon, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M. and Venter, J.C., 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5, e77.

Sharpton, T.J., 2014. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci 5, 209.

Sunagawa, S., Karsenti, E., Bowler, C. and Bork, P., 2015. Computational eco-systems biology in Tara Oceans: translating data into knowledge. Mol Syst Biol 11, 809.

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., Tolstoy, I. and

Zaslavsky, L., 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Res 43, D599-605.

Teeling, H. and Glockner, F.O., 2012. Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. Brief Bioinform 13, 728-42.

Thomas, T., Gilbert, J. and Meyer, F., 2012. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2, 3.

UniProt, C., 2015. UniProt: a hub for protein information. Nucleic Acids Res 43, D204-12.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. and Smith, H.O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

Whitman, W.B., Coleman, D.C. and Wiebe, W.J., 1998. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 95, 6578-83.

Woese, C.R. and Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74, 5088-90.

Wommack, K.E., Bhavsar, J., Polson, S.W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S. and Nasko, D.J., 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci 6, 427-39.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C.S., Li, H., Mashiyama, S.T., Joachimiak, M.P., van Belle, C., Chandonia, J.M., Soergel, D.A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B.J., Bafna, V., Friedman, R., Brenner, S.E., Godzik, A., Eisenberg, D., Dixon, J.E., Taylor, S.S., Strausberg, R.L., Frazier, M. and Venter, J.C., 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5, e16.

Abbreviations:
NGS , next-generation sequencing; DDBJ, DNA Data Bank of Japan; INSD, International Nucleotide Sequence Database; RDP, Ribosomal Database Project; MG-RAST, MetaGenomics Rapid Annotation using Subsystem Technology; GOLD, Genomes OnLine Database; MeganDB, MEtaGenome ANalyzer DataBase; MEGAN, MEtaGenome Analyzer

Table 1: List of reference databases for metagenomics data analysis by 16S ribosomal RNA genes

Database name	URL
RDP (Ribosomal Database Project)	http://rdp.cme.msu.edu/
Greengenes	http://greengenes.lbl.gov/
SILVA	http://www.arb-silva.de/
EzTaxon/EzTaxon-e	http://www.ezbiocloud.net/eztaxon

Table 2: List of the marine metagenome databases

Database name	Main Target data	Supporting Function	URL
iMicrobe	microbe	Tools	http://imicrobe.us/
VIROME	virus/microbe	Tools	http://virome.dbi.udel.edu/
EBI metagenomics	all	Pipeline	https://www.ebi.ac.uk/metagenomics/
IMG/M (Integrated Microbial Genomics and Metagenomics)	microbe	Pipeline	http://img.jgi.doe.gov/
MG-RAST (Metagenomic Rapid Annotations using Subsystems Technology)	all	Pipeline	http://metagenomics.anl.gov/
MeganDB	all	Tools	http://www.megan-db.org/