nature biotechnology

PERSPECTIVE

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Pelin Yilmaz[1,2*], Renzo Kottmann[1], Dawn Field[3], Rob Knight[4,5], James R Cole[6,7], Linda Amaral-Zettler[8], Jack A Gilbert[9–11], Ilene Karsch-Mizrachi[12], Anjanette Johnston[12], Guy Cochrane[13], Robert Vaughan[13], Christopher Hunter[13], Joonhong Park[14], Norman Morrison[3,15], Philippe Rocca-Serra[16], Peter Sterk[3], Manimozhiyan Arumugam[17], Mark Bailey[3], Laura Baumgartner[18], Bruce W Birren[19], Martin J Blaser[20], Vivien Bonazzi[21], Tim Booth[3], Peer Bork[17], Frederic D Bushman[22], Pier Luigi Buttigieg[1,2], Patrick S G Chain[7,23,24], Emily Charlson[22], Elizabeth K Costello[4], Heather Huot-Creasy[25], Peter Dawyndt[26], Todd DeSantis[27], Noah Fierer[28], Jed A Fuhrman[29], Rachel E Gallery[30], Dirk Gevers[19], Richard A Gibbs[31,32], Inigo San Gil[33], Antonio Gonzalez[34], Jeffrey I Gordon[35], Robert Guralnick[28,36], Wolfgang Hankeln[1,2], Sarah Highlander[31,37], Philip Hugenholtz[38], Janet Jansson[23,39], Andrew L Kau[35], Scott T Kelley[40], Jerry Kennedy[4], Dan Knights[34], Omry Koren[41], Justin Kuczynski[18], Nikos Kyrpides[23], Robert Larsen[4], Christian L Lauber[42], Teresa Legg[28], Ruth E Ley[41], Catherine A Lozupone[4], Wolfgang Ludwig[43], Donna Lyons[42], Eamonn Maguire[16], Barbara A Methé[44], Folker Meyer[10], Brian Muegge[35], Sara Nakielny[4], Karen E Nelson[44], Diana Nemergut[45], Josh D Neufeld[46], Lindsay K Newbold[3], Anna E Oliver[3], Norman R Pace[18], Giriprakash Palanisamy[47], Jörg Peplies[48], Joseph Petrosino[31,37], Lita Proctor[21], Elmar Pruesse[1,2], Christian Quast[1], Jeroen Raes[49], Sujeevan Ratnasingham[50], Jacques Ravel[25], David A Relman[51,52], Susanna Assunta-Sansone[16], Patrick D Schloss[53], Lynn Schriml[25], Rohini Sinha[22], Michelle I Smith[35], Erica Sodergren[54], Aymé Spor[41], Jesse Stombaugh[4], James M Tiedje[7], Doyle V Ward[19], George M Weinstock[54], Doug Wendel[4], Owen White[25], Andrew Whiteley[3], Andreas Wilke[10], Jennifer R Wortman[25], Tanya Yatsunenko[35] & Frank Oliver Glöckner[1,2]

Here we present a standard developed by the Genomic Standards Consortium (GSC) for reporting marker gene sequences—the minimum information about a marker gene sequence (MIMARKS). We also introduce a system for describing the environment from which a biological sample originates. The 'environmental packages' apply to any genome sequence of known origin and can be used in combination with MIMARKS and other GSC checklists. Finally, to establish a unified standard for describing sequence data and to provide a single point of entry for the scientific community to access and learn about GSC checklists, we present the minimum information about any (x) sequence (MIxS). Adoption of MIxS will enhance our ability to analyze natural genetic diversity documented by massive DNA sequencing efforts from myriad ecosystems in our ever-changing biosphere.

Without specific guidelines, most genomic, metagenomic and marker gene sequences in databases are sparsely annotated with the information required to guide data integration, comparative studies and knowledge generation. Even with complex keyword searches, it is currently impossible to reliably retrieve sequences that have originated from certain environments or particular locations on Earth—for example, all sequences from 'soil' or 'freshwater lakes' in a certain region of the world. Because public databases of the International Nucleotide Sequence Database Collaboration (INSDC; comprising DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (EBI-ENA) and GenBank (http://www.insdc.org/)) depend on author-submitted information to enrich the value of sequence data sets, we argue that the only way to change the current practice is to establish a standard of reporting that requires contextual data to be deposited at the time of sequence submission. The adoption of such a standard would elevate the quality, accessibility and utility of information that can be collected from INSDC or any other data repository.

The GSC has previously proposed standards for describing genomic sequences— the "minimum information about a genome sequence" (MIGS)—and metagenomic sequences—the "minimum information about a metagenome sequence" (MIMS)[1]. Here we introduce an extension of these standards for capturing information about marker genes. Additionally, we introduce 'environmental packages' that standardize sets of measurements and observations describing particular habitats that are applicable across all GSC checklists and beyond[2]. We define 'environment' as any location in which a sample or organism

*A list of affiliations appears at the end of the paper.

Published online 6 May 2011; doi:10.1038/nbt.1823

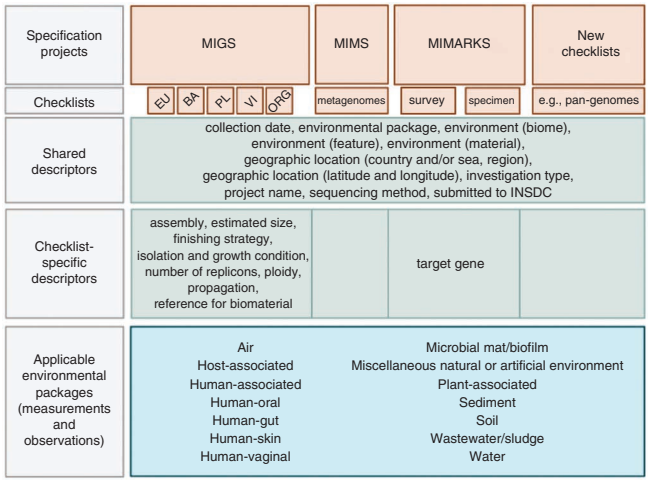

Figure 1 Schematic overview about the GSC MIxS standard (brown), including combination with specific environmental packages (blue). Shared descriptors apply to all MIxS checklists; however, each checklist has its own specific descriptors as well. Environmental packages can be applied to any of the checklists. EU, eukarya; BA, bacteria/archaea; PL, plasmid; VI, virus; ORG, organelle.

is found, e.g., soil, air, water, human-associated, plant-associated or laboratory. The original MIGS/MIMS checklists included contextual data about the location from which a sample was isolated and how the sequence data were produced. However, standard descriptions for a more comprehensive range of environmental parameters, which would help to better contextualize a sample, were not included. The environmental packages presented here are relevant to any genome sequence of known origin and are designed to be used in combination with MIGS, MIMS and MIMARKS checklists.

To create a single entry point to all minimum information checklists from the GSC and to the environmental packages, we propose an overarching framework, the MIxS standard (http://gensc.org/gc_wiki/index.php/MIxS). MIxS includes the technology-specific checklists from the previous MIGS and MIMS standards, provides a way of introducing additional checklists such as MIMARKS, and also allows annotation of sample data using environmental packages. A schematic overview of MIxS along with the MIxS environmental packages is shown in **Figure 1**.

Development of MIMARKS and the environmental packages

Over the past three decades, the 16S rRNA, 18S rRNA and internal transcribed spacer gene sequences (ITS) from *Bacteria*, *Archaea* and microbial *Eukaryotes* have provided deep insights into the topology of the tree of life[3,4] and the composition of communities of organisms that live in diverse environments, ranging from deep sea hydrothermal vents to ice sheets in the Arctic[5–16]. Numerous other phylogenetic marker genes have proven useful, including RNA polymerase subunits (*rpoB*), DNA gyrases (*gyrB*), DNA recombination and repair proteins (*recA*) and heat shock proteins (*HSP70*)[3]. Marker genes can also reveal key metabolic functions rather than phylogeny; examples include nitrogen cycling (*amoA, nifH, ntcA*)[17,18], sulfate reduction (*dsrAB*)[19] or phosphorus metabolism (*phnA, phnI, phnJ*)[20,21]. In this paper we define all phylogenetic and functional genes (or gene fragments) used to profile natural genetic diversity as 'marker genes'. MIMARKS (**Table 1**) complements the MIGS/MIMS checklists for genomes and metagenomes by adding two new checklists, a MIMARKS survey, for uncultured diversity marker gene surveys, and a MIMARKS specimen, for marker gene sequences obtained from any material identifiable by means of specimens. The MIMARKS extension adopts and incorporates the standards being developed by the Consortium for the Barcode of Life (CBOL)[22]. Therefore, the checklist can be universally applied to any marker gene, from small subunit rRNA to cytochrome oxidase I (COI), to all taxa, and to studies ranging from single individuals to complex communities.

Both MIMARKS and the environmental packages were developed by collating information from several sources and evaluating it in the framework of the existing MIGS/MIMS checklists. These include four independent community-led surveys, examination of the parameters reported in published studies and examination of compliance with optional features in INSDC documents. The overall goal of these activities was to design the backbone of the MIMARKS checklist, which describes the most important aspects of marker gene contextual data.

Results of community-led surveys

Four online surveys about descriptors for marker genes have been conducted to determine researcher preferences for core descriptors. The Department of Energy Joint Genome Institute and SILVA[23] surveys focused on general descriptor contextual data for a marker gene, whereas the Ribosomal Database Project (RDP)[24] focused on prevalent habitats for rRNA gene surveys, and the Terragenome Consortium[25] focused on soil metagenome project contextual data (**Supplementary Results 1**). The above recommendations were combined with an extensive set of contextual data items suggested by an International Census of Marine Microbes (ICoMM) working group that met in 2005. These collective resources provided valuable insights into community requests for contextual data items to be included in the MIMARKS checklist and the main habitats constituting the environmental packages.

Survey of published parameters

We reviewed published rRNA gene studies, retrieved from SILVA and the ICoMM database MICROBIS (The Microbial Oceanic Biogeographic Information System, http://icomm.mbl.edu/microbis/) to further supplement contextual data items that are included in the respective environmental packages. In total, 39 publications from SILVA and >40 ICoMM projects were scanned for contextual data items to constitute the core of the environmental package subtables (**Supplementary Results 1**).

In a final analysis step, we surveyed usage statistics of INSDC source feature key qualifier values of rRNA gene sequences contained in SILVA (**Supplementary Results 1**). Notably, <10% of the 1.2 million 16S rRNA gene sequences (SILVA release 100) were associated with even basic information such as latitude and longitude, collection date or PCR primers.

The MIMARKS checklist

The MIMARKS checklist provides users with an 'electronic laboratory notebook' containing core contextual data items required for consistent reporting of marker gene investigations. MIMARKS uses the MIGS/MIMS checklists with respect to the nucleic acid sequence source and sequencing contextual data, but extends them with further experimental contextual data such as PCR primers and conditions, or target gene name.

For clarity and ease of use, all items within the MIMARKS checklist are presented with a value syntax description, as well as a clear definition of the item. Whenever terms from a specific ontology are required as the value of an item, these terms can be readily found in the respective ontology browsers linked by URLs in the item definition. Although this version of the MIMARKS checklist does not

Table 1 The core items of the MIMARKS checklists, along with the value types, descriptions and requirement status

Item	Description	Report type: MIMARKS survey	Report type: MIMARKS specimen
Investigation			
Submitted to INSDC [boolean]	Depending on the study (large-scale, e.g., done with next-generation sequencing technology, or small-scale) sequences have to be submitted to SRA (Sequence Read Archives), DRA (DDBJ Sequence Read Archive) or through the classical Webin/Sequin systems to GenBank, ENA and DDBJ	M	M
Investigation type [mimarks-survey or mimarks-specimen]	Nucleic Acid Sequence Report is the root element of all MIMARKS compliant reports as standardized by Genomic Standards Consortium (GSC). This field is either MIMARKS survey or MIMARKS specimen	M	M
Project name	Name of the project within which the sequencing was organized	M	M
Environment			
Geographic location (latitude and longitude [float, point, transect and region])	The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system	M	M
Geographic location (depth [integer, point, interval, unit])	Please refer to the definitions of depth in the environmental packages	E	E
Geographic location (elevation of site [integer, unit]; altitude of sample [integer, unit])	Please refer to the definitions of either altitude or elevation in the environmental packages	E	E
Geographic location (country and/or sea [INSDC or GAZ]; region [GAZ])	The geographical origin of the sample as defined by the country or sea name. Country, sea or region names should be chosen from the INSDC list (http://insdc.org/country.html), or the GAZ (Gazetteer, v1.446) ontology (http://bioportal.bioontology.org/visualize/40651)	M	M
Collection date [ISO8601]	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated, that is, all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; except for 2008-01 and 2008, all are ISO6801 compliant	M	M
Environment (biome [EnvO])	In environmental biome level are the major classes of ecologically similar communities of plants, animals and other organisms. Biomes are defined based on factors such as plant structures, leaf types, plant spacing and other factors like climate. Examples include desert, taiga, deciduous woodland or coral reef. Environment Ontology (EnvO) (v1.53) terms listed under environmental biome can be found at http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00000428	M	M
Environment (feature [EnvO])	Environmental feature level includes geographic environmental features. Examples include harbor, cliff or lake. EnvO (v1.53) terms listed under environmental feature can be found at http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00002297	M	M
Environment (material [EnvO])	The environmental material level refers to the matter that was displaced by the sample, before the sampling event. Environmental matter terms are generally mass nouns. Examples include: air, soil or water. EnvO (v1.53) terms listed under environmental matter can be found at http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483	M	M
MIGS/MIMS/MIMARKS extension			
Environmental package [air, host-associated, human-associated, human-skin, human-oral, human-gut, human-vaginal, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated, sediment, soil, wastewater/sludge, water]	MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained. All environmental packages listed here are further defined in separate subtables. By giving the name of the environmental package, a selection of fields can be made from the subtables and can be reported	M	M
Nucleic acid sequence source			
Isolation and growth conditions [PMID, DOI or URL]	Publication reference in the form of PubMed ID (PMID), digital object identifier (DOI) or URL for isolation and growth condition specifications of the organism/material	–	M
Sequencing			
Target gene or locus (e.g., 16S rRNA, 18S rRNA, nif, amoA, rpo)	Targeted gene or locus name for marker gene study	M	M
Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony)	Sequencing method used, e.g., Sanger, pyrosequencing, ABI-solid	M	M

Items for the MIMARKS specification and their mandatory (M), status for both MIMARKS-survey and MIMARKS-specimen checklists. Furthermore, "–" denotes that an item is not applicable for a given checklist. E denotes that a field has environment-specific requirements. For example, whereas "depth" is mandatory for the environments water, sediment or soil, it is optional for human-associated environments. MIMARKS-survey is applicable to contextual data for marker gene sequences, obtained directly from the environment, without culturing or identification of the organisms. MIMARKS-specimen, on the other hand, applies to the contextual data for marker gene sequences from cultured or voucher-identifiable specimens. Both MIMARKS-survey and specimen checklists can be used for any type of marker gene sequence data, ranging from 16S, 18S, 23S, 28S rRNA to COI, hence the checklists are universal for all three domains of life. Item names are followed by a short description of the value of the item in parentheses and/or value type in brackets as a superscript. Whenever applicable, value types are chosen from a controlled vocabulary (CV) or an ontology from the Open Biological and Biomedical Ontologies (OBO) foundry (http://www.obofoundry.org/). This table only presents the very core of MIMARKS checklists, that is, only mandatory items for each checklist. **Supplementary Results 2** contains all MIMARKS items, the tables for environmental packages in the MIGS/MIMS/MIMARKS extension and GenBank structured comment name that should be used for submitting MIMARKS data to GenBank. In case of submitting to EBI-ENA, the full names can be used.

contain unit specifications, we recommend all units to be chosen from and follow the International System of Units (SI) recommendations. In addition, we strongly urge the community to provide feedback regarding the best unit recommendations for given parameters. Unit standardization across data sets will be vital to facilitate comparative studies in future. An Excel version of the MIMARKS checklist is provided on the GSC web site (http://gensc.org/gc_wiki/index.php/MIMARKS).

The MIxS environmental packages

Fourteen environmental packages provide a wealth of environmental and epidemiological contextual data fields for a complete description of sampling environments. The environmental packages can be combined with any of the GSC checklists (**Fig. 1** and **Supplementary Results 2**). Researchers within The Human Microbiome Project[26] contributed the host-associated and all human packages. The Terragenome Consortium contributed sediment and soil packages. Finally, ICoMM, Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites and the Max Planck Institute for Marine Microbiology contributed the water package. The MIMARKS working group developed the remaining packages (air, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated and wastewater/sludge). The package names describe high-level habitat terms in order to be exhaustive. The miscellaneous natural or artificial environment package contains a generic set of parameters, and is included for any other habitat that does not fall into the other thirteen categories. Whenever needed, multiple packages may be used for the description of the environment.

Examples of MIMARKS-compliant data sets

Several MIMARKS-compliant reports are included in **Supplementary Results 3**. These include a 16S rRNA gene survey from samples obtained in the North Atlantic, an 18S pyrosequencing tag study of anaerobic protists in a permanently anoxic basin of the North Sea, a *pmoA* survey from Negev Desert soils, a *dsrAB* survey of Gulf of Mexico sediments and a 16S pyrosequencing tag study of bacterial diversity in the western English Channel (SRA accession no. SRP001108).

Adoption by major database and informatics resources

Support for adoption of MIMARKS and the MIxS standard has spread rapidly. Authors of this paper include representatives from genome sequencing centers, maintainers of major resources, principal investigators of large- and small-scale sequencing projects, and individual investigators who have provided compliant data sets, showing the breadth of support for the standard within the community.

In the past, the INSDC has issued a reserved 'barcode' keyword for the CBOL[7]. Following this model, the INSDC has recently recognized the GSC as an authority for the MIxS standard and issued the standard with official keywords within INSDC nucleotide sequence records[27]. This greatly facilitates automatic validation of the submitted contextual data and provides support for data sets compliant with previous versions by including the checklist version as a keyword.

GenBank accepts MIxS metadata in tabular format using the sequin and tbl2asn submission tools, validates MIxS compliance and reports the fields in the structured comment block. The EBI-ENA Webin submission system provides prepared web forms for the submission of MIxS compliant data; it presents all of the appropriate fields with descriptions, explanations and examples, and validates the data entered. One tool that can aid submitting contextual data is MetaBar[28], a spreadsheet and web-based software, designed to assist users in the consistent acquisition, electronic storage and submission of contextual data associated with their samples in compliance with the MIxS standard. The online tool CDinFusion (http://www.megx.net/CDinFusion) was created to facilitate the combination of contextual data with sequence data, and generation of submission-ready files.

The next-generation Sequence Read Archive (SRA) collects and displays MIxS-compliant metadata in sample and experiment objects. There are several tools that are already available or under development to assist users in SRA submissions. The myRDP SRA PrepKit allows users to prepare and edit their submissions of reads generated from ultra-high-throughput sequencing technologies. A set of suggested attributes in the data forms assist researchers in providing metadata conforming to checklists such as MIMARKS. The Quantitative Insights Into Microbial Ecology (QIIME) web application (http://www.microbio.me/qiime) allows users to generate and validate MIMARKS-compliant templates. These templates can be viewed and completed in the users' spreadsheet editor of choice (e.g., Microsoft Excel). The QIIME web-platform also offers an ontology lookup and geo-referencing tool to aid users when completing the MIMARKS templates. The Investigation/Study/Assay (ISA) is a software suite that assists in the curation, reporting and local management of experimental metadata from studies using one or a combination of technologies, including high-throughput sequencing[29]. Specific ISA configurations (http://isa-tools.org/tools.html) have been developed to ensure MIxS compliance by providing templates and validation capability. Another tool, ISAconverter, produces SRA.xml documents, facilitating submission to the SRA repository. MIxS checklists are also registered with the BioSharing catalog of standards (http://biosharing.org/), set to progressively link minimal information specifications to the respective exchange formats, ontologies and compliant tools.

Further detailed guidance for submission processes can be found under the respective wiki pages (http://gensc.org/gc_wiki/index.php/MIxS) of the standard.

Maintenance of the MIxS standard

To allow further developments, extensions and enhancements of MIxS, we set up a public issue tracking system to track changes and accomplish feature requests (http://mixs.gensc.org/). New versions will be released annually. Technically, the MIxS standard, including MIMARKS and the environmental packages, is maintained in a relational database system at the Max Planck Institute for Marine Microbiology Bremen on behalf of the GSC. This provides a secure and stable mechanism for updating the checklist suite and versioning. In the future, we plan to develop programmatic access to this database to allow automatic retrieval of the latest version of each checklist for INSDC databases and for GSC community resources. Moreover, the Genomic Contextual Data Markup Language is a reference implementation of the GSC checklists by the GSC and now implements the full range of MIxS standards. It is based on XML Schema technology and thus serves as an interoperable data exchange format for infrastructures based on web services[30].

Conclusions and call for action

The GSC is an international body with a stated mission of working towards richer descriptions of the complete collection of genomes and metagenomes through the MIxS standard. The present report extends the scope of GSC guidelines to marker gene sequences and environmental packages and establishes a single portal where experimentalists

can gain access to and learn how to use GSC guidelines. The GSC is an open initiative that welcomes the participation of the wider community. This includes an open call to contribute to refinements of the MIxS standards and their implementations.

The adoption of the GSC standards by major data providers and organizations, as well as the INSDC, supports efforts to contextually enrich sequence data and complements recent efforts to enrich other (meta) 'omics data. The MIxS standard, including MIMARKS, has been developed to the point that it is ready for use in the publication of sequences. A defined procedure for requesting new features and stable release cycles will facilitate implementation of the standard across the community. Compliance among authors, adoption by journals and use by informatics resources will vastly improve our collective ability to mine and integrate invaluable sequence data collections for knowledge- and application-driven research. In particular, the ability to combine microbial community samples collected from any source, using the universal tree of life as a measure to compare even the most diverse communities, should provide new insights into the dynamic spatiotemporal distribution of microbial life on our planet and on the human body.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

Funding sources are listed in the **Supplementary Note**.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* **26**, 541–547 (2008).
2. Taylor, C.F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896 (2008).
3. Ludwig, W. & Schleifer, K.H. in *Microbial Phylogeny and Evolution, Concepts and Controversies.* (ed. Sapp, J.) 70–98 (Oxford University Press, New York, 2005).
4. Ludwig, W. *et al.* Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554–568 (1998).
5. Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).
6. Stahl, D.A. Analysis of hydrothermal vent associated symbionts by ribosomal RNA sequences. *Science* **224**, 409–411 (1984).
7. Ward, D.M., Weller, R. & Bateson, M.M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**, 63–65 (1990).
8. DeLong, E.F. Archaea in coastal marine environments. *Proc. Nat. Acad. Sci. USA* **89**, 5685–5689 (1992).
9. Diez, B., Pedros-Alio, C. & Massana, R. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941 (2001).
10. Fuhrman, J.A., McCallum, K. & Davis, A.A. Novel major archaebacterial group from marine plankton. *Nature* **356**, 148–149 (1992).
11. Hewson, I. & Fuhrman, J.A. Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. *Appl. Environ. Microbiol.* **70**, 3425–3433 (2004).
12. Huber, J.A., Butterfield, D.A. & Baross, J.A. Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. *Appl. Environ. Microbiol.* **68**, 1585–1594 (2002).
13. Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C. & Moreira, D. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607 (2001).
14. Moon-van der Staay, S.Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001).
15. Pace, N.R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
16. Rappe, M.S. & Giovannoni, S.J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
17. Francis, C.A., Beman, J.M. & Kuypers, M.M.M. New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J.* **1**, 19–27 (2007).
18. Zehr, J.P., Mellon, M.T. & Zani, S. New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (*nifH*) genes. *Appl. Environ. Microbiol.* **64**, 3444–3450 (1998).
19. Minz, D. *et al.* Diversity of sulfate-reducing bacteria in oxic and anoxic regions of a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase genes. *Appl. Environ. Microbiol.* **65**, 4666–4671 (1999).
20. Gilbert, J.A. *et al.* The seasonal structure of microbial communities in the Western English Channel. *Environ. Microbiol.* **11**, 3132–3139 (2009).
21. Martinez, A.W., Tyson, G. & DeLong, E.F. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ. Microbiol.* **12**, 222–238 (2009).
22. Hanner, R. Data Standards for BARCODE Records in INSDC (BRIs) (Database Working Group, Consortium for the Barcode of Life, 2009). <http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG_data_standards-Final.pdf>.
23. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
24. Cole, J.R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141–D145 (2009).
25. Vogel, T.M. *et al.* TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol.* **7**, 252 (2009).
26. Turnbaugh, P.J. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
27. Benson, D.A. *et al.* GenBank. *Nucleic Acids Res.* **36**, D25–D30 (2008).
28. Hankeln, W. *et al.* MetaBar—a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinformatics* **11**, 358 (2010).
29. Rocca-Serra, P. *et al.* ISA infrastructure: supporting standards-compliant experimental reporting and enabling curation at the community level. *Bioinformatics* **26**, 2354–2356 (2010).
30. Kottmann, R. *et al.* A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* **12**, 115–121 (2008).

[1]Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany. [2]Jacobs University Bremen gGmbH, Bremen, Germany. [3]Natural Environment Research Council Environmental Bioinformatics Centre, Wallington CEH, Oxford, UK. [4]Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA. [5]Howard Hughes Medical Institute, San Francisco, California, USA. [6]Ribosomal Database Project, Michigan State University, East Lansing, Michigan, USA. [7]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan, USA. [8]The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA. [9]Plymouth Marine Laboratory, Plymouth, UK. [10]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA. [11]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA. [12]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA. [13]European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [14]WCU Center for Green Metagenomics, School of Civil and Environmental Engineering, Yonsei University, Seoul, Republic of Korea. [15]School of Computer Science, University of Manchester, Manchester, UK. [16]Oxford e-Research Centre, University of Oxford, Oxford, UK. [17]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [18]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, USA. [19]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. [20]Department of Medicine and the Department of Microbiology, New York University Langone Medical Center, New York, New York, USA. [21]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. [22]Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. [23]DOE Joint Genome Institute, Walnut Creek, California, USA. [24]Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA. [25]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA. [26]Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium. [27]Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [28]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, USA. [29]Department of Biological Sciences, University of Southern California, Los Angeles, California, USA. [30]National Ecological Observatory Network, Boulder, Colorado, USA. [31]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [32]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. [33]Department of Biology, University of New Mexico, LTER Network Office, Albuquerque, New Mexico, USA. [34]Department of Computer Science, University of Colorado, Boulder, Colorado, USA. [35]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA. [36]University of Colorado Museum of Natural History, University of Colorado,

Boulder, Colorado, USA. [37]Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, USA. [38]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia. [39]Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [40]Department of Biology, San Diego State University, San Diego, California, USA. [41]Department of Microbiology, Cornell University, Ithaca, New York, USA. [42]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA. [43]Lehrstuhl für Mikrobiologie, Technische Universität München, Freising, Germany. [44]J. Craig Venter Institute, Rockville, Maryland, USA. [45]Department of Environmental Sciences, University of Colorado, Boulder, Colorado, USA. [46]Department of Biology, University of Waterloo, Ontario, Canada. [47]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. [48]Ribocon GmbH, Bremen, Germany. [49]VIB - Vrije Universiteit Brussel, Brussels, Belgium. [50]Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada. [51]Departments of Microbiology and Immunology and Department of Medicine, Stanford University School of Medicine, Stanford, California, USA. [52]Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA. [53]Department of Microbiology and Immunology, Ann Arbor, Michigan, USA. [54]The Genome Center, Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA. Correspondence should be addressed to F.O.G. (fog@mpi-bremen.de).