The "Minimum Information about an ENvironmental Sequence" (MIENS) specification

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[9,10,11], Ilene Karsch-Mizrachi[12], Anjanette Johnston[12], Guy Cochrane[13], Robert Vaughan[13], Christopher Hunter[13], Joonhong Park[14], Norman Morrison[15,16], Phillipe Rocca-Serra[13,17], Peter Sterk[3], Mani Arumugam[18], Laura Baumgartner[19], Bruce W. Birren[20], Martin J. Blaser[21], Vivien Bonazzi[22], Peer Bork[18], Pier Luigi Buttigieg[1,2], Patrick Chain[23,24], Elizabeth K. Costello[4], Heather Huot-Creasy[25], Peter Dawyndt[26], Todd DeSantis[27], Noah Fierer[28], Jed Fuhrman[29], Rachel E. Gallery[30], Richard A. Gibbs[31,32], Michelle Gwinn Giglio[25], Inigo San Gil[33], Elizabeth M. Glass[10], Antonio Gonzalez[34], Jeffrey I. Gordon[20], Robert Guralnick[28], Wolfgang Hankeln[1,2], Sarah Highlander[31,35], Philip Hugenholtz[23], Janet Jansson[36], Jerry Kennedy[4], Dan Knights[34], Omry Koren[35], Justin Kuczynski[19], Nikos Kyrpides[23], Robert Larsen[4], Christian L. Lauber[37], Teresa Legg[28], Ruth E. Ley[35], Catherine A. Lozupone[4], Wolfgang Ludwig[38], Donna Lyons[37], Eamonn Maguire[13,17], Barbara A. Methé[39], Folker Meyer[10], Sara Nakielny[4], Karen E. Nelson[39], Diana Nemergut[40], Josh D. Neufeld[41], Norman R. Pace[19], Giriprakash Palanisamy[42], Jörg Peplies[43], Jane Peterson[22], Joseph Petrosino[31,35], Lita Proctor[44], Jeroen Raes[45], Sujeevan Ratnasingham[46], Jacques Ravel[25], David A. Relman[47,48], Susanna Assunta-Sansone[13,17], Lynn Schriml[25], Erica Sodergren[49], Aymé Spor[35], Jesse Stombaugh[4], James M. Tiedje[7], Doyle V. Ward[20], George M. Weinstock[49], Doug Wendel[4], Owen White[25], Andreas Wilke[10], Jennifer Wortmann[25], Frank Oliver Glöckner[1,2]

[1] Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

[2] Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

[3] NERC Centre for Ecology and Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB, UK

[4] Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309, USA

[5] Howard Hughes Medical Institute, USA

[6] Ribosomal Database Project, Michigan State University, 2225A Biomedical and Physical Sciences Building, East Lansing, Michigan 48824-4320, USA

[7] Center for Microbial Ecology, Michigan State University, 540 Plant and Soil Sciences Building, East Lansing, Michigan 48824-1325, USA

[8] The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA

[9] Plymouth Marine Laboratory, Prospect Place, Plymouth, UK

[10] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

[11] Dept of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

[12] National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

[13] European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

[14] School of Civil and Environmental Engineering, Yonsei University, Seoul, 120-749,

Republic of Korea

[15] NERC Environmental Bioinformatics Centre, Oxford Centre for Ecology and Hydrology, Oxford OX1 3SR, UK

[16] Department of Computer Science, University of Manchester, Oxford Rd., Manchester, UK

[17] Oxford e-Research Centre, University of Oxford, Oxford OX1 3QG, UK

[18] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstr. 1, D–69117 Heidelberg, Germany

[19] Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA

[20] Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142

[21] Department of Medicine and the Department of Microbiology, New York University Langone Medical Center, New York, New York 10017, USA

[22] National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[23] DOE Joint Genome Institute, Walnut Creek, CA 94598, 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, MD 21201, USA

[26] Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium

[27] Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[28] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309, USA

[29] Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

[30] National Ecological Observatory Network (NEON), Boulder, CO 80301, USA

[31] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

[32] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

[33] Department of Biology, University of New Mexico, LTER Network Office, MSC03 2020, Albuquerque, NM 87131, USA

[34] Department of Computer Science, University of Colorado, Boulder, CO 80309, USA

[35] Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108, USA

[36] Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX

[37] Lawrence Berkeley National Laboratory, Earth Science Division, Berkeley, CA, USA

[38] Department of Microbiology, Cornell University, Ithaca NY 14853, USA

[39] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, USA

[40] Lehrstuhl für Mikrobiologie, Technische Universität München, D-853530 Freising, Germany

[41] J. Craig Venter Institute, Rockville, Maryland, United States of America

[42] Department of Environmental Sciences, University of Colorado, Boulder, CO 80309, USA

[43] Department of Biology, University of Waterloo, Ontario, N2L 3G1, Canada

[44] Environmental Sciences Division, Oak Ridge National Laboratory, Mail Stop 6407 Oak Ridge, TN, USA

[45] Ribocon GmbH, D-28359 Bremen, Germany

[46] The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA

[47] VIB - Vrije Universiteit Brussel, 1050 Brussels, Belgium

[48] Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, University of Guelph, 50 Stone Road, Guelph, ON, Canada N1G 2W1

[49] Departments of Microbiology and Immunology and of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

[50] Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304, USA

[51] The Genome Center, Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA

Summary

We present the Genomic Standards Consortium's (GSC) "Minimum Information about an ENvironmental Sequence" (MIENS) standard for describing marker genes. Adoption of MIENS will enhance our ability to analyze natural genetic diversity across the Tree of Life as it is currently being documented by massive DNA sequencing efforts from myriad ecosystems in our ever-changing biosphere.

Big Data need Standards

The term Big Data is increasingly being used to describe the vast capacity of high-throughput experimental methodologies, especially next-generation sequencing, to generate data [1,2]. Sharing and re-use of such data, and translating such data into knowledge, requires widely-adopted standards that are best developed within the auspices of international working groups [3]. Here we describe a new standard, developed by a large and diverse community of researchers, to describe one of the most abundant and useful types of sequence data – that of marker gene data sets.

The wealth of marker gene data sets

The adoption of phylogenetic marker genes as molecular proxies for tracking and cataloguing the diversity of microorganisms has revolutionized the way we view the biological world, and provided us with insights into how life has evolved and how different organisms are genetically related to each other. In the 1970s, studies of small subunit (SSU) ribosomal RNA (rRNA) genes from environmental samples led to the discovery of the domain *Archaea* [4] and to the proposal for a three domain classification of life [5]. Following Darwin's insight that all life is related, SSU rRNA gene surveys allow organisms from any communities, no matter how diverse, to be compared using the same universal phylogenetic tree. This rRNA gene-based molecular approach to characterizing natural communities of organisms provided, for the first time, culture-independent access to the diversity and distribution of microorganisms '*in situ*'. As a result, we are now acutely aware that the vast majority (90-99%) of microorganisms have evaded isolation using existing cultivation methods [6-8].

143 Over the past three decades, the 16S rRNA, 18S rRNA and internal transcribed spacer
144 gene sequences (ITS) from *Bacteria*, *Archaea*, and microbial *Eukaryotes* have provided
145 deep insights into the topology of the tree of life [9-12] and the composition of communities
146 of organisms that live in diverse environments, which range from deep sea hydrothermal
147 vents to ice sheets in the Arctic [13-24].
148 Numerous other phylogenetic marker genes have also proven useful [25]: Currently, around
149 40 such phylogenetic marker genes are in wide use, representing well-conserved,
150 housekeeping genes that include initiation factors, for example, RNA polymerase
151 subunits (*rpoB*), DNA gyrases (*gyrB*), DNA recombination and repair proteins (*recA*) and
152 heat shock proteins (*HSP70*) [10,26]. Most of these genes support or complement the
153 currently accepted topology of the Tree of Life. Combinations of these genes can also be
154 used in multi-locus sequence typing (MLST) approaches, increasing phylogenetic
155 resolution and differentiating between closely related species of the same genus [27,28].
156 Marker genes can also reveal key metabolic functions rather than phylogeny; examples
157 include nitrogen cycling (*amoA, nifH, ntcA*) [29,30], sulfate reduction (*dsrAB*) [31] or
158 phosphorus metabolism (*phnA, phnI, phnJ*) [32-34].
159 The molecular approach has been extended beyond microorganisms by its application to
160 phylogeny and systematics of higher *Eukaryotes.* The Barcode of Life Initiative (BOLI)
161 has adapted the molecular approach with the standardized use of a specific gene
162 sequence: the 680 base-pair region of mitochondrial cytochrome c oxidase I ("COI"), as a
163 means of rapid species identification and discrimination [35].
164 In this paper we collectively define all of these different phylogenetic and functional
165 genes (or gene fragments) as 'marker genes' as they are used to profile natural genetic

diversity across the Tree of Life, and argue that a small amount of additional effort invested in describing them with specific guidelines in our public databases will revolutionize the types of studies that can be performed with these large data resources. This effort is timely, given the need to determine how climate change and various other anthropogenic perturbations of our biosphere are affecting biodiversity, and how marked changes in our cultural traditions and lifestyles are affecting human microbial ecology.

The collective value of marker gene sequences

The quality and quantity of marker gene sequence data used to make phylogenetic assignments, to infer metabolic traits, and to assess biogeographic distributions continues to increase rapidly due to the availability of next generation sequencing (NGS) technologies powering the ability to study increasingly complex and/or divergent ecosystems.

Some recent large-scale molecular studies have endeavored to use environmental variables to explain the diversity and distribution of microbes. For example, a clear correlation between phylogenetic similarity and similar living conditions was observed using data in available SSU sequence repositories and culture collections [36]. In addition, two separate global environmental studies established a latitudinal diversity gradient for marine *Bacteria* [37,38]. Furthermore, it was shown that temporally-driven environmental factors, such as temperature and nutrients, correlate with local seasonal succession of marine microbial communities [39]. In a cross-habitat study, salinity and pH have been suggested to influence bacterial and archaeal community compositions, respectively [40,41]. In a different habitat, the human body, it has been suggested that the microbial

community composition varies systematically across body habitats, individuals and time [42].

For multicellular organisms, modeling approaches to predict global distributions of marine species have been applied in projects such as AquaMaps [43]. Combination of such efforts with the potential of COI to unveil historical processes may successfully be applied in determining factors responsible for the contemporary geographic distributions of these organisms [44].

Unfortunately, only a few of these large-scale environmental surveys of biodiversity and biogeography have relied on *existing* marker gene sequence data sets found in the public databases [36,40,45]. Mainly due to the lack of specific guidelines, most marker gene sequences in databases are sparsely annotated with the information that would be required to underpin data integration, comparative studies, and knowledge generation. Even with complex keyword searches, it is currently impossible to reliably retrieve marker gene 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.

With appropriate sequence and contextual data integration and analysis new potential explanations for observed distribution and abundance patterns of microorganisms can be unraveled. For example, in human health and the study of epidemiology, it would also be desirable to have additional contextual data to help monitor the origins and regional spreading of pandemics [46] and study the variation of the human microbiota [47-49]. Combining clinical and environmental datasets could provide new insight into where the trillions of bacteria that inhabit our body come from, and could help predict new

outbreaks of disease or assist in understanding the normal ecology of occasional pathogens. Already known correlations of some microbial taxa in different environments, such as depth in the marine environment [50,51], and pH in the soil environment [52], can be extended further. Finally, micro- and macro-organismal taxonomic knowledge can be greatly enhanced with the preservation of 'voucher specimens', which serve as the basis of study and as a reference. These may be cultures, tissue lines, DNA, or even images, depending on the organisms and the traits involved. The literature is filled with discoveries that could not be validated due to lack of vouchers.

The MIENS Specification

Few of the publicly available marker gene datasets contain contextual information about the environment such as geographic location, sampling time, habitat, or about experimental procedures used to obtain the DNA sequences. Such information may or may not be available in associated publications but the 'costs' in terms of time and energy to collect this by hand or with semi-automated systems from the literature are prohibitive [53]. Public databases of the International Nucleotide Sequence Database Collaboration (INSDC; comprising of DDBJ, ENA, and GenBank; http://www.insdc.org) depend on information submitted by authors to enrich the value of these sequences. 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 [3]. The adoption of such a standard would elevate the quality, accessibility, and utility of information that can be collected from INSDC.

Here we present a reporting guideline for marker genes (MIENS: Minimum Information

about an ENvironmental Sequence), which is based on the "Minimum Information about a (Meta) Genome Sequence" (MIGS/MIMS) specification issued by the Genomic Standards Consortium (GSC) [54]. Since its proposal at the sixth GSC meeting in 2008 [55], the consortium has been working to build a consensus on an ideal and minimum set of contextual data that should be reported for marker genes retrieved from the environment.

The proposed MIENS standard (Table 1) extends the MIGS/MIMS specification for genomes and metagenomes by adding two new report types, a "MIENS-survey" and a "MIENS-culture", the former being the checklist of choice for uncultured diversity marker gene surveys, the latter designed for marker gene sequences obtained from cultured organisms or any material identifiable via voucher specimens.

A specific focus of the extended requirements is the sets of measurements and observations describing particular habitats, termed 'environmental packages'.

The MIENS checklist adopts and incorporates the standards being developed by the Consortium for the Barcode of Life (CBOL) (http://www.barcoding.si.edu/PDF/DWG_data_standards-Final.pdf). Therefore, the specification can be universally applied to any marker gene, from SSU rRNA to COI, to cultured and uncultured organisms, to all taxa and to studies ranging from single individuals to complex communities.

The MIENS checklist was developed by collating information from several sources and evaluating it in the framework of the existing MIGS/MIMS specification. 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 MIENS specification that describes the most important aspects of marker gene contextual data,

and that would encourage users to deposit this contextual data in a standardized fashion.

Results of community-led surveys

Community surveys are an excellent way to determine researcher preferences for core descriptors. To date, there have been four online surveys about descriptors for marker genes. In the same manner as the Department of Energy Joint Genome Institute's (DOE-JGI) user survey focusing on general descriptor contextual data for marker genes in 2005, the Ribosomal Database Project (RDP) [56,57], SILVA [58] and the Terragenome Consortium (http://www.terragenome.org) conducted three more user surveys focusing on prevalent habitats for rRNA gene surveys, general descriptor contextual data for rRNA gene sequences and soil metagenome project contextual data, respectively (supplementary information 1). Additionally, following a special session during the 2005 International Census of Marine Microbes (ICoMM), an extensive set of contextual data items were selected, and were analyzed along with survey results.

These results of these user surveys provided valuable insights into community requests for contextual data items to be included in the MIENS specification and main habitats constituting the environmental packages.

Survey of published parameters

We reviewed published rRNA gene studies, retrieved via SILVA and the ICoMM database MICROBIS (http://icomm.mbl.edu/microbis) to further supplement contextual data items that are included in the respective environmental packages. In total, thirty-nine publications from SILVA; including twenty-three publications with more than 500

sequences, and thirteen others retrieved with habitat-specific study queries; and over 40 ICoMM projects were scanned for contextual data items to constitute the core of the environmental package sub-tables (supplementary information 1).

Survey of INSDC source feature qualifiers

As a final analysis step, we surveyed usage statistics of INSDC source feature key qualifier values of rRNA gene sequences contained in SILVA (supplementary information 1). Most striking of these results is that <10% of the 1.2 million 16S rRNA gene sequences (SILVA release 100) were associated with even basic information such as latitude/longitude, collection date or PCR primers.

The MIENS checklist in full

The MIENS specification provides users with an 'electronic laboratory notebook' containing core contextual data items required for consistent reporting of marker gene investigations. A number of experts in a wide array of topics, guided by a solid rationalization procedure at each step along the way, led the development of these contextual data items.

Project details are hosted in the 'Investigation' section of MIENS, facilitating access to the outline of contextual data of a marker gene survey. The 'Environment' section provides the geospatial, temporal and environmental context. Fourteen 'environmental-packages' were developed, with the assistance from user surveys, publication reviews and expert communities working on their respective environments, and were integrated into the 'MIMS/MIENS extension' section. These packages provide a wealth of

304 environmental and epidemiological contextual data fields for a complete description of
305 sampling environments (supplementary information 2). Researchers within The Human
306 Microbiome Project [59] contributed the host associated and all human packages. The
307 Terragenome Consortium contributed sediment and soil packages. Finally, ICoMM,
308 Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research
309 Sites (MIRADA-LTERS), and the Max Planck Institute for Marine Microbiology
310 contributed the water package. The MIENS working group developed the remaining
311 packages (air, microbial mat/biofilm, miscellaneous natural or artificial environment,
312 plant-associated, and wastewater/sludge). The package names describe high-level habitat
313 terms in order to be exhaustive. The miscellaneous natural or artificial environment
314 package contains a generic set of parameters, and is included for any other habitat that
315 does not fall into the other thirteen categories. Whenever needed, multiple packages may
316 be used for the description of the environment.
317 The MIGS/MIMS specifications are applicable to MIENS with respect to the nucleic acid
318 sequence source and sequencing contextual data, but have been complemented with
319 further experimental contextual data such as PCR primers and conditions, or target
320 gene/locus.
321 For clarity and ease of use, all items within the MIENS specification are presented with a
322 value syntax description, as well as a clear definition of the item. Whenever terms from a
323 specific ontology are required as the value of an item, these terms can be readily found in
324 the respective ontology browsers, which are linked by URLs in the item definition.
325 Although this version of the MIENS specification does not contain unit specifications, we
326 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. To facilitate comparative studies, unit standardization across data sets will be vital in future versions of MIENS.

Accessing the MIGS/MIMS/MIENS checklists

The MIGS/MIMS/MIENS checklists are maintained in a relational database system on behalf of the GSC community. This provides a secure and stable mechanism for updating the checklist suite and versioning. An excel version of the checklist is also provided to the community on the GSC web site at: http://gensc.org/gc_wiki/index.php/MIENS. The checklist is updated on the GSC web site as development work is carried out on the database end. In the future, we plan to develop programmatic access to this database in order to allow automatic retrieval of the latest version of each checklist for INSDC databases and for GSC community resources.

MIENS Adoption by Major Database and Informatics Resources

A variety of efforts are under way to aid sequence submitters in compliance. In the past, the INSDC has issued a reserved 'BARCODE' keyword for the Consortium for the Barcoding of Life (CBOL) [60,61]. Following this model, the INSDC has recently recognized the GSC as an authority for the MIGS/MIMS/MIENS standards and issued it with an official keyword within INSDC nucleotide sequence records [62]. This greatly facilitates automatic validation of the submitted contextual data and provides support for datasets compliant with previous versions by including the checklist version in the keyword.

GenBank accepts MIENS metadata in tabular format using the sequin and tbl2asn submission tools, validates MIENS compliance, and reports the MIENS fields in the structured comment block. The ENA Webin submission system provides prepared web forms for the submission of MIENS compliant data; it presents all of the appropriate fields with descriptions, explanations and examples, in addition to validation of the data entered in the forms. An example which can aid in submission via Sequin or Webin systems is MetaBar [63]; a spreadsheet and web-based software tool designed to assist users in the consistent acquisition, electronic storage and submission of contextual data associated with their samples in compliance with the MIGS/MIMS/MIENS specifications.

The next-generation Sequence Read Archives (SRA) collects and displays MIENS compliant metadata in the sample and experiment objects. There are several tools that are already available or under development to assist users in SRA and ERA 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 the MIMS and MIENS specifications. The Investigation/Study/Assay (ISA) Infrastructure is a flexible, freely available software suite that assists in the curation, reporting, and local management of experimental metadata from studies employing one or a combination of technologies, including high-throughput sequencing. Specific ISA configurations (available from http://gensc.org/gc_wiki/index.php/Adopters#ISA_infrastructure) have been developed to ensure MIENS compliance by providing templates and validation capability while another tool, ISAconverter, produces SRA.xml documents, thereby

facilitating submission to ERA and SRA repositories [64].

The SILVA, RDP, Greengenes and the ICoMM resources have participated in the development of MIENS, and are now taking the standardization one step further by establishing tools and resources to aid in compliance.

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

Examples of MIENS compliant datasets

Several MIENS compliant reports are included in the supplementary information 3. These include; a 16S rRNA gene survey from samples obtained in the North Atlantic, an 18S pyrotag study of anaerobic protists in permanently anoxic basin of the North Sea, a *pmoA* survey from desert soils of Negev Desert, Israel, a *dsrAB* survey from marine sediments from the Gulf of Mexico, and finally a 16S pyrotag study of bacterial diversity in the Western English Channel (publicly accessible via SRA study accession number SRP001108). Two further MIENS compliant 16S submissions are available in INSDC under the accession numbers GU949561.1 and GU949562.1.

MIENS – a 'living standard'

MIENS, as well as MIGS/MIMS, are 'living checklists' and not final specifications. Therefore, further developments, extensions, and enhancements will be recognized, and improved versions of the checklists, if necessitated, will be released annually, while preserving the validity of former versions. A public ticketing system will be set up to track changes and feature requests. The final decisions about their implementation will be

done by the MIENS working group.

Conclusions and Call for Action

The GSC (www.gensc.org) is an international working body with a stated mission of working towards richer descriptions of our complete collection of genomes and metagenomes. With the development of the MIENS specification, this mission has been extended to marker gene sequences as well. 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 MIENS specification or its implementation.

The adoption of the MIENS standard by major data providers and organizations as well as the three primary public sequence data repositories (INSDC) with an active poll for MIENS compliant data underlines and seconds the efforts to contextually enrich our marker gene collection, and complements the recent efforts to contextually enrich other (meta) omics data. The MIENS checklist has been developed to the point that it is ready to be used in the publication of sequences. A defined procedure for requesting new features and the stable release cycles will facilitate implementation of the standard across the community. Widespread 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 yardstick to compare even the most diverse communities, should provide new insights into the dynamic spatial and temporal distribution of microbial life on our planet and even on our own bodies.

References

1 Community cleverness required. *Nature* **455**, 1-1 (2008).

2 Field, D. *et al.*, 'Omics Data Sharing. *Science* **326**, 234-236 (2009).

3 Taylor, C. F. *et al.*, Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* **26**, 889-896 (2008).

4 Woese, C. R. and Fox, E., Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Nat Acad Sci USA* **74**, 5088-5090 (1977).

5 Woese, C. R., Kandler, O., and Wheelis, M. L., Towards a natural system of organisms: proposal for the domains *Archaea, Bacteria,* and *Eucarya*. *Proc Nat Acad Sci USA* **87**, 4576-4579 (1990).

6 Amann, R. I., Ludwig, W., and Schleifer, K. H., Phylogenetic identification and in-situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59**, 143-169 (1995).

7 Curtis, T. P., Sloan, W. T., and Scannell, J. W., Estimating prokaryotic diversity and its limits. *Proc Nat Acad Sci USA* **99**, 10494-10499 (2002).

8 Turroni, F. *et al.*, Human gut microbiota and bifidobacteria: from composition to functionality. *Antonie van Leeuwenhoek* **94**, 35-50 (2008).

9 Ludwig, W. *et al.*, Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554-568 (1998).

10 Ludwig, W. and Schleifer, K. H., in *Microbial phylogeny and evolution, concepts and controversies*, edited by J. Sapp (Oxford university press, New York, 2005), pp. 70-98.

11 Ciccarelli, F. D. *et al.*, Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287 (2006).

12 Teeling, H. and Glöckner, F. O., RibAlign: a software tool and database for eubacterial phylogeny based on concatenated ribosomal protein subunits. *BMC Bioinformatics* **7** (2006).

13 Stahl, D. A., Analysis of hydrothermal vent associated symbionts by ribosomal RNA sequences. *Science* **224**, 409-411 (1984).

14 Pace, N. R., Stahl, D. A., Olsen, G. J., and Lane, D. J., Analyzing natural

microbial populations by rRNA sequences. *ASM News* **51**, 4-12 (1985).

15 Olsen, G. J. *et al.*, Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**, 337-365 (1986).

16 Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., and Field, K. G., Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60-63 (1990).

17 Ward, D. M., Weller, R., and Bateson, M. M., 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**, 63-65 (1990).

18 López-García, P., López-López, A., Moreira, D., and Rodríguez-Valera, F., Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. *Fems Microbiol Ecol* **36**, 193-202 (2001).

19 Moon-van der Staay, S. Y., De Wachter, R., and Vaulot, D., Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607-610 (2001).

20 DeLong, E. F., *Archaea* in coastal marine environments. *Proc Nat Acad Sci USA* **89**, 5685-5689 (1992).

21 Pace, N. R., A molecular view of microbial diversity and the biosphere. *Science* **276**, 734-740 (1997).

22 Huber, J. A., Butterfield, D. A., and Baross, J. A., Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. *Appl Environ Microbiol* **68**, 1585-1594 (2002).

23 Rappe, M. S. and Giovannoni, S. J., The uncultured microbial majority. *Annu Rev Microbiol* **57**, 369-394 (2003).

24 Hewson, I. and Fuhrman, J. A., Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. *Appl Environ Microbiol* **70**, 3425-3433 (2004).

25 Doolittle, W. F., Fun With Genealogy. *Proc Nat Acad Sci USA* **94**, 12751-12753 (1997).

26 Huynen, M. A. and Bork, P., Measuring genome evolution. *Proc Nat Acad Sci USA* **95**, 5849-5856 (1998).

27 Ivars-Martinez, E. *et al.*, Biogeography of the ubiquitous marine bacterium *Alteromonas macleodii* determined by multilocus sequence analysis. *Mol Ecol* **17**, 4092-

4106 (2008).

28 Cole, J. R., Konstantinidis, K., Farris, R. J., and Tiedje, J. M., in *Environmental Molecular Microbiology*, edited by W.-T. Liu and J.K. Jansson (Caister Academic Press UK, 2010), pp. 1-19.

29 Zehr, J. P., Mellon, M. T., and Zani, S., New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (nifH) genes. *Appl Environ Microbiol* **64**, 3444-3450 (1998).

30 Francis, C. A., Beman, J. M., and 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).

31 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).

32 Thomas, S. *et al.*, Evidence for phosphonate usage in the coral holobiont. *Isme J* (2009).

33 Gilbert, J. A. *et al.*, Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters. *Environ Microbiol* **11**, 111-125 (2009).

34 Martinez, A., W. Tyson, G., and DeLong, E., F., Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol* **9999** (2009).

35 Hebert, P. D. N., Cywinska, A., Ball, S. L., and Dewaard, J. R., Biological identifications through DNA barcodes. *Proc R Soc Lond B Biol Sci* **270**, 313-321 (2003).

36 von Mering, C. *et al.*, Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**, 1126-1130 (2007).

37 Pommier, T. *et al.*, Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16**, 867-880 (2007).

38 Fuhrman, J. A. *et al.*, A latitudinal diversity gradient in planktonic marine bacteria. *Proc Nat Acad Sci USA* **105**, 7774-7778 (2008).

39 Gilbert, J., A. *et al.*, The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* **11**, 3132-3139 (2009).

40 Lozupone, C. A. and Knight, R., Global patterns in bacterial diversity. *Proc Nat*

512 *Acad Sci USA* **104**, 11436-11440 (2007).
513 41 Auguet, J.-C., Barberan, A., and Casamayor, E. O., Global ecological patterns in
514 uncultured Archaea. *Isme J* **4**, 182-190 (2010).
515 42 Costello, E. K. *et al.*, Bacterial community variation in human body habitats
516 across space and time. *Science* **326**, 1694-1697 (2009).
517 43 Kaschner, K. *et al.*, AquaMaps: Predicted range maps for aquatic species,
518 Available at http://www.aquamaps.org/, (2008).
519 44 Workshops Report and Recommendations DNA Barcoding of Marine
520 Biodiversity (MarBOL) presented at the MarBOL Workshops, 2009 (unpublished).
521 45 Tamames, J. *et al.*, Environmental distribution of prokaryotic taxa. *BMC*
522 *Microbiology* **10**, 85.
523 46 Schriml, L. M. *et al.*, GeMInA, Genomic Metadata for Infectious Agents, a
524 geospatial surveillance pathogen database. *Nucl Acids Res* **38**, D754-D764 (2010).
525 47 Palmer, C. *et al.*, Development of the Human Infant Intestinal Microbiota. *PLoS*
526 *Biol* **5**, e177 (2007).
527 48 Ravel, J. *et al.*, Vaginal microbiome of reproductive-age women. *Proc Nat Acad*
528 *Sci USA* **e-pub ahead of print** (2010).
529 49 Qin, J. *et al.*, A human gut microbial gene catalogue established by metagenomic
530 sequencing. *Nature* **464**, 59-65 (2010).
531 50 DeLong, E. F. *et al.*, Community genomics among stratified microbial
532 assemblages in the ocean's interior. *Science* **311**, 496-503 (2006).
533 51 Moreira, D., Rodriguez-Valera, F., and Lopez-Garcia, P., Metagenomic analysis
534 of mesopelagic Antarctic plankton reveals a novel deltaproteobacterial group.
535 *Microbiology* **152**, 505-517 (2006).
536 52 Lauber, C. L., Hamady, M., Knight, R., and Fierer, N., Soil pH as a predictor of
537 soil bacterial community structure at the continental scale: a pyrosequencing-based
538 assessment. *Appl Environ Microbiol* **75**, 5111-5120 (2009).
539 53 Hirschman, L. *et al.*, Habitat-Lite: a GSC case study based on free text terms for
540 environmental metadata. *OMICS* **12**, 129-136 (2008).
541 54 Field, D. *et al.*, The minimum information about a genome sequence (MIGS)
542 specification. *Nat Biotechnol* **26**, 541-547 (2008).

55 Field, D. *et al.*, Meeting reports from the Genomic Standards Consortium (GSC) Workshops 6 and 7. *SIGS* **1**, 68-71 (2009).

56 Cole, J. R. *et al.*, The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucl Acids Res* **35**, D169-172 (2007).

57 Cole, J. R. *et al.*, The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucl Acids Res* **37**, D141-145 (2009).

58 Pruesse, E. *et al.*, SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Acids Res* **35**, 7188-7196 (2007).

59 Turnbaugh, P. J. *et al.*, The Human Microbiome Project. *Nature* **449**, 804-810 (2007).

60 Benson, D. A. *et al.*, GenBank. *Nucl. Acids Res.* **35**, D21-25 (2007).

61 Benson, D. A. *et al.*, GenBank. *Nucl. Acids Res.* **36**, D25-30 (2008).

62 Lynette, H. *et al.*, Meeting report: Metagenomics, Metadata and Meta-analysis" (M3) Workshop at the Pacific Symposium on Biocomputing 2010. *SIGS* **2**, 357-360 (2010).

63 Hankeln, W. *et al.*, MetaBar - a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinformatics* **11**, 358 (2010).

64 Rocca-Serra, P. *et al.*, ISA infrastructure: supporting standards-compliant experimental reporting and enabling curation at the community level. *Bioinformatics* **in press** (2010).

		Report type	
		MIENS survey	**MIENS culture**
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 (Short Read Archives), ENA (European Nucleotide Archive), DRA (DDBJ Read Archive) or via the classical Webin/Sequin systems to Genbank, ENA and DDBJ	M	M
Investigation type [survey or culture]	Nucleic Acid Sequence Report is the root element of all MIENS compliant reports as standardized by Genomic Standards Consortium (GSC). This field is either MIENS survey or MIENS culture	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 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 i.e. 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: 2008-01; 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. EnvO terms listed under environmental biome can be found from the link: http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=ENVO&termId=ENVO%3A00000428&termName=biome	M	M
Environment (feature [EnvO])	Environmental feature level includes geographic environmental features. Examples include: harbor, cliff, or lake. EnvO terms listed under environmental feature can be found from the link: http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=ENVO&termId=ENVO%3A00002297&termName=environmental%20feature	M	M

Environment (material [EnvO])	The environmental material level refers to the matter that was displaced by the sample, prior to the sampling event. Environmental matter terms are generally mass nouns. Examples include: air, soil, or water. EnvO terms listed under environmental matter can be found from the link: http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=ENVO&termId=ENVO%3A00010483&termName=environmental%20matter	M	M
MIGS/MIMS/MIENS 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/MIENS 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, V6, ITS)	Targeted gene, locus or gene region name for marker gene study	M	M
Sequencing method (e.g. dideoxysequencing, pyrosequencing, polony) [OBI]	Sequencing method used; e.g. Sanger, pyrosequencing, ABI-solid. This field accepts OBI terms, for a browser of OBI terms please see http://bioportal.bioontology.org/visualize/40547	M	M

Table 1. Items for the MIENS specification and their mandatory (M), conditionally mandatory (C) (the item is mandatory only when applicable to the study) or recommended (X) status for both MIENS-survey and MIENS-culture checklists. '-' denotes that an item is not applicable for a given checklist. 'E' denotes that a field has environment-specific requirements. For example, while 'depth' is mandatory for environments water, sediment or soil; it is optional for human-associated environments. 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) from the OBO foundry

(http://www.obofoundry.org), EFO (ArrayExpress Environmental Factor Ontology), EnvO (Environment Ontology), CABRI (Common Access to Biological Resources and Information), OBI (Ontology for Biomedical Investigations). Items that are mandatory or conditionally mandatory for both MIENS report types are indicated by shaded rows and constitute the core of the MIENS specification. This table only presents the very core of MIENS checklists, i.e. only mandatory items for each checklist. Supplementary information 2 in spreadsheet format contains all MIENS items, the tables for environmental packages in the MIMS/MIENS extension, and GenBank structured comment name that should be used for submitting MIENS data to GenBank.