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], Tim Booth[3], Peer Bork[18], Frederic D. Bushman[23], Pier Luigi Buttigieg[1,2], Patrick S. G. Chain[7,24,25], Emily Charlson[23], Elizabeth K. Costello[4], Heather Huot-Creasy[26], Peter Dawyndt[27], Todd DeSantis[28], Noah Fierer[29], Jed Fuhrman[30], Rachel E. Gallery[31], Dirk Gevers[20], Richard A. Gibbs[32,33], Michelle Gwinn Giglio[26], Inigo San Gil[34], Antonio Gonzalez[35], Jeffrey I. Gordon[36], Robert Guralnick[29], Wolfgang Hankeln[1,2], Sarah Highlander[32,37], Philip Hugenholtz[24], Janet Jansson[38], Scott T. Kelley[39], Jerry Kennedy[4], Dan Knights[35], Omry Koren[40], Justin Kuczynski[19], Nikos Kyrpides[24], Robert Larsen[4], Christian L. Lauber[41], Teresa Legg[29], Ruth E. Ley[40], Catherine A. Lozupone[4], Wolfgang Ludwig[42], Donna Lyons[41], Eamonn Maguire[13,17], Barbara A. Methé[43], Folker Meyer[10], Sara Nakielny[4], Karen E. Nelson[43], Diana Nemergut[44], Lindsay K. Neubold[3], Josh D. Neufeld[45], Anna E. Oliver[3], Norman R. Pace[19], Giriprakash Palanisamy[46], Jörg Peplies[47], Jane Peterson[22], Joseph Petrosino[32,37], Lita Proctor[48], Elmar Pruesse[1,2], Christian Quast[1], Jeroen Raes[49], Sujeevan Ratnasingham[50], Jacques Ravel[26], David A. Relman[51,52], Susanna Assunta-Sansone[13,17], Patrick D. Schloss[53], Lynn Schriml[26], Rohini Sinha[23], Erica Sodergren[54], Aymé Spor[40], Jesse Stombaugh[4], James M. Tiedje[7], Doyle V. Ward[20], George M.

Weinstock[54], Doug Wendel[4], Owen White[26], Andrew Whitely[3], Andreas Wilke[10], Jennifer R. Wortman[26], 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 Department of Microbiology, University of Pennsylvania School of Medicine, 426A Johnson Pavilion, 3610 Hamilton Walk, Philadelphia, PA 19104

24 DOE Joint Genome Institute, Walnut Creek, CA 94598, USA

25 Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA

26 Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

27 Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium

28 Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

29 Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309, USA

30 Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

31 National Ecological Observatory Network (NEON), Boulder, CO 80301, USA

32 Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

33 Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

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

35 Department of Computer Science, University of Colorado, Boulder, CO 80309, USA

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

37 Department of Molecular Virology and Microbiology, Baylor College of Medicine,

Houston, TX

38 Lawrence Berkeley National Laboratory, Earth Science Division, Berkeley, CA, USA

39 Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego , CA 92182-4614 USA

40 Department of Microbiology, Cornell University, Ithaca NY 14853, USA

41 Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, USA

42 Lehrstuhl für Mikrobiologie, Technische Universität München, D-853530 Freising, Germany

43 J. Craig Venter Institute, Rockville, Maryland, United States of America

44 Department of Environmental Sciences, University of Colorado, Boulder, CO 80309, USA

45 Department of Biology, University of Waterloo, Ontario, N2L 3G1, Canada

46 Environmental Sciences Division, Oak Ridge National Laboratory, Mail Stop 6407 Oak Ridge, TN, USA

47 Ribocon GmbH, D-28359 Bremen, Germany

48 The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA

49 VIB - Vrije Universiteit Brussel, 1050 Brussels, Belgium

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

51 Departments of Microbiology and Immunology and of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

52 Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304, USA

53 Department of Microbiology and Immunology, 5641 Medical Science Bldg. II, 1150 West Medical Center Dr., Ann Arbor, Michigan 48109-5620

54 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.

Acronyms

amoA: ammonia monooxygenase-alpha subunit

BOLI: Barcode of Life Initiative

CBOL: Consortium for the Barcode of Life

COI: cytochrome c oxidase I

DDBJ: DNA DataBank of Japan

DOE-JGI: Department of Energy Joint Genome Institute

DOI: Digital Object Identifier

DRA: DDBJ Sequence Read Archive

dsrAB: dissimilatory sulfite reductase

ENA: European Nucleotide Archive

EnvO: Environment Ontology

GAZ: Gazetteer

GCDML: Genomic Contextual Data Markup Language

GSC: Genomic Standards Consortium

gyrA: DNA gyrase (type II topoisomerase), subunit A

HSP70: 70 kilodalton heat shock protein

ICoMM: International Census of Marine Microbes

INSDC: International Nucleotide Sequence Database Collaboration

ISA: Investigation/Study/Assay Infrastructure

ISO: International Organization for Standardization

ITS: internal transcribed spacer region

LSU: large subunit

MICROBIS: The Microbial Oceanic Biogeographic Information System

MIENS: Minimum Information about an Environmental Sequence

MIGS/MIMS: Minimum Information about a Genome/Metagenome Sequence

MIRADA-LTERS: Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites

MLST: multi-locus sequence typing

NGS: next generation sequencing

nifH: dinitrogenase reductase

ntcA: nitrogen regulator gene

OBO: Open Biological and Biomedical Ontologies

phnA: phosphonoacetate hydrolase gene

phnJ: carbon-phosphorous lyase complex subunit

PMID: Pubmed ID

RDP: Ribosomal Database Project

recA: recombinase A subunit

rpoB: beta subunit of the bacterial RNA polymerase

rRNA: ribosomal RNA

SI: International System of Units

SRA: Sequence Read Archive

SSU: small subunit

URL: Uniform Resource Locator

WGS84: World Geodetic System 84

XML Schema: Extensible Markup Language Schema

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].

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 [9-12] and the composition of communities of organisms that live in diverse environments, which range from deep sea hydrothermal vents to ice sheets in the Arctic [13-27].

Numerous other phylogenetic marker genes have also proven useful [28]: Currently, around 40 such phylogenetic marker genes are in wide use, representing well-conserved, housekeeping genes that include initiation factors, for example, RNA polymerase subunits (*rpoB*), DNA gyrases (*gyrB*), DNA recombination and repair proteins (*recA*) and heat shock proteins (*HSP70*) [10,29]. Combinations of these genes can also be used in multi-locus sequence typing (MLST) approaches, increasing phylogenetic resolution and differentiating between closely related species of the same genus [30,31]. Marker genes can also reveal key metabolic functions rather than phylogeny; examples include nitrogen cycling (*amoA, nifH, ntcA*) [32,33], sulfate reduction (*dsrAB*) [34] or phosphorus metabolism (*phnA, phnI, phnJ*) [35-37].

The molecular approach has been extended beyond microorganisms by its application to phylogeny and systematics of higher *Eukaryotes.* The Barcode of Life Initiative (BOLI) adapted the molecular approach with the standardized use of a specific gene sequence: the 680 base-pair region of mitochondrial cytochrome c oxidase I (COI), as a means of rapid species identification and discrimination [38].

In this paper we collectively define all of these different phylogenetic and functional 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

222 invested in describing them with specific guidelines in our public databases will
223 revolutionize the types of studies that can be performed with these large data resources.
224 This effort is timely, given the need to determine how climate change and various other
225 anthropogenic perturbations of our biosphere are affecting biodiversity, and how marked
226 changes in our cultural traditions and lifestyles are affecting human microbial ecology.
227
228 ***The collective value of marker gene sequences***
229 The quality and quantity of marker gene sequence data used to make phylogenetic
230 assignments, to infer metabolic traits, to unravel succession as a function of the
231 environment, and to assess biogeographic distributions continues to increase rapidly due
232 to the availability of next generation sequencing (NGS) technologies. Clearly, specific
233 associations of microbial dynamics with the environment and geography were achieved
234 for cultured microorganisms long before the advent of metagenomics and NGS
235 technologies [39-42]. However, with the new powerful technologies at our service, it is
236 possible to unravel the diversity and function of the uncultured majority as well as to
237 study increasingly complex and/or divergent ecosystems. For example, a clear correlation
238 between phylogenetic similarity and similar living conditions was observed using data in
239 available SSU sequence repositories and culture collections [43]. In addition, two separate
240 global environmental studies established a latitudinal diversity gradient for marine
241 *Bacteria* [44,45]. Furthermore, it was shown that temporally-driven environmental factors,
242 such as temperature and nutrients, correlate with local seasonal succession of marine
243 microbial communities [46]. In a cross-habitat study, salinity and pH have been suggested
244 to influence bacterial and archaeal community compositions, respectively [47,48]. In the

245 human body, it has been suggested that the microbial community composition varies
246 systematically across body habitats, individuals and time [49]. A recent study combined
247 habitat type and 16S rRNA based operational taxonomic units (OTUs) in a graph-
248 theoretic approach to demonstrate that different habitats harbor unique assemblages of
249 co-occurring microorganisms [50]. For multicellular organisms, modeling approaches to
250 predict global distributions of marine species have been applied in projects such as
251 AquaMaps [51]. Combination of such efforts with the potential of COI to unveil historical
252 processes may successfully be applied in determining factors responsible for the
253 contemporary geographic distributions of these organisms [52].
254 Unfortunately, only a few of these large-scale environmental surveys of biodiversity and
255 biogeography have relied on *existing* marker gene sequence data sets found in the public
256 databases [43,47,50,53]. Mainly due to the lack of specific guidelines, most marker gene
257 sequences in databases are sparsely annotated with the information that would be
258 required to underpin data integration, comparative studies, and knowledge generation.
259 Even with complex keyword searches, it is currently impossible to reliably retrieve
260 marker gene sequences that have originated from certain environments or particular
261 locations on Earth; for example, all sequences from ‘soil’ or ‘freshwater lakes’ in a
262 certain region of the world.
263 In human health and the study of epidemiology, it would also be desirable to have
264 additional contextual data to help monitor the origins and regional spreading of
265 pandemics [54] and study the variation of the human microbiota [55-57]. Combining clinical
266 and environmental datasets could provide new insight into where the trillions of bacteria
267 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 with different environmental conditions, such as depth in the marine environment [58,59], and pH in the soil environment [60], can be extended further. Careful integration of bacterial, archaeal and eukaryotic SSU and LSU rRNA sequence data with their geographical and environmental context can shed light on new mechanisms by which organisms from these three domains interact.

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 [61]. Public databases of the International Nucleotide Sequence Database Collaboration (INSDC; comprised of DDBJ (DNA Data Bank of Japan), ENA (European Nucleotide Archive), 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

291 a (Meta) Genome Sequence" (MIGS/MIMS) specification issued by the Genomic
292 Standards Consortium (GSC) [62]. Since its proposal at the sixth GSC meeting in 2008 [63],
293 the consortium has been working to build a consensus on an ideal and minimum set of
294 contextual data that should be reported for marker genes retrieved from the environment.
295 The proposed MIENS standard (Table 1) extends the MIGS/MIMS specification for
296 genomes and metagenomes by adding two new report types, a "MIENS-survey" and a
297 "MIENS-culture", the former being the checklist of choice for uncultured diversity
298 marker gene surveys, the latter designed for marker gene sequences obtained from
299 cultured organisms or any material identifiable via voucher specimens.
300 A specific focus of the extended requirements is the sets of measurements and
301 observations describing particular habitats, termed 'environmental packages'.
302 The MIENS checklist adopts and incorporates the standards being developed by the
303 Consortium for the Barcode of Life (CBOL) (http://www.barcoding.si.edu/PDF/
304 DWG_data_standards-Final.pdf). Therefore, the specification can be universally applied
305 to any marker gene, from SSU rRNA to COI, to cultured and uncultured organisms, to all
306 taxa and to studies ranging from single individuals to complex communities.
307 The MIENS checklist was developed by collating information from several sources and
308 evaluating it in the framework of the existing MIGS/MIMS specification. These include
309 four independent community-led surveys, examination of the parameters reported in
310 published studies, and examination of compliance with optional features in INSDC
311 documents. The overall goal of these activities was to design the backbone of the MIENS
312 specification that describes the most important aspects of marker gene contextual data,
313 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) [64,65], SILVA [66] and the Terragenome Consortium [67] 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.

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

Survey of published parameters

We reviewed published rRNA gene studies, retrieved via 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, 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 less than 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 environmental and epidemiological contextual data fields for a complete description of

sampling environments (supplementary information 2). Researchers within The Human Microbiome Project [68] 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 (MIRADA-LTERS), and the Max Planck Institute for Marine Microbiology contributed the water package. The MIENS 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.

The MIGS/MIMS specifications are applicable to MIENS with respect to the nucleic acid sequence source and sequencing contextual data, but have been complemented with further experimental contextual data such as PCR primers and conditions, or target gene/locus.

For clarity and ease of use, all items within the MIENS specification 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, which are linked by URLs in the item definition. Although this version of the MIENS specification does not 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. 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. Moreover, the Genomic Contextual Data Markup Language (GCDML) is a reference implementation of the MIGS/MIMS/MIENS checklists by the GSC. It is based on the XML Schema technology and thus serves as an interoperable data exchange format for Web Service based infrastructures [69].

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 CBOL [70,71]. 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 [72]. 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 of a tool that can aid in submission via Sequin or Webin systems is MetaBar [73], 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 Archive (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 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 the SRA repository [74].

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 the 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 issue tracking system, which can be

reached via http://mixs.gensc.org/, is set up to track changes and accomplish feature requests. The final decisions about their implementation will be concluded by the MIENS working group.

Conclusions and Call for Action

The GSC 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., Lane, D. J., Olsen, G. J., and Pace, N. R. 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 DeLong, E. F. *Archaea* in coastal marine environments. *Proc Nat Acad Sci USA* **89**, 5685-5689 (1992).

19 Fuhrman, J. A., McCallum, K., and Davis, A. A. Novel major archaebacterial group from marine plankton. *Nature* **356**, 148-149 (1992).

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

21 Diez, B., Pedros-Alio, C., and 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).

22 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).

23 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).

24 Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C., and Moreira, D. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603-607 (2001).

25 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).

26 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).

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

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

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

30 Ivars-Martinez, E. *et al.* Biogeography of the ubiquitous marine bacterium

546 *Alteromonas macleodii* determined by multilocus sequence analysis. *Mol Ecol* **17**, 4092-
547 4106 (2008).
548 31 Cole, J. R., Konstantinidis, K., Farris, R. J., and Tiedje, J. M. in *Environmental*
549 *Molecular Microbiology*, edited by W.-T. Liu and J.K. Jansson (Caister Academic Press
550 UK, 2010), pp. 1-19.
551 32 Zehr, J. P., Mellon, M. T., and Zani, S. New nitrogen-fixing microorganisms
552 detected in oligotrophic oceans by amplification of nitrogenase (nifH) genes. *Appl*
553 *Environ Microbiol* **64**, 3444-3450 (1998).
554 33 Francis, C. A., Beman, J. M., and Kuypers, M. M. M. New processes and players
555 in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia
556 oxidation. *Isme J* **1**, 19-27 (2007).
557 34 Minz, D. *et al.* Diversity of sulfate-reducing bacteria in oxic and anoxic regions of
558 a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase
559 genes. *Appl Environ Microbiol* **65**, 4666-4671 (1999).
560 35 Gilbert, J. A. *et al.* Potential for phosphonoacetate utilization by marine bacteria
561 in temperate coastal waters. *Environ Microbiol* **11**, 111-125 (2009).
562 36 Martinez, A., W. Tyson, G., and DeLong, E., F. Widespread known and novel
563 phosphonate utilization pathways in marine bacteria revealed by functional screening and
564 metagenomic analyses. *Environ Microbiol* **9999** (2009).
565 37 Thomas, S. *et al.* Evidence for phosphonate usage in the coral holobiont. *Isme J*
566 **4**, 459-461 (2010).
567 38 Hebert, P. D. N., Cywinska, A., Ball, S. L., and Dewaard, J. R. Biological
568 identifications through DNA barcodes. *Proc R Soc Lond B Biol Sci* **270**, 313-321 (2003).

569 39 ZoBell, C. E. and Johnson, F. H. The influence of hydrostatic pressure on the
570 growth and viability of terrestrial and marine bacteria. *J Bacteriol* **57**, 179 (1949).
571 40 Brock, T. D. and Brock, M. L. Relationship between Environmental Temperature
572 and Optimum Temperature of Bacteria along a Hot Spring Thermal Gradient. *J Appl*
573 *Microbiol* **31**, 54-58 (1968).
574 41 Cho, J.-C. and Tiedje, J. M. Biogeography and Degree of Endemicity of
575 Fluorescent Pseudomonas Strains in Soil. *Appl Environ Microbiol* **66**, 5448-5456 (2000).
576 42 Pomeroy, L. R. and Wiebe, W. J. Temperature and substrates as interactive
577 limiting factors for marine heterotrophic bacteria. *Aquat Microb Ecol* **23**, 187-204 (2001).
578 43 von Mering, C. *et al.* Quantitative phylogenetic assessment of microbial
579 communities in diverse environments. *Science* **315**, 1126-1130 (2007).
580 44 Pommier, T. *et al.* Global patterns of diversity and community structure in marine
581 bacterioplankton. *Mol Ecol* **16**, 867-880 (2007).
582 45 Fuhrman, J. A. *et al.* A latitudinal diversity gradient in planktonic marine bacteria.
583 *Proc Nat Acad Sci USA* **105**, 7774-7778 (2008).
584 46 Gilbert, J., A. *et al.* The seasonal structure of microbial communities in the
585 Western English Channel. *Environ Microbiol* **11**, 3132-3139 (2009).
586 47 Lozupone, C. A. and Knight, R. Global patterns in bacterial diversity. *Proc Nat*
587 *Acad Sci USA* **104**, 11436-11440 (2007).
588 48 Auguet, J.-C., Barberan, A., and Casamayor, E. O. Global ecological patterns in
589 uncultured Archaea. *Isme J* **4**, 182-190 (2010).
590 49 Costello, E. K. *et al.* Bacterial community variation in human body habitats across
591 space and time. *Science* **326**, 1694-1697 (2009).

50 Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res* **20**, 947-959 (2010).

51 Kaschner, K. *et al.* AquaMaps: Predicted range maps for aquatic species, Available at http://www.aquamaps.org/, (2008).

52 Workshops Report and Recommendations DNA Barcoding of Marine Biodiversity (MarBOL) presented at the MarBOL Workshops, 2009 (unpublished).

53 Tamames, J. *et al.* Environmental distribution of prokaryotic taxa. *BMC Microbiology* **10**, 85.

54 Schriml, L. M. *et al.* GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database. *Nucl Acids Res* **38**, D754-D764 (2010).

55 Palmer, C. *et al.* Development of the Human Infant Intestinal Microbiota. *PLoS Biol* **5**, e177 (2007).

56 Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc Nat Acad Sci USA* **e-pub ahead of print** (2010).

57 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).

58 DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496-503 (2006).

59 Moreira, D. Rodriguez-Valera, F., and Lopez-Garcia, P., Metagenomic analysis of mesopelagic Antarctic plankton reveals a novel deltaproteobacterial group. *Microbiology* **152**, 505-517 (2006).

60 Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. Soil pH as a predictor of

615 soil bacterial community structure at the continental scale: a pyrosequencing-based
616 assessment. *Appl Environ Microbiol* **75**, 5111-5120 (2009).
617 61 Hirschman, L. *et al.* Habitat-Lite: a GSC case study based on free text terms for
618 environmental metadata. *OMICS* **12**, 129-136 (2008).
619 62 Field, D. *et al.* The minimum information about a genome sequence (MIGS)
620 specification. *Nat Biotechnol* **26**, 541-547 (2008).
621 63 Field, D. *et al.* Meeting reports from the Genomic Standards Consortium (GSC)
622 Workshops 6 and 7. *SIGS* **1**, 68-71 (2009).
623 64 Cole, J. R. *et al.* The ribosomal database project (RDP-II): introducing myRDP
624 space and quality controlled public data. *Nucl Acids Res* **35**, D169-172 (2007).
625 65 Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new
626 tools for rRNA analysis. *Nucl Acids Res* **37**, D141-145 (2009).
627 66 Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked
628 and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Acids Res* **35**,
629 7188-7196 (2007).
630 67 Vogel, T. M. *et al.* TerraGenome: a consortium for the sequencing of a soil
631 metagenome. *Nat Rev Micro* **7**, 252-252 (2009).
632 68 Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804-810
633 (2007).
634 69 Kottmann, R. *et al.* A standard MIGS/MIMS compliant XML schema: Toward
635 the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS*
636 **12**, 115-121 (2008).
637 70 Benson, D. A. *et al.* GenBank. *Nucl. Acids Res.* **35**, D21-25 (2007).

638 71 Benson, D. A. *et al.* GenBank. *Nucl. Acids Res.* **36**, D25-30 (2008).

639 72 Hirschman, L. *et al.* Meeting report: Metagenomics, Metadata and Meta-analysis"
640 (M3) Workshop at the Pacific Symposium on Biocomputing 2010. *SIGS* **2**, 357-360
641 (2010).

642 73 Hankeln, W. *et al.* MetaBar - a tool for consistent contextual data acquisition and
643 standards compliant submission. *BMC Bioinformatics* **11**, 358 (2010).

644 74 Rocca-Serra, P. *et al.* ISA infrastructure: supporting standards-compliant
645 experimental reporting and enabling curation at the community level. *Bioinformatics* **26**,
646 2354-2356 (2010).

647

648

		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 (Sequence Read Archives), DRA (DDBJ Sequence 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 (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 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. Environment Ontology (EnvO) (v1.53) terms listed under environmental biome can be found from the link: 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 from the link: 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, prior to 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 from the link: http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483	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)	Sequencing method used; e.g. Sanger, pyrosequencing, ABI-solid.	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. MIENS-survey is applicable to contextual data for marker gene sequences, obtained directly from the environment, without culturing or identification of the organisms. MIENS-culture, on the other hand, applies to the contextual data for marker gene sequences from cultured or voucher-identifiable specimens. Both MIENS-survey and culture 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. '-' 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), or an ontology from the Open Biological and Biomedical Ontologies (OBO) foundry (http://www.obofoundry.org). 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.