

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

24 Ward<sup>19</sup>, George M. Weinstock<sup>54</sup>, Doug Wendel<sup>4</sup>, Owen White<sup>25</sup>, Andrew Whiteley<sup>3</sup>,  
25 Andreas Wilke<sup>10</sup>, Jennifer R. Wortman<sup>25</sup>, Tanya Yatsunenko<sup>36</sup>, Frank Oliver Glöckner<sup>1,2</sup>  
26  
27  
28 1 Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine  
29 Microbiology, D-28359 Bremen, Germany  
30 2 Jacobs University Bremen gGmbH, D-28759 Bremen, Germany  
31 3 Natural Environment Research Council Environmental Bioinformatics Centre,  
32 Wallington CEH, Oxford OX10 8BB, UK  
33 4 Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado  
34 80309, USA  
35 5 Howard Hughes Medical Institute, San Francisco, California 94143-2208, USA  
36 6 Ribosomal Database Project, Michigan State University, East Lansing, Michigan  
37 48824-4320, USA  
38 7 Center for Microbial Ecology, Michigan State University, East Lansing, Michigan  
39 48824-1325, USA  
40 8 The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution,  
41 Marine Biological Laboratory, Woods Hole, Massachusetts, USA  
42 9 Plymouth Marine Laboratory, Plymouth PL1 3DH, UK  
43 10 Mathematics and Computer Science Division, Argonne National Laboratory,  
44 Argonne, Illinois 60439, USA  
45 11 Department of Ecology and Evolution, University of Chicago, Chicago, Illinois  
46 60637, USA  
47 12 National Center for Biotechnology Information (NCBI), National Library of

48 Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA  
49 13 European Molecular Biology Laboratory (EMBL) Outstation, European  
50 Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge  
51 CB10 1SD, UK  
52 14 School of Civil and Environmental Engineering, Yonsei University, Seoul 120-749,  
53 Republic of Korea  
54 15 School of Computer Science, University of Manchester, Manchester M13 9PL, UK  
55 16 Oxford e-Research Centre, University of Oxford, Oxford OX1 3QG, UK  
56 17 Structural and Computational Biology Unit, European Molecular Biology Laboratory,  
57 D-69117 Heidelberg, Germany  
58 18 Department of Molecular, Cellular and Developmental Biology, University of  
59 Colorado, Boulder, Colorado 80309, USA  
60 19 Broad Institute of Massachusetts Institute of Technology and Harvard University,  
61 Cambridge, Massachusetts 02142, USA  
62 20 Department of Medicine and the Department of Microbiology, New York University  
63 Langone Medical Center, New York 10017, USA  
64 21 National Human Genome Research Institute, National Institutes of Health, Bethesda,  
65 Maryland 20892, USA  
66 22 Department of Microbiology, University of Pennsylvania School of Medicine,  
67 Philadelphia 19104, USA  
68 23 DOE Joint Genome Institute, Walnut Creek, California 94598, USA  
69 24 Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico,  
70 USA

71 25 Institute for Genome Sciences, University of Maryland School of Medicine,  
72 Baltimore, Maryland 21201, USA

73 26 Department of Applied Mathematics and Computer Science, Ghent University, 9000  
74 Ghent, Belgium

75 27 Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory,  
76 Berkeley, California, USA

77 28 Department of Ecology and Evolutionary Biology, University of Colorado, Boulder,  
78 Colorado 80309, USA

79 29 University of Colorado Museum of Natural History, University of Colorado, Boulder,  
80 Colorado 80309, USA

81 30 Department of Biological Sciences, University of Southern California, Los Angeles,  
82 California 90089, USA

83 31 National Ecological Observatory Network (NEON), Boulder, Colorado 80301, USA

84 32 Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas  
85 77030, USA

86 33 Department of Molecular and Human Genetics, Baylor College of Medicine, Houston,  
87 Texas 77030, USA

88 34 Department of Biology, University of New Mexico, LTER Network Office,  
89 Albuquerque, New Mexico 87131, USA

90 35 Department of Computer Science, University of Colorado, Boulder, Colorado 80309,  
91 USA

92 36 Center for Genome Sciences and Systems Biology, Washington University School of  
93 Medicine, St. Louis, Missouri 63108, USA

94 37 Department of Molecular Virology and Microbiology, Baylor College of Medicine,  
95 Houston, Texas 77030, USA

96 38 Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences,  
97 The University of Queensland, Brisbane QLD 4072, Australia

98 39 Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley,  
99 California, USA

100 40 Department of Biology, San Diego State University, San Diego, California 92182-  
101 4614, USA

102 41 Department of Microbiology, Cornell University, Ithaca, New York 14853, USA

103 42 Cooperative Institute for Research in Environmental Sciences, University of Colorado,  
104 Boulder, Colorado 80302, USA

105 43 Lehrstuhl für Mikrobiologie, Technische Universität München, D-853530 Freising,  
106 Germany

107 44 J. Craig Venter Institute, Rockville, Maryland 20850-3213, USA

108 45 Department of Environmental Sciences, University of Colorado, Boulder, Colorado  
109 80309, USA

110 46 Department of Biology, University of Waterloo, Ontario, N2L 3G1, Canada

111 47 Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge,  
112 Tennessee 37830-8020, USA

113 48 Ribocon GmbH, D-28359 Bremen, Germany

114 49 VIB - Vrije Universiteit Brussel, 1050 Brussels, Belgium

115 50 Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, University of  
116 Guelph, Guelph, Ontario N1G 2W1, Canada

117 51 Departments of Microbiology and Immunology and of Medicine, Stanford University  
118 School of Medicine, Stanford, California 94305, USA

119 52 Veterans Affairs Palo Alto Health Care System, Palo Alto, California 94304, USA

120 53 Department of Microbiology and Immunology, Ann Arbor, Michigan 48109-5620,  
121 USA

122 54 The Genome Center, Department of Genetics, Washington University in St. Louis  
123 School of Medicine, St. Louis, Missouri 63108, USA

124

125

126 **Here we present a standard developed by the Genomic Standards Consortium**  
127 **(GSC) to describe marker gene sequences—the minimum information about a**  
128 **marker gene sequence (MIMARKS). We also introduce a system for describing the**  
129 **environment from which a biological sample originates. The “environmental**  
130 **packages” apply to any sequence whose origin is known and can therefore be used**  
131 **in combination with MIMARKS or other GSC checklists. Finally, to establish a**  
132 **unified standard for describing sequence data and to provide a single point of entry**  
133 **for the scientific community to access and learn about GSC checklists, we establish**  
134 **the minimum information about any (x) sequence (MIxS). Adoption of MIxS will**  
135 **enhance our ability to analyze natural genetic diversity across the Tree of Life as it**  
136 **is currently being documented by massive DNA sequencing efforts from myriad**  
137 **ecosystems in our ever-changing biosphere.**

- 138 **Abbreviations**
- 139 CBOL: Consortium for the Barcode of Life
- 140 COI: cytochrome c oxidase I
- 141 DDBJ: DNA DataBank of Japan
- 142 DOI: Digital Object Identifier
- 143 DRA: DDBJ Sequence Read Archive
- 144 ENA: European Nucleotide Archive
- 145 EnvO: Environment Ontology
- 146 GAZ: Gazetteer
- 147 GCDML: Genomic Contextual Data Markup Language
- 148 GSC: Genomic Standards Consortium
- 149 ICoMM: International Census of Marine Microbes
- 150 INSDC: International Nucleotide Sequence Database Collaboration
- 151 ISA: Investigation/Study/Assay Infrastructure
- 152 ISO: International Organization for Standardization
- 153 MICROBIS: The Microbial Oceanic Biogeographic Information System
- 154 MIMARKS: Minimum Information about a MARKer Gene Sequence
- 155 MIGS/MIMS: Minimum Information about a Genome/Metagenome Sequence
- 156 MIRADA-LTERs: Microbial Inventory Research Across Diverse Aquatic Long Term  
157 Ecological Research Sites
- 158 OBO: Open Biological and Biomedical Ontologies
- 159 PMID: Pubmed ID
- 160 RDP: Ribosomal Database Project
- 161 *rRNA*: ribosomal RNA
- 162 SI: International System of Units
- 163 SRA: Sequence Read Archive

- 164 SSU: small subunit
- 165 URL: Uniform Resource Locator
- 166 WGS84: World Geodetic System 84
- 167 XML Schema: Extensible Markup Language Schema
- 168



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

183           The GSC has previously proposed standards for describing genomic sequences,  
184 the “minimum information about a genome sequence” (MIGS), and metagenomic  
185 sequences, the “minimum information about a metagenome sequence” (MIMS)<sup>1</sup>. Here  
186 we introduce an extension of these standards for capturing information about marker  
187 genes, MIMARKS. Additionally, we introduce “environmental packages” that  
188 standardize sets of measurements and observations describing particular habitats that are  
189 applicable across all GSC checklists and beyond<sup>2</sup>. We define “environment” as any  
190 location in which a sample or organism is found, e.g., soil, air, water, human-associated,  
191 plant-associated, or laboratory. The original MIGS/MIMS checklists included contextual

192 data about the location from which a sample was isolated and how the sequence data was  
193 produced. However, standard descriptions for a more comprehensive range of  
194 environmental parameters, which would help to better contextualize a sample, were not  
195 included. The environmental packages presented here are relevant to any genome  
196 sequence of known origin, and would usefully be combined with many projects described  
197 by MIGS, MIMS or MIMARKS.

198 To create a single entry point to all minimum information checklists from the  
199 GSC and to the environmental packages, we propose an overarching framework, the  
200 MIxS standard [\[AU: ADD URL\]](#). MIxS is a new standard that includes the technology-  
201 specific checklists from the previous MIGS and MIMS standards, provides a way of  
202 introducing additional checklists such as MIMARKS, and also allows annotation of  
203 sample data using environmental packages. A schematic overview of MIxS along with  
204 the MIxS environmental packages is shown in **Figure 1**.

205

## 206 **The development of MIMARKS and the environmental packages**

207 Over the past three decades, the 16S rRNA, 18S rRNA and internal transcribed  
208 spacer gene sequences (ITS) from *Bacteria*, *Archaea*, and microbial *Eukaryotes* have  
209 provided deep insights into the topology of the tree of life<sup>3, 4</sup> and the composition of  
210 communities of organisms that live in diverse environments, which range from deep sea  
211 hydrothermal vents to ice sheets in the Arctic<sup>5-16</sup>. Numerous other phylogenetic marker  
212 genes have also proven useful, including RNA polymerase subunits (*rpoB*), DNA gyrases  
213 (*gyrB*), DNA recombination and repair proteins (*recA*) and heat shock proteins (*HSP70*)<sup>3</sup>.  
214 Marker genes can also reveal key metabolic functions rather than phylogeny; examples

215 include nitrogen cycling (*amoA*, *nifH*, *ntcA*)<sup>17, 18</sup>, sulfate reduction (*dsrAB*)<sup>19</sup> or  
216 phosphorus metabolism (*phnA*, *phnI*, *phnJ*)<sup>20, 21</sup>. In this paper we collectively define all of  
217 these different phylogenetic and functional genes (or gene fragments) as “marker genes”  
218 as they are used to profile natural genetic diversity across the Tree of Life, and argue that  
219 a small amount of additional effort invested in describing them with specific guidelines in  
220 our public databases will revolutionize the study types that can be performed with these  
221 large data resources. This effort is timely, given the need to determine how climate  
222 change and various other anthropogenic perturbations of our biosphere are affecting  
223 biodiversity, and how marked changes in our cultural traditions and lifestyles are  
224 affecting human microbial ecology, and, ultimately, human health.

225 MIMARKS (**Table 1**) complements the MIGS/MIMS checklists for genomes and  
226 metagenomes by adding two new checklists, a MIMARKS-survey, for uncultured  
227 diversity marker gene surveys, and a MIMARKS-specimen, for marker gene sequences  
228 obtained from any material identifiable via specimens. The MIMARKS extension adopts  
229 and incorporates the standards being developed by the Consortium for the Barcode of  
230 Life (CBOL)  
231 ([http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG\\_data\\_standards-  
232 Final.pdf](http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG_data_standards-Final.pdf)). Therefore, the checklist can be universally applied to any marker gene, from  
233 SSU rRNA to COI, to all taxa, and to studies ranging from single individuals to complex  
234 communities.

235 Both MIMARKS and the environmental packages were developed by collating  
236 information from several sources and evaluating it in the framework of the existing  
237 MIGS/MIMS checklists. These include four independent community-led surveys,

238 examination of the parameters reported in published studies, and examination of  
239 compliance with optional features in INSDC documents. The overall goal of these  
240 activities was to design the backbone of the MIMARKS checklist, which describes the  
241 most important aspects of marker gene contextual data.

#### 242 *Results of community-led surveys*

243 To date, four online surveys about descriptors for marker genes have been conducted to  
244 determine researcher preferences for core descriptors. The Department of Energy Joint  
245 Genome Institute and SILVA<sup>22</sup> surveys focused on general descriptor contextual data for  
246 a marker gene, whereas the Ribosomal Database Project (RDP)<sup>23</sup> focused on prevalent  
247 habitats for rRNA gene surveys, and the Terragenome Consortium<sup>24</sup> focused on soil  
248 metagenome project contextual data (supplementary information 1). The above  
249 recommendations were joined by an extensive set of contextual data items suggested by  
250 an International Census of Marine Microbes (ICoMM) working group that met in 2005.  
251 These collective resources provided valuable insights into community requests for  
252 contextual data items to be included in the MIMARKS checklist and the main habitats  
253 constituting the environmental packages.

#### 254 *Survey of published parameters*

255 We reviewed published rRNA gene studies, retrieved via SILVA and the ICoMM  
256 database MICROBIS (The Microbial Oceanic Biogeographic Information System)  
257 (<http://icomm.mbl.edu/microbis>) to further supplement contextual data items that are  
258 included in the respective environmental packages. In total, 39 publications from SILVA  
259 and >40 ICoMM projects were scanned for contextual data items to constitute the core of  
260 the environmental package sub-tables (supplementary information 1).

261 *Survey of INSDC source feature qualifiers*

262 In a final analysis step, we surveyed usage statistics of INSDC source feature key  
263 qualifier values of rRNA gene sequences contained in SILVA (supplementary  
264 information 1). Notably, less than 10% of the 1.2 million 16S rRNA gene sequences  
265 (SILVA release 100) were associated with even basic information such as  
266 latitude/longitude, collection date or PCR primers.

267 *The MIMARKS checklist*

268 The MIMARKS checklist provides users with an “electronic laboratory notebook”  
269 containing core contextual data items required for consistent reporting of marker gene  
270 investigations. MIMARKS uses the MIGS/MIMS checklists with respect to the nucleic  
271 acid sequence source and sequencing contextual data, but extends them with further  
272 experimental contextual data such as PCR primers and conditions, or target gene name.  
273 For clarity and ease of use, all items within the MIMARKS checklist are presented with a  
274 value syntax description, as well as a clear definition of the item. Whenever terms from a  
275 specific ontology are required as the value of an item, these terms can be readily found in  
276 the respective ontology browsers linked by URLs in the item definition. Although this  
277 version of the MIMARKS checklist does not contain unit specifications, we recommend  
278 all units to be chosen from and follow the International System of Units (SI)  
279 recommendations. In addition, we strongly urge the community to provide feedback  
280 regarding the best unit recommendations for given parameters. To facilitate comparative  
281 studies, unit standardization across data sets will be vital in future. An Excel<sup>®</sup> version of  
282 the MIMARKS checklist is provided to the community on the GSC web site at:  
283 [http://gensc.org/gc\\_wiki/index.php/MIMARKS](http://gensc.org/gc_wiki/index.php/MIMARKS).

284 ***The MlXS environmental packages***

285 Fourteen environmental packages provide a wealth of environmental and epidemiological  
286 contextual data fields for a complete description of sampling environments. Furthermore,  
287 the environmental packages can be combined with any of the GSC checklists (figure 1  
288 and supplementary information 2). Researchers within The Human Microbiome Project<sup>25</sup>  
289 contributed the host-associated and all human packages. The Terragenome Consortium  
290 contributed sediment and soil packages. Finally, ICoMM, Microbial Inventory Research  
291 Across Diverse Aquatic Long Term Ecological Research Sites (MIRADA-LTERs), and  
292 the Max Planck Institute for Marine Microbiology contributed the water package. The  
293 MIMARKS working group developed the remaining packages (air, microbial  
294 mat/biofilm, miscellaneous natural or artificial environment, plant-associated, and  
295 wastewater/sludge). The package names describe high-level habitat terms in order to be  
296 exhaustive. The miscellaneous natural or artificial environment package contains a  
297 generic set of parameters, and is included for any other habitat that does not fall into the  
298 other thirteen categories. Whenever needed, multiple packages may be used for the  
299 description of the environment.

300 ***Examples of MIMARKS-compliant datasets***

301 Several MIMARKS-compliant reports are included in Supplementary Information 3.  
302 These include a 16S rRNA gene survey from samples obtained in the North Atlantic, a  
303 18S pyrosequencing tag study of anaerobic protists in a permanently anoxic basin of the  
304 North Sea, a *pmoA* survey from Negev Desert soils, a *dsrAB* survey of Gulf of Mexico  
305 sediments, and a 16S pyrosequencing tag study of bacterial diversity in the Western  
306 English Channel (accessible via SRA study accession number SRP001108).

307 **Adoption by major database and informatics resources**

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

313 In the past, the INSDC has issued a reserved “BARCODE” keyword for the  
314 CBOL<sup>26</sup>. Following this model, the INSDC has recently recognized the GSC as an  
315 authority for the MIxS standard and issued it with official keywords within INSDC  
316 nucleotide sequence records<sup>27</sup>. This greatly facilitates automatic validation of the  
317 submitted contextual data and provides support for datasets compliant with previous  
318 versions by including the checklist version as a keyword.

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

330           The next-generation Sequence Read Archive (SRA) collects and displays MIxS-  
331 compliant metadata in sample and experiment objects. There are several tools that are  
332 already available or under development to assist users in SRA submissions. The myRDP  
333 SRA PrepKit allows users to prepare and edit their submissions of reads generated from  
334 ultra-high-throughput sequencing technologies. A set of suggested attributes in the data  
335 forms assist researchers in providing metadata conforming to checklists such as  
336 MIMARKS. The Quantitative Insights Into Microbial Ecology ("QIIME") web  
337 application (<http://www.microbio.me/qiime>) allows users to generate and validate  
338 MIMARKS-compliant templates. These templates can be viewed and completed in the  
339 users' spreadsheet editor of choice (e.g. Microsoft Excel<sup>®</sup>). The QIIME web-platform also  
340 offers an ontology lookup and geo-referencing tool to aid users when completing the  
341 MIMARKS templates. The Investigation/Study/Assay (ISA) is a software suite that  
342 assists in the curation, reporting, and local management of experimental metadata from  
343 studies employing one or a combination of technologies, including high-throughput  
344 sequencing<sup>29</sup>. Specific ISA configurations (available from <http://isa-tools.org/tools.html>)  
345 have been developed to ensure MIxS compliance by providing templates and validation  
346 capability. Another tool, ISAconverter, produces SRA.xml documents, facilitating  
347 submission to the SRA repository.

348           Further detailed guidance for submission processes can be found under the  
349 respective wiki pages ([http://gensc.org/gc\\_wiki/index.php/MIGS/MIMS/MIMARKS](http://gensc.org/gc_wiki/index.php/MIGS/MIMS/MIMARKS)) of  
350 the standard.

### 351 **Maintenance of the MIxS standard**

352 To allow further developments, extensions, and enhancements of MIxS, we set up a



353 public issue tracking system to track changes and accomplish feature requests  
354 (<http://mixs.genc.org/>). New versions will be released annually. Technically, the MIxS  
355 standard, including MIMARKS and the environmental packages, is maintained in a  
356 relational database system at the Max Planck Institute for Marine Microbiology Bremen  
357 on behalf of the GSC. This provides a secure and stable mechanism for updating the  
358 checklist suite and versioning. In future, we plan to develop programmatic access to this  
359 database in order to allow automatic retrieval of the latest version of each checklist for  
360 INSDC databases and for GSC community resources. Moreover, the Genomic Contextual  
361 Data Markup Language (GCDML) is a reference implementation of the GSC checklists  
362 by the GSC and now implements the full range of MIxS standards. It is based on XML  
363 Schema technology and thus serves as an interoperable data exchange format for Web  
364 Service based infrastructures<sup>30</sup>.

365

## 366 **Conclusions and call for action**

367 The GSC is an international body with a stated mission of working towards richer  
368 descriptions of the complete collection of genomes and metagenomes through the MIxS  
369 standard. The present report extends the scope of GSC guidelines to marker gene  
370 sequences and environmental packages and establishes a single portal where  
371 experimentalists can gain access to and learn how to use GSC guidelines. The GSC is an  
372 open initiative that welcomes the participation of the wider community. This includes an  
373 open call to contribute to refinements of the MIxS standards and their implementations.  
374 The adoption of the GSC standards by major data providers and organizations, as well as  
375 the INSDC, underlines and seconds the efforts to contextually enrich our sequence data

376 collection, and complements the recent efforts to enrich other (meta) omics data. The  
377 MIxS standard, including MIMARKS, has been developed to the point that it is ready for  
378 use in the publication of sequences. A defined procedure for requesting new features and  
379 stable release cycles will facilitate implementation of the standard across the community.  
380 Compliance among authors, adoption by journals and use by informatics resources will  
381 vastly improve our collective ability to mine and integrate invaluable sequence data  
382 collections for knowledge- and application-driven research. In particular, the ability to  
383 combine microbial community samples collected from any source, using the universal  
384 Tree of Life as a measure to compare even the most diverse communities, should provide  
385 new insights into the dynamic spatiotemporal distribution of microbial life on our planet  
386 and in/on the human body.

387

388 **Figure Legend**

389 **Figure 1:** Schematic overview about the GSC MIxS standard (brown), including  
390 combination with specific environmental packages (blue). Shared descriptors apply to all  
391 MIxS checklists, however each checklist has its own specific descriptors as well.  
392 Environmental packages can be applied to any of the checklists. (EU: *Eukarya*, BA:  
393 *Bacteria/Archaea*, PL: Plasmid, VI: Virus, ORG: Organelle).

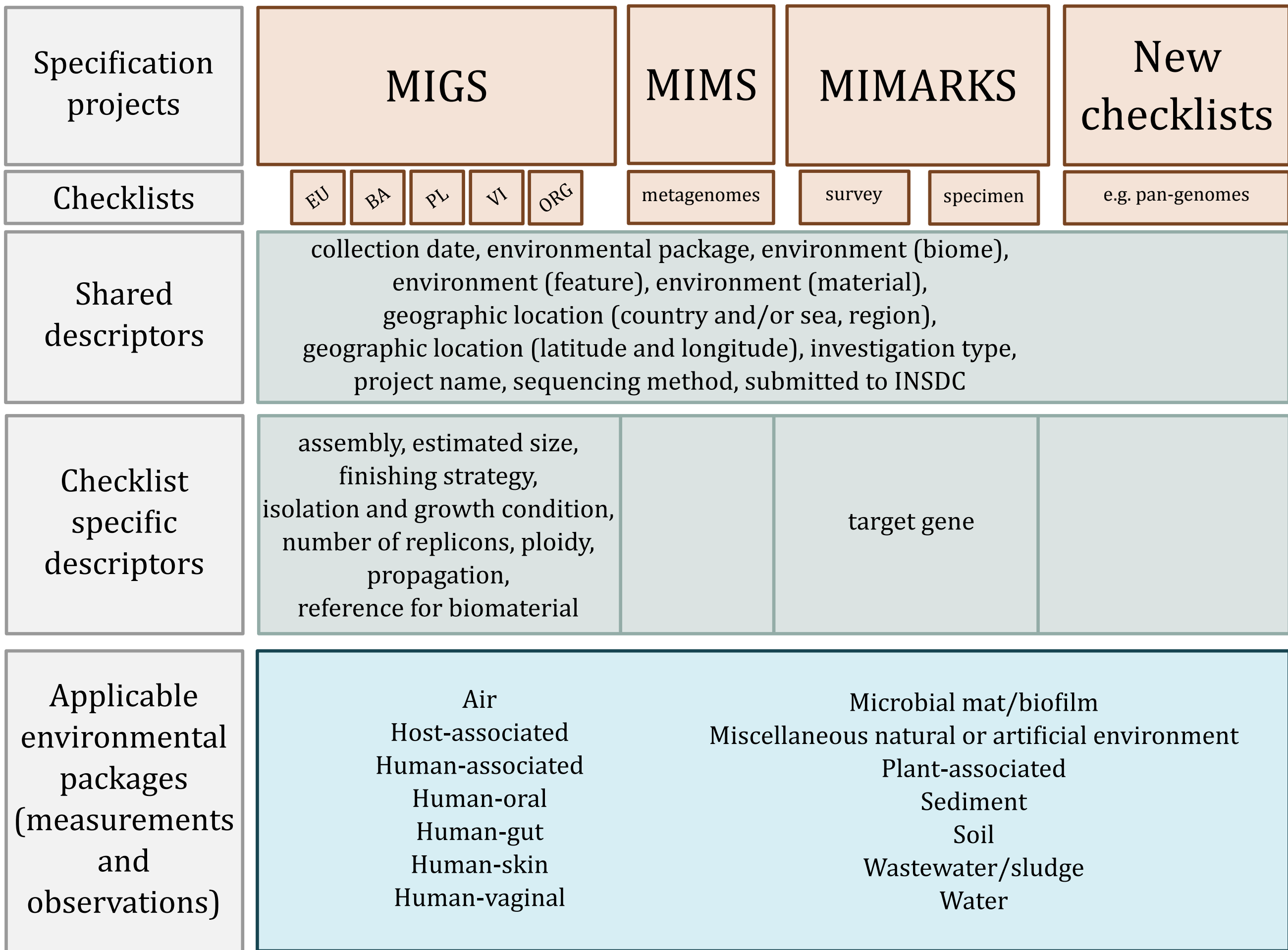
394

- 395 1. Field, D. *et al.* The minimum information about a genome sequence (MIGS)  
396 specification. *Nat. Biotechnol.* **26**, 541-547 (2008).
- 397 2. Taylor, C.F. *et al.* Promoting coherent minimum reporting guidelines for  
398 biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**,  
399 889-896 (2008).
- 400 3. Ludwig, W. & Schleifer, K.H. in *Microbial phylogeny and evolution, concepts*  
401 *and controversies*. (ed. J. Sapp) 70-98 (Oxford university press, New York, USA,  
402 2005).
- 403 4. Ludwig, W. *et al.* Bacterial phylogeny based on comparative sequence analysis.  
404 *Electrophoresis* **19**, 554-568 (1998).
- 405 5. Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. Genetic diversity in  
406 Sargasso Sea bacterioplankton. *Nature* **345**, 60-63 (1990).
- 407 6. Stahl, D.A. Analysis of hydrothermal vent associated symbionts by ribosomal  
408 RNA sequences. *Science* **224**, 409-411 (1984).
- 409 7. Ward, D.M., Weller, R. & Bateson, M.M. 16S rRNA sequences reveal numerous  
410 uncultured microorganisms in a natural community. *Nature* **345**, 63-65 (1990).

- 411 8. DeLong, E.F. Archaea in coastal marine environments. *Proc. Nat. Acad. Sci. USA*  
412 **89**, 5685-5689 (1992).
- 413 9. Diez, B., Pedros-Alio, C. & Massana, R. Study of genetic diversity of eukaryotic  
414 picoplankton in different oceanic regions by small-subunit rRNA gene cloning  
415 and sequencing. *Appl. Environ. Microbiol.* **67**, 2932-2941 (2001).
- 416 10. Fuhrman, J.A., McCallum, K. & Davis, A.A. Novel major archaeobacterial group  
417 from marine plankton. *Nature* **356**, 148-149 (1992).
- 418 11. Hewson, I. & Fuhrman, J.A. Richness and diversity of bacterioplankton species  
419 along an estuarine gradient in Moreton Bay, Australia. *Appl. Environ. Microbiol.*  
420 **70**, 3425-3433 (2004).
- 421 12. Huber, J.A., Butterfield, D.A. & Baross, J.A. Temporal changes in archaeal  
422 diversity and chemistry in a mid-ocean ridge subseafloor habitat. *Appl. Environ.*  
423 *Microbiol.* **68**, 1585-1594 (2002).
- 424 13. Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C. & Moreira, D.  
425 Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*  
426 **409**, 603-607 (2001).
- 427 14. Moon-van der Staay, S.Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA  
428 sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*  
429 **409**, 607-610 (2001).
- 430 15. Pace, N.R. A molecular view of microbial diversity and the biosphere. *Science*  
431 **276**, 734-740 (1997).
- 432 16. Rappe, M.S. & Giovannoni, S.J. The uncultured microbial majority. *Annu. Rev.*  
433 *Microbiol.* **57**, 369-394 (2003).

- 434 17. Francis, C.A., Beman, J.M. & Kuypers, M.M.M. New processes and players in  
435 the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia  
436 oxidation. *ISME. J.* **1**, 19-27 (2007).
- 437 18. Zehr, J.P., Mellon, M.T. & Zani, S. New nitrogen-fixing microorganisms detected  
438 in oligotrophic oceans by amplification of nitrogenase (*nifH*) genes. *Appl.*  
439 *Environ. Microbiol.* **64**, 3444-3450 (1998).
- 440 19. Minz, D. *et al.* Diversity of sulfate-reducing bacteria in oxic and anoxic regions of  
441 a microbial mat characterized by comparative analysis of dissimilatory sulfite  
442 reductase genes. *Appl. Environ. Microbiol.* **65**, 4666-4671 (1999).
- 443 20. Gilbert, J., A. *et al.* The seasonal structure of microbial communities in the  
444 Western English Channel. *Environ. Microbiol.* **11**, 3132-3139 (2009).
- 445 21. Martinez, A., W. Tyson, G. & DeLong, E., F. Widespread known and novel  
446 phosphonate utilization pathways in marine bacteria revealed by functional  
447 screening and metagenomic analyses. *Environ. Microbiol.* **12**, 222-238 (2009).
- 448 22. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked  
449 and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids*  
450 *Res.* **35**, 7188-7196 (2007).
- 451 23. Cole, J.R. *et al.* The Ribosomal Database Project: improved alignments and new  
452 tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141-145 (2009).
- 453 24. Vogel, T.M. *et al.* TerraGenome: a consortium for the sequencing of a soil  
454 metagenome. *Nat. Rev. Microbiol.* **7**, 252-252 (2009).
- 455 25. Turnbaugh, P.J. *et al.* The Human Microbiome Project. *Nature* **449**, 804-810  
456 (2007).

- 457 26. Benson, D.A. *et al.* GenBank. *Nucl. Acids Res.* **36**, D25-30 (2008).
- 458 27. Hirschman, L. *et al.* Meeting report: Metagenomics, Metadata and Meta-analysis”  
459 (M3) Workshop at the Pacific Symposium on Biocomputing 2010. *SIGS 2*, 357-  
460 360 (2010).
- 461 28. Hankeln, W. *et al.* MetaBar - a tool for consistent contextual data acquisition and  
462 standards compliant submission. *BMC Bioinformatics* **11**, 358 (2010).
- 463 29. Rocca-Serra, P. *et al.* ISA infrastructure: supporting standards-compliant  
464 experimental reporting and enabling curation at the community level.  
465 *Bioinformatics* **26**, 2354-2356 (2010).
- 466 30. Kottmann, R. *et al.* A standard MIGS/MIMS compliant XML schema: Toward  
467 the development of the Genomic Contextual Data Markup Language (GCDML).  
468 *OMICS* **12**, 115-121 (2008).
- 469
- 470



		Report type	
		MIMARKS-survey	MIMARKS-specimen
<b>Investigation</b>			
Submitted to INSDC <sup>[boolean]</sup>	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 <sup>[mimarks-survey or mimarks-specimen]</sup>	Nucleic Acid Sequence Report is the root element of all MIMARKS compliant reports as standardized by Genomic Standards Consortium (GSC). This field is either MIMARKS survey or MIMARKS specimen	M	M
Project name	Name of the project within which the sequencing was organized	M	M
<b>Environment</b>			
Geographic location (latitude and longitude <sup>[float, point, transect and region]</sup> )	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 <sup>[integer, point, interval, unit]</sup> )	Please refer to the definitions of depth in the environmental packages	E	E
Geographic location (elevation of site <sup>[integer, unit]</sup> ; altitude of sample <sup>[integer, unit]</sup> )	Please refer to the definitions of either altitude or elevation in the environmental packages	E	E
Geographic location (country and/or sea <sup>[INSDC or GAZ]</sup> ; region <sup>[GAZ]</sup> )	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 ( <a href="http://insdc.org/country.html">http://insdc.org/country.html</a> ), or the GAZ (Gazetteer, v1.446) ontology ( <a href="http://bioportal.bioontology.org/visualize/40651">http://bioportal.bioontology.org/visualize/40651</a> )	M	M
Collection date <sup>[ISO8601]</sup>	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 <sup>[EnvO]</sup> )	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: <a href="http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A0000428">http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A0000428</a>	M	M
Environment (feature <sup>[EnvO]</sup> )	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: <a href="http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00002297">http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00002297</a>	M	M
Environment (material <sup>[EnvO]</sup> )	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: <a href="http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483">http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483</a>	M	M
<b>MIGS/MIMS/MIMARKS Extension</b>			
Environmental package [air, host-associated, human-associated, human-skin, human-oral, human-gut, human-vaginal, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated, sediment, soil, wastewater/sludge, water]	MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained. All environmental packages listed here are further defined in separate subtables. By giving the name of the environmental package, a selection of fields can be made from the subtables and can be reported	M	M
<b>Nucleic acid sequence source</b>			
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
<b>Sequencing</b>			
Target gene or locus (e.g. 16S rRNA, 18S rRNA, nif, amoA, rpo)	Targeted gene or locus name for marker gene study	M	M
Sequencing method (e.g. dideoxysequencing, pyrosequencing, polony)	Sequencing method used; e.g. Sanger, pyrosequencing, ABI-solid.	M	M

Table 1. Items for the MIMARKS specification and their mandatory (M), conditionally mandatory (C) (the item is mandatory only when applicable to the study) or recommended (X) status for both MIMARKS-survey and MIMARKS-specimen checklists. Furthermore, “-” denotes that an item is not applicable for a given checklist. “E” denotes that a field has environment-specific requirements. For example, while “depth” is mandatory for environments water, sediment or soil; it is optional for human-associated environments. **MIMARKS-survey** is applicable to contextual data for marker gene sequences, obtained directly from the environment, without culturing or identification of the organisms. **MIMARKS-specimen**, on the other hand, applies to the contextual data for marker gene sequences from cultured or voucher-identifiable specimens. Both MIMARKS-survey and specimen checklists can be used for any type of marker gene sequence data, ranging from 16S, 18S, 23S, 28S rRNA to COI, hence the checklists are universal for all three domains of life.

Item names are followed by a short description of the value of the item in parentheses and/or value type in brackets as a superscript. Whenever applicable, value types are chosen from a controlled vocabulary (CV), or an ontology from the Open Biological and Biomedical Ontologies (OBO) foundry (<http://www.obofoundry.org>). This table only presents the very core of MIMARKS checklists, i.e. only mandatory items for each checklist. Supplementary information 2 in spreadsheet format contains all MIMARKS items, the tables for environmental packages in the MIGS/MIMS/MIMARKS extension, and GenBank structured comment name that should be used for submitting MIMARKS data to GenBank. In case of submitting to EBI/ENA the full names can be used.