Minimum Information about a Biosynthetic Gene cluster.

Marnix H Medema, Renzo Kottmann, Yi Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene de Bruijn, Yi Chooi, Jan Claesen, R Cameron Coates, et al.

OPEN

Minimum Information about a Biosynthetic Gene cluster

A wide variety of enzymatic pathways that produce specialized metabolites in bacteria, fungi and plants are known to be encoded in biosynthetic gene clusters. Information about these clusters, pathways and metabolites is currently dispersed throughout the literature, making it difficult to exploit. To facilitate consistent and systematic deposition and retrieval of data on biosynthetic gene clusters, we propose the Minimum Information about a Biosynthetic Gene cluster (MIBiG) data standard.

Living organisms produce a range of secondary metabolites with exotic chemical structures and diverse metabolic origins. Many of these secondary metabolites find use as natural products in medicine, agriculture and manufacturing. Research on natural product biosynthesis is undergoing an extensive transformation, driven by technological developments in genomics, bioinformatics, analytical chemistry and synthetic biology. It has now become possible to computationally identify thousands of biosynthetic gene clusters (BGCs) in genome sequences, and to systematically explore and prioritize them for experimental characterization[1,2]. A BGC can be defined as a physically clustered group of two or more genes in a particular genome that together encode a biosynthetic pathway for the production of a specialized metabolite (including its chemical variants). It is becoming possible to carry out initial experimental characterization of hundreds of such natural products, using high-throughput approaches powered by rapid developments in mass spectrometry[3–5] and chemical structure elucidation[6]. At the same time, single-cell sequencing and metagenomics are opening up access to new and uncharted branches of the tree of life[7–9], enabling scientists to tap into a previously undiscovered wealth of BGCs. Furthermore, synthetic biology allows the redesign of BGCs for effective heterologous expression in preengineered hosts, which will ultimately empower the construction of standardized high-throughput platforms for natural product discovery[10,11].

In this changing research environment, there is an increasing need to access all the experimental and contextual data on characterized BGCs for comparative analysis, for function prediction and for collecting building blocks for the design of novel biosynthetic pathways. For this purpose, it is paramount that this information be available in a standardized and systematic format, accessible in the same intuitive way as, for example, genome annotations or protein structures. Currently, the situation is far from ideal, with information on natural product biosynthetic pathways scattered across hundreds of scientific articles in a wide variety of journals; it requires in-depth reading of papers to confidently discern which of the molecular functions associated with a gene cluster or pathway have been experimentally verified and which have been predicted solely on the basis of biosynthetic logic or bioinformatic algorithms. Although some valuable existing manually curated databases have data models in place to store some of this information[12–14], all are specialized towards certain subcategories of BGCs and include just a limited number of parameters defined by the interests of a subset of the scientific community. To enable the future development of databases with universal value, a generally applicable community standard is required that specifies the exact annotation and metadata parameters agreed upon by a wide range of scientists, as well as the possible types of evidence that are associated with each variable in publications and/or patents. Such a standard will be of great value for the consistent storage of data and will thus alleviate the tedious process of manually gathering information on BGCs. Moreover, a comprehensive data standard will allow future data infrastructures to enable the integration of multiple types of data, which will generate new insights that would otherwise not be attainable.

The Genomic Standards Consortium (GSC)[15] (**Box 1**) previously developed the Minimum Information about any Sequence (MIxS) framework[16]. This extensible 'minimum information' standardization framework includes the Minimum Information about a Genome Sequence (MIGS)[17] and the Minimum Information about a MARKer gene Sequence (MIMARKS)[16] standards. MIxS is a flexible framework that can be expanded upon to serve a wide variety of purposes. The GSC facilitates the community effort of maintaining and extending MIxS, and stimulates compliance among the community.

Here, we introduce the "Minimal Information about a Biosynthetic Gene cluster" (MIBiG) specification as a coherent extension of the GSC's MIxS standards framework. MIBiG provides a comprehensive and standardized specification of BGC annotations and gene cluster–associated metadata that will allow their systematic deposition in databases. Through a community annotation of BGCs that have been experimentally characterized and described in the literature during previous decades, we have constructed an MIBiG-compliant seed dataset. Moreover, a large part of the research community has committed to continue submitting data on newly characterized gene clusters in the MIBiG format in the future. Together, the MIBiG standard and the resulting MIBiG-compliant data sets will allow data infrastructures to be developed that will facilitate key future developments in natural product research.

Design of the MIBiG standard

The MIBiG standard covers general parameters that are applicable to each and every gene cluster as well as compound type–specific parameters that apply only to specific classes of pathways (**Fig. 1**). Notably, the standard has been designed to be suitable for

A full list of authors and affiliations appears at the end of the paper.

Box 1 | The Genomic Standards Consortium and its MIxS framework

The Genomic Standards Consortium (GSC, http://gensc.org/) was founded in 2005 as an open-membership working body with the purpose of promoting the standardization of genome descriptions as well as the exchange and integration of genomic data.

The GSC initiates and coordinates the design of and compliance with several minimum information standards (also known as checklists). An overarching framework has been designed to connect and standardize these checklists themselves: the Minimum Information about any (x) Sequence (MIxS)[16]. MIxS consists of three layers.

First, the MIxS standard includes a number of shared descriptors that are relevant to all types of nucleotide sequences, such as collection date, environmental origin, geographical coordinates of the location of origin of the sample and sequencing method.

Second, for a wide range of different environmental origins, so-called 'environmental packages' are available that constitute checklists of measurements and observations that are specific to each environment: for example, for host-associated microbial DNA samples, the taxonomy of the host, the habitat of the host and several phenotypic characteristics of the host can be collected. In this manner, rich contextual information on the context of each microbial sample is stored.

Third, several checklists are available for specific sequence types, each having their own checklist-specific descriptors. Previous checklists include the Minimum Information about a Genome Sequence (MIGS)[17], Minimum Information about a Metagenome Sequence (MIMS)[17] and Minimum Information about a MARKer gene Sequence (MIMARKS)[16].

In spring 2013, the MIBiG project proposal was accepted by the board of the GSC to form a new checklist within the MIxS framework. Besides a number of general descriptors, it also includes pathway type–specific packages that function analogously, for different classes of biosynthetic pathways, to the way the MIxS environmental packages do for different environmental origins.

biosynthetic pathways from any taxonomic origin, including those from bacteria, archaea, fungi and plants.

The general parameters cover important data items that are universally applicable. First, they include identifiers of the publications associated with the characterization of the gene cluster, so that the full description of the experimental results that support the entire entry can be accessed easily.

The second key group of general parameters describes the associated genomic locus (or loci) and its accession numbers and coordinates, as deposited in or submitted to one of the databases of the International Nucleotide Sequence Database Collaboration (INSDC): the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (EBI-ENA) or GenBank, all of which share unified accession numbers. The INSDC accession numbers are also used to link each MIBiG entry (which is given a separate MIBiG accession number) and its annotations to the corresponding nucleotide sequence(s) computationally; hence, a GenBank/ENA/DDBJ submission of the underlying nucleotide sequence is always required to file a MIBiG submission.

The third group of general parameters describes the chemical compounds produced from the encoded pathway, including their structures, molecular masses, biological activities and molecular targets. Additionally, these parameters allow documentation of miscellaneous chemical moieties that are connected to the core scaffold of the molecule (but synthesized independently) and the genes associated with their biosynthesis; this will facilitate the design of tools for the straightforward comparison of such 'sub-clusters', which are frequently present in different variants across multiple parent BGCs.

Finally, there is a group of general parameters describing experimental data on genes and operons in a gene cluster, including gene knockout phenotypes, experimentally verified gene functions and operons verified by techniques such as RNA-seq.

Beside the general parameters, the MIBiG standard contains dedicated class-specific checklists for gene clusters encoding pathways to produce polyketides, nonribosomal peptides (NRPs), ribosomally synthesized and post-translationally modified peptides (RiPPs), terpenes, saccharides and alkaloids. These include items such as acyltransferase domain substrate specificities and starter units for polyketide BGCs, release/cyclization types and adenylation domain substrate specificities for NRP BGCs, precursor peptides and peptide modifications for RiPP BGCs, and glycosyltransferase specificities for saccharide BGCs. Where applicable, the standard was made compliant with earlier community agreements, such as the recently published classification of RiPPs[18]. Hybrid BGCs that cover multiple

genomic STANDARDS consortium — MIBiG

General parameters

Biosynthetic class(es)*	Functional category of each gene	Chemical database identifier(s)
Key publications*	Tailoring enzymes/reactions*	Compound structure(s)*
Number of loci*	Knockout mutant phenotypes	Compound molecular formula(e)
Complete/partial cluster*	Operon architecture	Exact molecular mass(es)
Nucleotide accession + coordinates*	Functional sub-clusters	Compound activity*
MIxS environmental metadata	Compound name(s)*	Compound molecular target
Custom gene names	Synonyms for compound name(s)	Chemical moieties

Compound type–specific parameters

Polyketides

Polyketide subclass*
Polyketide synthase type*
Linear/cyclic*
Starter unit*
Polyketide length**
PKS genes*
Number of iterations (if iterative)**
Iterative PKS subtype + cyclization type**
Trans-acyltransferase genes
Cyclases/aromatases*
Thioesterase type (+ thioesterases)**
Release/cyclization type*
Modular PKS architecture**
Module skipping/iteration**
Reductive domains**
Non-reductive modifying PKS domains**
Ketoreductase stereochemistries
AT/CAL domain substrate specificities**

Nonribosomal peptides

NRP subclass*
Linear/cyclic*
NRPS genes*
Thioesterase type (+ thioesterases)*
Release/cyclization type*
Modular NRPS architecture
Module skipping/iteration**
Scaffold-modifying NRPS domains
Condensation domain subtypes
A domain substrate specificities
Epimerization of amino acid

Terpenoids

Terpene subclass*
Terpenoid scaffold size*
Final isoprenoid precursor*
Terpene synthases/cyclases
Prenyltransferases

RiPPs

RiPP subclass*
Linear/cyclic*
Precursor-encoding gene(s)*
Core peptide sequence
Leader peptide length
Follower peptide length
Cleavage recognition site sequence
Recognition motif in leader peptide
Peptidases involved in cleavage
Crosslinks and crosslink types

Saccharides

Saccharide subclass*
Glycosyltransferase (GT) genes
GT substrate specificities
Sub-clusters for sugar biosynthesis

Alkaloids

Alkaloid subclass*

Figure 1 | Schematic overview of the MIBiG standard. The MIBiG standard is composed of general and compound class–specific parameters. Wherever relevant, evidence coding is used to indicate the experimental support for items in the checklist. Fields annotated with an asterisk are absolutely mandatory; fields with two asterisks are conditionally mandatory.

Figure 2 | An example MIBiG entry, describing the relatively simple hybrid NRPS-PKS biosynthetic gene cluster for isoflavipucine/dihydroisoflavipucine from *Aspergillus terreus*. Fields without information have been omitted, and some JSON field abbreviations have been modified for clarity. The full entry is available from http://mibig.secondarymetabolites.org/repository/BGC0001122/BGC0001122.json.

biochemical classes can be described by simply entering information on each of the constituent compound types: the checklists have been designed in such a way that this does not lead to conflicts. Importantly, the modularity of the checklist system allows for the straightforward addition of further class-specific checklists when new types of molecules are discovered in the future.

The combination of general and compound-specific MIBiG parameters, together with the MIxS checklist, provides a complete description of the chemical, genomic and environmental dimensions that characterize a biosynthetic pathway (**Fig. 2**). A minimal set of key parameters is mandatory, while other parameters are optional. For many parameters, a specific ontology has been designed in order to standardize the inputs and to make it easier to categorize and search the resulting data.

Whenever possible, parameters are linked to a system of evidence attribution that specifies the kinds of experiments performed to arrive at the conclusions indicated by the chosen parameter values. Hence, each annotation entered during submission is assigned a specific evidence code: for example, when annotating the substrate specificity of a nonribosomal peptide synthetase (NRPS) adenylation domain, the submitter can choose between 'activity assay', 'structure-based inference' and 'sequence-based prediction' as evidence categories to support a given specificity.

During the design of the standard, great care was taken to make it compatible with unusual biosynthetic pathways, such as branched or module-skipping polyketide synthase (PKS) and NRPS assembly lines. Also, to ensure that the standard is compliant with the current state of the art in the various subfields of natural product research, we conducted an online community survey at an early stage of standard development (see **Supplementary Data Set 1**). Feedback was provided by 61 principal investigators from 16 different countries (most of whom also coauthored this paper), including at least ten leading experts for each major class of biosynthetic pathways covered.

Addressing key research needs

Adoption of the MIBiG standard will allow the straightforward collation of all annotations and experimental data on each BGC, which would otherwise be dispersed across multiple scientific articles and resources. Moreover, there are at least three additional key ways in which MIBiG will facilitate new scientific and technological developments: it will enable researchers to systematically connect genes to chemistry (and vice versa), to better understand secondary metabolite biosynthesis and the compounds produced in their ecological and environmental context, and to effectively use synthetic biology to engineer newly designed BGC configurations underpinned by an evidence-based parts registry (**Fig. 3**).

First, the comprehensive dataset generated through MIBiG-compliant submissions will enable researchers to systematically connect genes and chemistry. Not only will it allow individual researchers to predict enzyme functions by comparing enzyme-coding genes in newly identified BGCs to a thoroughly documented dataset, it will also facilitate general advances in chemistry predictions. Substrate specificities of PKS acyltransferase domains and NRPS adenylation domains, as well as their evidence codes, will be registered automatically for all gene clusters. This will enable automated updating of the training sets for key chemistry prediction algorithms[19–21], which can then be curated by the degree of evidence available, increasing the accuracy of predictions of core peptide and polyketide scaffolds. Also, because groups of genes associated with the biosynthesis of specific chemical moieties (such as sugars and nonproteinogenic amino acids) will be registered consistently, a continuously growing dataset of such sub-clusters will be available to use as a basis for chemical structure predictions.

In addition, MIBiG has the potential to greatly enhance the understanding of secondary metabolite biosynthesis in its ecological and environmental context: the connection of MIBiG to the MIxS standard should stimulate researchers to supply MIxS data on the genome and metagenome sequences that contain the BGCs. This will generate opportunities for a range of analyses, such as the biogeographical mapping of secondary metabolite biosynthesis[22], thereby identifying locations and ecosystems harboring rich biosynthetic diversity. But even if the contextual data associated with the genome sequences cannot always be made MIxS compliant (perhaps because the origin of a strain can no longer be traced), the MIBiG standard itself provides a comprehensive reference dataset for annotating large-scale MIxS-compliant metagenomic data from projects such as the Earth Microbiome Project[23], *Tara* Oceans[24] and Ocean Sampling Day[25]. This will enable scientists to obtain a better understanding of the distribution of BGCs in the environment. Altogether, the standard will play a significant role in guiding sampling efforts for future natural product discovery.

Finally, the data resulting from MIBiG-compliant submissions will provide an evidence-based parts registry for the engineering of biosynthetic pathways. Synthetic biologists need a toolbox containing genetic parts that have been experimentally characterized. The MIBiG standard, through its systematic annotation of gene function by evidence coding, knockout mutant phenotypes and substrate specificities, will streamline the identification of all available candidate genes and proteins available to perform a desired function, together with the pathway context in which they natively occur. In this manner, it will provide a comprehensive catalog of parts that can be used for the modification of existing biosynthetic pathways or the *de novo* design of new pathways.

Community annotation effort

To accelerate the usefulness of new MIBiG-compliant data submissions, we initiated this project by annotating a significant portion of the experimental data on the hundreds of BGCs that have been characterized in recent decades. The resulting data will allow immediate contextualization of new submissions (see below) and comparative

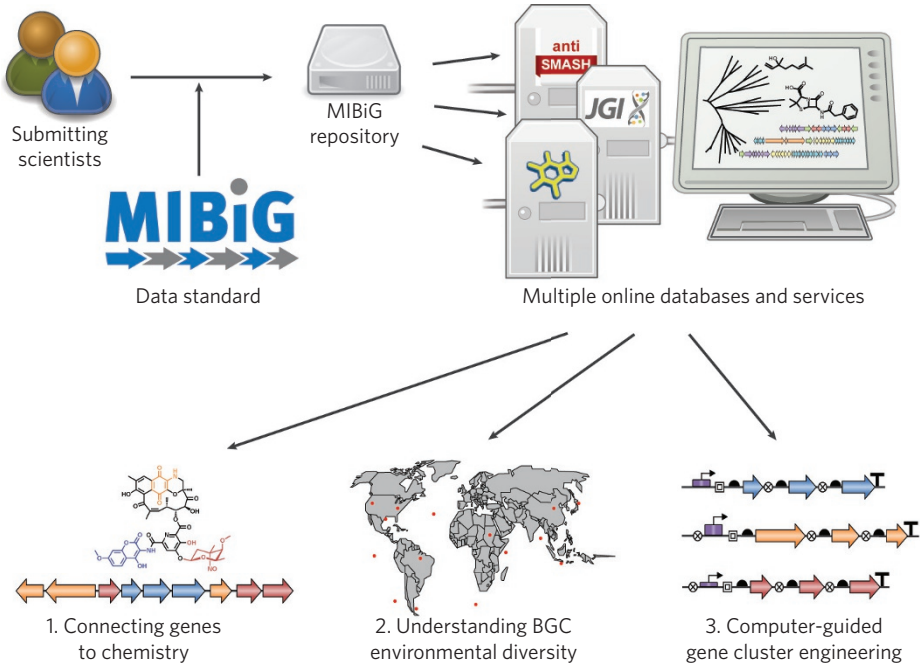

Figure 3 | The MIBiG data standard and submission system will lead to a continuously growing dataset (stored in the online MIBiG repository) that will be loaded into several databases and web services. The lower part of the figure shows the threefold potential of MIBiG for the study of BGCs, which will make it possible to (1) systematically connect genes and chemistry by identifying which genes are responsible for the biosynthesis of which chemical moieties; (2) understand the natural genetic diversity of BGCs within their environmental and ecological context, by combining MIBiG- and MIxS-derived metadata sets; and (3) develop an evidence-based parts registry for engineering biosynthetic pathways and gene clusters through synthetic biology.

analysis of any newly characterized BGCs with a rich source of MIBiG-compliant data. Moreover, this annotation effort offered an ideal opportunity to evaluate the MIBiG standard in practice on a diverse range of BGCs. Hence, we carefully mined the literature to obtain a set of 1,170 experimentally characterized gene clusters: 303 PKS, 189 NRPS, 147 hybrid NRPS-PKS, 169 RiPP, 78 terpene, 123 saccharide, 21 alkaloid and 140 other BGCs. Compared to the 288 BGCs currently deposited in ClusterMine360[12] and the 103 BGCs deposited in DoBISCUIT[14], this presents a significant advance in terms of comprehensiveness. We then annotated each of these 1,170 BGCs with a minimal number of parameters (genomic locus, publications, chemical structure and biosynthetic class and subclass). Subsequently, in a community initiative involving 81 academic research groups and several companies worldwide, we performed a fully MIBiG-compliant reannotation of 405 of these BGCs according to the information available in earlier publications and laboratory archives. (All participants of this annotation effort are either listed as coauthors of this article or mentioned in the Acknowledgments, depending on the size of their contribution.)

An initial visualization of the full data set arising from this reannotation is publicly available online at http://mibig.secondarymetabolites.org. Altogether, these submitted entries will function as a very useful seed dataset for the development of databases on secondary metabolism. Future data curation efforts will strive to achieve a fully MIBiG-compliant annotation of the remaining 765 BGCs that are currently annotated with a more restricted set of parameters.

Planned implementation

To allow straightforward and user-friendly access, the MIBiG standard will be implemented by multiple databases and web services for genome data and secondary metabolite research. For example, the MIBiG-curated dataset has already been integrated into the antiSMASH tool in the form of a new module[26] that compares any identified BGCs with the full MIBiG-compliant dataset of known BGCs. Moreover, a full-fledged database is currently under development that will be tightly integrated with antiSMASH and will build on the previously published ClusterMine360 framework[12]. Additionally, MIBiG-compliant data will be integrated into the recently released Integrated Microbial Genomes Atlas of Biosynthetic Clusters (IMG-ABC) database from the Joint Genome Institute (https://img.jgi.doe.gov/ABC/)[27]. Regular exchange of data will take place between the MIBiG repository and the IMG-ABC, antiSMASH and ClusterMine databases. Additional cross-links with the chemical databases ChemSpider[28], chEMBL[29] and chEBI[30] are being developed so that researchers can easily find the full MIBiG annotation of the BGC responsible for the biosynthesis of given molecules. Finally, all community-curated data are freely available and downloadable in JSON format for integration into other software tools or databases, without any need to request permission, as long as the source is acknowledged.

For submission of new MIBiG-compliant data by scientists in the field, we prepared an interactive online submission form (available from http://mibig.secondarymetabolites.org), which was extensively tested through the community annotation effort. Data can also be submitted through the BioSynML plug-in[26] (http://www.biosynml.de) that was recently built for use in the Geneious software. In this way, MIBiG-compliant data can easily be integrated with the in-house BGC content management systems of individual laboratories or companies. Finally, it will be possible to submit updates to existing MIBiG entries based on peer-reviewed articles through dedicated web forms.

Future perspectives

The MIBiG coordinating team within the GSC is committed to ensuring the continued support and curation of the MIBiG standard, in cooperation with its partners. Compliance with the standard and interoperability with other standards and databases will also be guaranteed within the GSC. In order to stay relevant and viable, MIBiG is projected to be a 'living' standard: updates will be made as needed to remain technologically and scientifically current.

Coordination with relevant journals will be sought to make MIBiG submission of BGCs (evidenced by MIBiG accession codes) a standard item to check during manuscript review. To stimulate submission of MIBiG data during the process of publishing new biosynthetic gene clusters, unique MIBiG accession numbers are provided for each BCG that can be used during article review (including for data embargoed until after publication). The research community represented by this paper commits itself to submitting MIBiG-compliant data sets as well as updates to existing entries when publishing new experimental results on

BGCs. We encourage the larger community to join in this endeavor.

References

1. Cimermancic, P. *et al. Cell* **158**, 412–421 (2014).
2. Doroghazi, J.R. *et al. Nat. Chem. Biol.* **10**, 963–968 (2014).
3. Kersten, R.D. *et al. Nat. Chem. Biol.* **7**, 794–802 (2011).
4. Kersten, R.D. *et al. Proc. Natl. Acad. Sci. USA* **110**, E4407–E4416 (2013).
5. Gubbens, J. *et al. Chem. Biol.* **21**, 707–718 (2014).
6. Inokuma, Y. *et al. Nature* **495**, 461–466 (2013).
7. Charlop-Powers, Z., Milshteyn, A. & Brady, S.F. *Curr. Opin. Microbiol.* **19**, 70–75 (2014).
8. Wilson, M.C. & Piel, J. *Chem. Biol.* **20**, 636–647 (2013).
9. Wilson, M.C. *et al. Nature* **506**, 58–62 (2014).
10. Shao, Z. *et al. ACS Synth. Biol.* **2**, 662–669 (2013).
11. Yamanaka, K. *et al. Proc. Natl. Acad. Sci. USA.* **111**, 1957–1962 (2014).
12. Conway, K.R. & Boddy, C.N. *Nucleic Acids Res.* **41**, D402–D407 (2013).
13. Anand, S. *et al. Nucleic Acids Res.* **38**, W487–W496 (2010).
14. Ichikawa, N. *et al. Nucleic Acids Res.* **41**, D408–D414 (2013).
15. Field, D. *et al. PLoS Biol.* **9**, e1001088 (2011).
16. Yilmaz, P. *et al. Nat. Biotechnol.* **29**, 415–420 (2011).
17. Field, D. *et al. Nat. Biotechnol.* **26**, 541–547 (2008).
18. Arnison, P.G. *et al. Nat. Prod. Rep.* **30**, 108–160 (2013).
19. Röttig, M. *et al. Nucleic Acids Res.* **39**, W362–W367 (2011).
20. Khayatt, B.I., Overmars, L., Siezen, R.J. & Francke, C. *PLoS One* **8**, e62136 (2013).
21. Baranašić, D. *et al. J. Ind. Microbiol. Biotechnol.* **41**, 461–467 (2014).
22. Charlop-Powers, Z. *et al. Elife* **4**, e05048 (2015).
23. Gilbert, J.A., Jansson, J.K. & Knight, R. *BMC Biol.* **12**, 69 (2014).
24. Bork, P. *et al. Science* **348**, 873 (2015).
25. Kopf, A. *et al. GigaScience* **4**, 27 (2015).
26. Weber, T. *et al. Nucleic Acids Res.* **43**, W237–W243 (2015).
27. Hadjithomas, M. *et al. mBio* **6**, e00932-15 (2015).
28. Pence, H.E. & Williams, A. *J. Chem. Educ.* **87**, 1123–1124 (2010).
29. Bento, A.P. *et al. Nucleic Acids Res.* **42**, D1083–D1090 (2014).
30. Hastings, J. *et al. Nucleic Acids Res.* **41**, D456–D463 (2013).

Acknowledgments

M.H.M. was supported by a Rubicon fellowship of the Netherlands Organization for Scientific Research (NWO; Rubicon 825.13.001). The work of R.K. was supported by the European Union's Seventh Framework Programme (Joint Call OCEAN.2011–2: Marine microbial diversity—new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no. 287589 (Micro B3). M.C. was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) studentship (BB/J014478/1). The GSC is supported by funding from the Natural Environment Research Council (UK), the National Institute for Energy Ethics and Society (NIEeS; UK), the Gordon and Betty Moore Foundation, the National Science Foundation (NSF; US) and the US Department of Energy. The Manchester Synthetic Biology Research Centre, SYNBIOCHEM, is supported by BBSRC/Engineering and Physical Sciences Research Council (EPSRC) grant BB/M017702/1. We thank P. d'Agostino, P.R. August, R. Chau, C.D. Deane, S. Diethelm, L. Fernandez-Martinez, A. El Gamal, C. Garcia De Gonzalo, T.H. Grossman, C.-J. Huang, S. Kodani, A.L. Leandrini, I.A. MacNeil, M. Metelev, E.M. Molly, C. Olano, M. Ortega, L. Ray, K. Reynolds, A. Ross, I.N. Silva, R. Teufel, G. Thibodeaux, J. Tietz and D. Widdick for their contributions in the community annotation. We thank R. Baltz, M. Bibb, C. Boddy, C. Corre, E. Dittmann, H. Gramajo, N. Ichikawa, H. Ikeda, P. Jensen, C. Khosla, R. Li, M. Marahiel, D. Mohanty, C. Moore, W. Nierman, D.-C. Oh, E. Schmidt, Y. Shen, D. Stevens, B. Tudzynski and S. Van Lanen for useful comments on an early draft version of the community standard. We are grateful to three anonymous referees for their constructive suggestions.

Author contributions

M.H.M., R.B., E.T. and F.O.G. initiated and coordinated the MIBiG standardization project. M.H.M. designed the first draft of the standard. M.H.M., R.K., P.Y. and F.O.G. coordinated integration within the MIxS framework. M.H.M. and R.K. constructed the submission system. M.H.M. and M.C. curated submitted community annotation entries. M.H.M., R.K., P.Y., M.C., R.B., E.T. and F.O.G. wrote the first draft of the paper. All authors contributed to the design of the standard, contributed to the community annotation of MIBiG entries and provided feedback on an early draft of the paper.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available in the online version of the paper.

Marnix H Medema[1,115,*], Renzo Kottmann[1], Pelin Yilmaz[1], Matthew Cummings[2], John B Biggins[3], Kai Blin[4], Irene de Bruijn[5], Yit Heng Chooi[6–8], Jan Claesen[9,10], R Cameron Coates[11], Pablo Cruz-Morales[12], Srikanth Duddela[13], Stephanie Düsterhus[14], Daniel J Edwards[15], David P Fewer[16], Neha Garg[17], Christoph Geiger[14], Juan Pablo Gomez-Escribano[18], Anja Greule[19], Michalis Hadjithomas[11], Anthony S Haines[20], Eric J N Helfrich[21], Matthew L Hillwig[22], Keishi Ishida[23], Adam C Jones[24], Carla S Jones[25], Katrin Jungmann[13], Carsten Kegler[26], Hyun Uk Kim[4,27], Peter Kötter[14], Daniel Krug[13], Joleen Masschelein[28,29], Alexey V Melnik[17], Simone M Mantovani[30], Emily A Monroe[31], Marcus Moore[32], Nathan Moss[30], Hans-Wilhelm Nützmann[33], Guohui Pan[34], Amrita Pati[11], Daniel Petras[35], F Jerry Reen[36], Federico Rosconi[37], Zhe Rui[38,39], Zhenhua Tian[40], Nicholas J Tobias[26], Yuta Tsunematsu[23,41], Philipp Wiemann[42], Elizabeth Wyckoff[43,44], Xiaohui Yan[34], Grace Yim[45], Fengan Yu[46–49], Yunchang Xie[50], Bertrand Aigle[51], Alexander K Apel[52,53], Carl J Balibar[54], Emily P Balskus[55], Francisco Barona-Gómez[12], Andreas Bechthold[19], Helge B Bode[26,56], Rainer Borriss[57], Sean F Brady[3], Axel A Brakhage[23], Patrick Caffrey[58], Yi-Qiang Cheng[59], Jon Clardy[60], Russell J Cox[61,62], René De Mot[63], Stefano Donadio[64], Mohamed S Donia[65], Wilfred A van der Donk[66,67], Pieter C Dorrestein[17,30,68], Sean Doyle[69], Arnold J M Driessen[70], Monika Ehling-Schulz[71], Karl-Dieter Entian[14], Michael A Fischbach[9,10], Lena Gerwick[30], William H Gerwick[17,30], Harald Gross[52,53], Bertolt Gust[52,53], Christian Hertweck[23,72], Monica Höfte[73], Susan E Jensen[74], Jianhua Ju[50], Leonard Katz[75], Leonard Kaysser[52,53], Jonathan L Klassen[76], Nancy P Keller[42,77], Jan Kormanec[78], Oscar P Kuipers[79], Tomohisa Kuzuyama[80], Nikos C Kyrpides[11,81], Hyung-Jin Kwon[82], Sylvie Lautru[83], Rob Lavigne[28], Chia Y Lee[84], Bai Linquan[85,86], Xinyu Liu[22], Wen Liu[40], Andriy Luzhetskyy[13], Taifo Mahmud[87], Yvonne Mast[88], Carmen Méndez[89,90], Mikko Metsä-Ketelä[91], Jason Micklefield[92], Douglas A Mitchell[66], Bradley S Moore[17,30], Leonilde M Moreira[93], Rolf Müller[13], Brett A Neilan[94], Markus Nett[23], Jens Nielsen[4,95], Fergal O'Gara[36,96], Hideaki Oikawa[97], Anne Osbourn[33], Marcia S Osburne[98], Bohdan Ostash[99], Shelley M Payne[43,44], Jean-Luc Pernodet[83], Miroslav Petricek[100], Jörn Piel[21], Olivier Ploux[101], Jos M Raaijmakers[5], José A Salas[89], Esther K Schmitt[102], Barry Scott[103],

Ryan F Seipke[104], Ben Shen[34,105], David H Sherman[46–49], Kaarina Sivonen[16], Michael J Smanski[106,107], Margherita Sosio[64], Evi Stegmann[53,88], Roderich D Süssmuth[35], Kapil Tahlan[32], Christopher M Thomas[20], Yi Tang[6,7], Andrew W Truman[18], Muriel Viaud[108], Jonathan D Walton[109], Christopher T Walsh[110], Tilmann Weber[4], Gilles P van Wezel[111], Barrie Wilkinson[18], Joanne M Willey[112], Wolfgang Wohlleben[53,88], Gerard D Wright[45], Nadine Ziemert[53,88], Changsheng Zhang[50], Sergey B Zotchev[113], Rainer Breitling[2], Eriko Takano[2] & Frank Oliver Glöckner[1,114]

[1]Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany. [2]Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, UK. [3]Laboratory of Genetically Encoded Small Molecules, Howard Hughes Medical Institute, The Rockefeller University, New York, New York, USA. [4]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark. [5]Netherlands Institute of Ecology (NIOO-KNAW), Department of Microbial Ecology, Wageningen, the Netherlands. [6]Department of Chemical and Biomolecular Engineering, University of California Los Angeles, Los Angeles, California, USA. [7]Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, California, USA. [8]School of Chemistry and Biochemistry, University of Western Australia, Perth, Western Australia, Australia. [9]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA. [10]California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, USA. [11]Department of Energy (DOE) Joint Genome Institute, Walnut Creek, California, USA. [12]Evolution of Metabolic Diversity Laboratory, Unidad de Genómica Avanzada (Langebio), Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav-IPN), Irapuato, Guanajuato, México. [13]Helmholtz Institute for Pharmaceutical Research, Helmholtz Centre for Infection Research and Department of Pharmaceutical Biotechnology, Saarland University, Saarbrücken, Germany. [14]Institute for Molecular Biosciences, Goethe University, Frankfurt am Main, Germany. [15]Department of Chemistry and Biochemistry, California State University, Chico, California, USA. [16]Microbiology and Biotechnology Division, Department of Food and Environmental Sciences, University of Helsinki, Helsinki, Finland. [17]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA. [18]Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Norwich, UK. [19]Department of Pharmaceutical Biology and Biotechnology, Albert-Ludwigs-University of Freiburg, Freiburg, Germany. [20]School of Biosciences, University of Birmingham, Birmingham, UK. [21]Institute of Microbiology, Eidgenössische Technische Hochschule (ETH) Zürich, Zürich, Switzerland. [22]Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [23]Leibniz Institute for Natural Product Research and Infection Biology (HKI), Jena, Germany. [24]Gordon and Betty Moore Foundation, Palo Alto, California, USA. [25]Sustainable Studies Program, Roosevelt University Chicago, Illinois, USA. [26]Merck Stiftungsprofessur für Molekulare Biotechnologie, Goethe Universität Frankfurt, Fachbereich Biowissenschaften, Frankfurt, Germany. [27]BioInformatics Research Center, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea. [28]Laboratory of Gene Technology, KU Leuven, Heverlee, Belgium. [29]Laboratory of Food Microbiology, KU Leuven, Heverlee, Belgium. [30]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA. [31]Department of Biology, William Paterson University, Wayne, New Jersey, USA. [32]Department of Biology, Memorial University of Newfoundland, St. John's, Newfoundland, Canada. [33]Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Norwich, UK. [34]Department of Chemistry, The Scripps Research Institute, Jupiter, Florida, USA. [35]Institut für Chemie, Technische Universität Berlin, Berlin, Germany. [36]BIOMERIT Research Centre, School of Microbiology, University College Cork–National University of Ireland, Cork, Ireland. [37]Departamento de Bioquímica y Genómica Microbianas, IBCE, Montevideo, Uruguay. [38]Energy Biosciences Institute, University of California Berkeley, Berkeley, California, USA. [39]Department of Chemical and Biomolecular Engineering, University of California Berkeley, Berkeley, California, USA. [40]State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, China. [41]Department of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. [42]Department of Medical Microbiology and Immunology, University of Wisconsin–Madison, Madison, Wisconsin, USA. [43]Department of Molecular Biosciences, The University of Texas, Austin, Texas, USA. [44]Institute for Cellular and Molecular Biology, The University of Texas, Austin, Texas, USA. [45]Department of Biochemistry and Biomedical Sciences, The M.G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada. [46]Life Sciences Institute, University of Michigan, Ann Arbor, Michigan, USA. [47]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan, USA. [48]Department of Chemistry, University of Michigan, Ann Arbor, Michigan, USA. [49]Department of Microbiology & Immunology, University of Michigan, Ann Arbor, Michigan, USA. [50]Key Laboratory of Tropical Marine Bio-resources and Ecology, Guangdong Key Laboratory of Marine Materia Medica, RNAM Center for Marine Microbiology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China. [51]Dynamique des Génomes et Adaptation Microbienne, Université de Lorraine and Institut National de la Recherche Agronomique (INRA), Unité Mixte de Recherche (UMR) 1128, Vandœuvre-lès-Nancy, France. [52]Pharmaceutical Institute, Department of Pharmaceutical Biology, University of Tübingen, Tübingen, Germany. [53]German Centre for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany. [54]Infectious Disease Research, Merck Research Laboratories, Kenilworth, New Jersey, USA. [55]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, USA. [56]Buchmann Institute for Molecular Life Sciences (BMLS), Goethe Universität Frankfurt, Frankfurt, Germany. [57]Fachbereich Phytomedizin, Albrecht Thaer Institut, Humboldt Universität Berlin, Berlin, Germany. [58]UCD School of Biomolecular and Biomedical Science, University College Dublin, Dublin, Ireland. [59]UNT System College of Pharmacy, University of North Texas Health Science Center, Fort Worth, Texas, USA. [60]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [61]Institut für Organische Chemie, Leibniz Universität Hannover, Hannover, Germany. [62]School of Chemistry, University of Bristol, Bristol, UK. [63]Centre of Microbial and Plant Genetics, Faculty of Bioscience Engineering, University of Leuven, Heverlee, Belgium. [64]Naicons Srl, Milano, Italy. [65]Department of Molecular Biology, Princeton University, Princeton, New Jersey, USA. [66]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana-Champaign, Illinois, USA. [67]Howard Hughes Medical Institute, USA. [68]Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, California, USA. [69]Department of Biology, Maynooth University, Maynooth, County Kildare, Ireland. [70]Department of Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, Groningen, the Netherlands. [71]Functional Microbiology, Institute of Microbiology, Department of Pathobiology, University of Veterinary Medicine Vienna, Vienna, Austria. [72]Friedrich Schiller University, Jena, Germany. [73]Department of Crop Protection, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium. [74]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. [75]Synthetic Biology Engineering Research Center (SynBERC), University of California Emeryville, Emeryville, California, USA. [76]Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, USA. [77]Department of Bacteriology, University of Wisconsin–Madison, Madison, Wisconsin, USA. [78]Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovak Republic. [79]Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, the Netherlands. [80]Biotechnology Research Center, The University of Tokyo, Tokyo, Japan. [81]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia. [82]Division of Bioscience and Bioinformatics, Myongji University, Yongin-si, Gyeonggi-Do, South Korea. [83]Institute of Integrative Biology of the Cell (I2BC), Commissariat à l'Energie Atomique (CEA), Centre National de la Recherche Scientifique (CNRS), Université Paris Sud, Orsay,

France. [84]*Department of Microbiology and Immunology, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA.* [85]*State Key Laboratory of Microbial Metabolism, Shanghai Jiao Tong University, Shanghai, China.* [86]*School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, China.* [87]*Department of Pharmaceutical Sciences, Oregon State University, Corvallis, Oregon, USA.* [88]*Microbiology/Biotechnology, Interfaculty Institute of Microbiology and Infection Medicine, Faculty of Science, University of Tübingen, Tübingen, Germany.* [89]*Departamento de Biología Funcional, Universidad de Oviedo, Oviedo, Spain.* [90]*Instituto Universitario de Oncología del Principado de Asturias (I.U.O.P.A), Universidad de Oviedo, Oviedo, Spain.* [91]*Department of Biochemistry, University of Turku, Turku, Finland.* [92]*School of Chemistry, University of Manchester, Manchester, UK.* [93]*Institute for Bioengineering and Biosciences, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal.* [94]*School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia.* [95]*Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden.* [96]*Curtin University, School of Biomedical Sciences, Perth, Western Australia, Australia.* [97]*Division of Chemistry, Graduate School of Science, Hokkaido University, Sapporo, Japan.* [98]*Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, Massachusetts, USA.* [99]*Department of Genetics and Biotechnology, Ivan Franko National University of Lviv, Lviv, Ukraine.* [100]*Institute of Microbiology, Academy of Sciences of the Czech Republic (ASCR), Prague, Czech Republic.* [101]*Laboratoire Interdisciplinaire des Energies de Demain (LIED), UMR 8236 CNRS, Université Paris Diderot, Paris, France.* [102]*Novartis Institutes for BioMedical Research, Novartis Campus, Basel, Switzerland.* [103]*Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand.* [104]*Astbury Centre for Structural Molecular Biology, School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK.* [105]*Molecular Therapeutics and Natural Products Library Initiative, The Scripps Research Institute, Jupiter, Florida, USA.* [106]*Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota–Twin Cities, Saint Paul, Minnesota, USA.* [107]*BioTechnology Institute, University of Minnesota–Twin Cities, Saint Paul, Minnesota, USA.* [108]*Unité BIOlogie et GEstion des Risques en agriculture (BIOGER), Institut National de la Recherche Agronomique (INRA), Grignon, France.* [109]*Department of Energy Great Lakes Bioenergy Research Center and Department of Energy Plant Research Laboratory, Michigan State University, East Lansing, Michigan, USA.* [110]*Chemistry, Engineering & Medicine for Human Health (ChEM-H) Institute, Stanford University, Stanford, California, USA.* [111]*Molecular Biotechnology, Institute of Biology, Leiden University, Leiden, the Netherlands.* [112]*Hofstra North Shore–Long Island Jewish School of Medicine, Hempstead, New York, USA.* [113]*Department of Biotechnology, Norwegian University of Science and Technology, Trondheim, Norway.* [114]*Jacobs University Bremen gGmbH, Bremen, Germany.* [115]*Present address: Bioinformatics Group, Wageningen University, Wageningen, the Netherlands.*
**e-mail: marnix.medema@wur.nl*