

The Genomic Standards Consortium

Dawn Field^{1*}, Linda Amaral-Zettler², Guy Cochrane³, James R. Cole⁴, Peter Dawyndt⁵, George M. Garrity⁶, Jack Gilbert^{7,8}, Frank Oliver Glöckner⁹, Lynette Hirschman¹⁰, Ilene Karsch-Mizrachi¹¹, Hans-Peter Klenk¹², Rob Knight¹³, Renzo Kottmann⁹, Nikos Kyrpides¹⁴, Folker Meyer^{7,15}, Inigo San Gil¹⁶, Susanna-Assunta Sansone¹⁷, Lynn M. Schriml¹⁸, Peter Sterk¹⁹, Tatiana Tatusova¹¹, David W. Ussery²⁰, Owen White¹⁸, John Wooley²¹

1 Centre for Ecology & Hydrology, Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire, United Kingdom, **2** The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, United States of America, **3** European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **4** Center for Microbial Ecology, Michigan State University, East Lansing, Michigan, United States of America, **5** Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium, **6** Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, United States of America, **7** Argonne National Laboratory, Argonne, Illinois, United States of America, **8** Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America, **9** Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, Bremen, Germany, **10** Information Technology Center, The MITRE Corporation, Bedford, Massachusetts, United States of America, **11** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **12** DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany, **13** Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, United States of America, **14** DOE Joint Genome Institute, Walnut Creek, California, United States of America, **15** Computation Institute, University of Chicago, Chicago, Illinois, United States of America, **16** LTER Network Office, Department of Biology, University of New Mexico, Albuquerque, New Mexico, United States of America, **17** University of Oxford, Oxford e-Research Centre, Oxford, United Kingdom, **18** Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, United States of America, **19** Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **20** Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark, **21** University of California San Diego, La Jolla, California, United States of America

Abstract: A vast and rich body of information has grown up as a result of the world's enthusiasm for 'omics technologies. Finding ways to describe and make available this information that maximise its usefulness has become a major effort across the 'omics world. At the heart of this effort is the Genomic Standards Consortium (GSC), an open-membership organization that drives community-based standardization activities. Here we provide a short history of the GSC, provide an overview of its range of current activities, and make a call for the scientific community to join forces to improve the quality and quantity of contextual information about our public collections of genomes, metagenomes, and marker gene sequences.

of information unthinkable even a few years ago. Each data set represents an organism or community with a unique biological history, sampling location, environmental context, and set of biologically interesting traits. Hence, each of these data sets makes a unique contribution to the ongoing creation of our public online catalogue of life.

We are now witnessing the rapid democratization of access to sequencing capacity—an immense opportunity for the global community, if proper stewardship of these data keeps pace [2,3]. This stewardship must include enriching public sequence databases with the biological context of these sequences (Box 1), which will in turn necessitate the adoption of a fresh attitude to reporting results, both in our papers and our submissions to the public databases. Large, well-contextualized genome, metagenome, and marker gene data sets (e.g., ribosomal gene

surveys) provide ideal opportunities for comparison and contrasting using computational means to solve a wide range of questions in biology (including questions in medicine, physiology, developmental biology, biogeochemistry, evolution, ecology, etc.).

These data sets should be treated as part of a larger whole—a catalogue of life on earth—that will allow us to observe, as we sample in time and space, how life changes. A range of ongoing and proposed megasequencing projects also promise to make great inroads into this grand vision (i.e., the Genomic Encyclopedia of *Bacteria* and *Archaea* [GEBA] [4], Human Microbiome Project [HMP] [5], Microbial Earth Project [<http://genome.jgi-psf.org/programs/bacteria-archaea/MEP/index.jsf>], Earth Microbiome Project [6], Genomes 10K [7], Tara Oceans [<http://oceans.taraexpeditions.org/>], Malaspina [http://en.wikipedia.org/wiki/Malaspina_Expedition_

Introduction

We currently have thousands of genomes, hundreds of metagenomes, and tens of thousands of marker gene data sets in the public domain, and these numbers are rapidly increasing [1]. Next-generation sequencing technologies promise to further fill the public databases with a bounty

Citation: Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, et al. (2011) The Genomic Standards Consortium. PLoS Biol 9(6): e1001088. doi:10.1371/journal.pbio.1001088

Published: June 21, 2011

Copyright: © 2011 Field et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: NERC International Opportunities Fund Award NE/3521773/1 and NE/E007325/1 (<http://www.nerc.ac.uk/funding/>) and National Science Foundation grant RCN4GSC, DBI-0840989 (<http://www.nsf.gov/funding/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: GSC, Genomic Standards Consortium; MlxS, Minimum Information about any (x) Sequence

* E-mail: dfield@ceh.ac.uk

The Community Page is a forum for organizations and societies to highlight their efforts to enhance the dissemination and value of scientific knowledge.

Box 1. When the Cost of a Bacterial Genome Sequence Is Almost Nothing, That Organism's Contextual Information Is Increasingly Valuable

Consider the scenario where a new *E. coli* sequence has been obtained from a futuristic handheld device (like a *Star Trek* tricorder) that generates the complete genome in seconds. While the genome sequence may only be slightly different from strains already in the public databases, the metadata associated with this bug is both unique and crucial. Where and when was the *E. coli* isolated? Was it transmitted as a food-borne pathogen? Did it hospitalize the patient from whom it was isolated? Was it part of a larger infectious outbreak? Knowledge that a pathogen was isolated from diseased patients or healthy controls will readily assist in intervention strategies derived from machine-readable data.

2010], Sorcerer II Global Ocean Sampling expedition [8]).

How must we now change the way we think about these data sets to prepare to integrate and co-analyze these large suites of related and contrasting data? Clearly, these data must be stored in robust comprehensive electronic systems that link to specific environments, diseases, or physiological states such that these relationships are electronically retrievable. To achieve this goal we urgently need shared standards that are both easy to use and scientifically robust.

The Genomic Standards Consortium

The GSC was established in late 2005 [9,10] to tackle the challenge of working towards better descriptions of genomes and metagenomes through community-level, consensus-driven solutions. The GSC's mission is to work towards 1) the implementation of new genomic standards, 2) methods of capturing and exchanging the information captured in these standards (metadata, or contextual data) and 3) harmonization of information collection and analysis efforts across the wider genomics community.

The GSC fulfils this mission by holding face-to-face meetings, forming working groups, and building consensus products that can be widely used in this community. Thus far, the GSC has created a standard, the Minimum Information about any (x) Sequence (MIxS), that includes three minimum information checklists for describing genomes, metagenomes, and environmental marker sequences (MIGS/MIMS/MIMARKS) upon submission to the public databases and publication [11,12]. MIxS requires core information on habitat, geolocation, and sequencing methodology as well as fields specific to data type and a range of optional environmental packages to capture core measurements defining a broad range of habitats, including water, soil, and host-

associated habitats. The International Nucleotide Sequence Database Collaboration (INSDC; DDBJ/EMBL/GenBank) has created a GSC "keyword" (MIxS) to mark the richer entries complying with this standard.

Other working groups are dedicated to 1) the maintenance of an extensible markup language (GCDML) that provides a reference implementation of the MIxS checklists [13], 2) development of tools and software, 3) compliance and curation, and 4) biodiversity. Those requiring help complying with MIxS (curation support) should contact the compliance working group, and those requiring technical assistance in implementing/adopting these standards in software or database projects should contact the developer's working group (technical support). The developer and compliance groups work closely together, for example, to support compliance through a range of portals, including GOLD [1], MG-Rast [14], CAMERA [15], IMG/m [16], the RDP [17], SILVA [18], megx.net [19], and the ISA software suite [20]. The Biodiversity group works with communities to make sure that GSC standards evolve in harmony with standards for describing taxonomy and biodiversity.

The GSC has also stepped forward to create a journal designed to underpin the emerging field of standards development in the biological sciences [21]. The *Standards in Genomic Sciences* journal now serves as a formal voice for the GSC and supports the publication of standardized genome, metagenome, and pan-genome reports and other standards-supportive publications like Standard Operating Procedures (SOPs) [22] from the scientific community at large.

The GSC is now maturing into a hub for the coordination of large-scale projects. Two projects running under the GSC umbrella are the Microbial Earth Project, which calls for the coordinated sequencing of over 9,000 type strains (<http://genome.jgi-psf.org/programs/bacteria-archaea/MEP/index.jsf>), and the M5 project,

which calls for the coordinated development of a next-generation computational infrastructure (http://gensc.org/gc_wiki/index.php/M5) [23].

The GSC also works closely with a range of related communities and helped drive the formation of the Environment Ontology [24], the Minimum Information for Biological and Biomedical Investigations (MIBBI) initiative [2], and most recently the BioSharing forum [3, 25].

A Call for Participation and Adoption

The Internet has resulted in a Cambrian explosion of productivity and data sharing through the adoption of a huge stack of agreed-upon protocols (standards) that allow many devices and programs to communicate to the transformative benefit of the everyday user [26]. Enabling access to user-generated content is key to harnessing the resources of a distributed community: Flickr has over 5 billion photographs uploaded, and Wikipedia has over 3.5 million English articles as of this writing. Standards for organizing sequence data will be similarly needed as sequencing instruments themselves, especially as these instruments are more and more commoditized and owned by individuals rather than institutions.

The tagline of the GSC is "Innovation through Collaboration". For any standard to create a lasting impact requires substantial input from the wider scientific community, including adoption and support. The GSC urges researchers interested in pushing the boundaries of genomic science through collaboration to join and contribute expertise to building the GSC roadmap for the future. Membership in the GSC and all working groups is currently defined by participation. The GSC has a Board and several standing committees in addition to its working groups. For more information on the GSC, please see <http://gensc.org/>.

Conclusions

The GSC is working to become the authoritative working body in the area of genomics for the development and adoption of standards. We anticipate that the need for a collaborative body in which to build consensus at the community level and undertake large-scale projects will only increase with time, as in many ways the era of genomics is just beginning. In the future, sequence generation will only increase as access is further democratized. On one extreme, it will be like any other

industrial commodity and will be out-sourced into a global manufacturing marketplace. On the other, mid- to large-scale sequencing will be as locally accessible as a benchtop microscope or PCR machine is to a typical university researcher. Making these diverse streams of data accessible in a coherent framework will require new, standardized ways of describing, storing, and exchanging this information. The

framework required to do this will involve acceptance of profound sociological and technological changes in how we do business in the genomic sciences.

Acknowledgments

The GSC acknowledges all participants in past GSC meetings for their thoughtful contributions. The GSC also acknowledges a range of funding sources for its past meetings, including

NERC, NIEES, NSF, the Gordon and Betty Moore Foundation and DOE. In particular, funding from NERC helped launch the GSC and allow essential infrastructure to be built. Funding from the NSF in the form of the Research Co-ordination Network (*RCN4GSC*) is supporting exchange visits of early career scientists and working group activities.

References

- Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, et al. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 38: D346–D354.
- Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, et al. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* 26: 889–896.
- Field D, Sansone SA, Collis A, Booth T, Dukes P, et al. (2009) 'Omics data sharing. *Science* 326: 234–236.
- Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462: 1056–1060.
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. *Nature* 449: 804–810.
- Gilbert J, Meyer F, Jansson J, Gordon J, Pace N, et al. (2010) The Earth Microbiome Project: Meeting report of the “1st EMP meeting on sample selection and acquisition” at Argonne National Laboratory October 6th 2010. *Stand Genomic Sci* 3: 249–253.
- (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* 100: 659–674.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5: e77. doi:10.1371/journal.pbio.0050077.
- Field D, Hughes J (2005) Cataloguing our current genome collection. *Microbiology* 151: 1016–1019.
- Field D, Garrity G, Morrison N, Selengut J, Sterk P, et al. (2005) eGenomics: cataloguing our complete genome collection. *Comp Funct Genomics* 6: 363–368.
- Field D, Garrity G, Gray T, Morrison N, Selengut J, et al. (2008) The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 26: 541–547.
- Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, et al. (2011) The “Minimum Information about a MARKer gene Sequence” (MIMARKS) specification. *Nat Biotechnol* 29: 415–420.
- Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, et al. (2008) A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 12: 115–121.
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386.
- Sun S, Chen J, Li W, Altintas I, Lin A, et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAM-ERA resource. *Nucleic Acids Res* 39: D546–D551.
- Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, et al. (2010) The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res* 38: D382–D390.
- Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 37: D141–D145.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35: 7188–7196.
- Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, et al. (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* 38: D391–395.
- Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, et al. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 26: 2354–2356.
- Garrity GM, Field D, Kyrpides N, Hirschman L, Sansone SA, et al. (2008) Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 12: 157–160.
- Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, et al. (2008) Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. *OMICS* 12: 137–141.
- (2009) Metagenomics versus Moore's law. *Nat Meth* 6: 623.
- Morrison N, Wood AJ, Hancock D, Shah S, Hakes L, et al. (2006) Annotation of environmental OMICS data: application to the transcriptomics domain. *OMICS* 10: 172–178.
- Field D, Sansone S, Delong EF, Sterk P, Friedberg I, et al. (2010) Meeting report: BioSharing at ISMB 2010. *Stand Genomic Sci* 3: 254–258.
- Berners-Lee T (22 November 2010) Long live the web: a call for continued open standards and neutrality. *Scientific American*.