

Towards a richer description of our complete collection of genomes and metagenomes: the “Minimum Information about a Genome Sequence” (MIGS) specification

Dawn Field¹, George Garrity², Tanya Gray^{1,3}, Norman Morrison^{3,4}, Jeremy Selengut⁵, Peter Sterk⁶, Tatiana Tatusova⁷, Nicholas Thomson⁸, Michael J. Allen⁹, Michael Ashburner¹⁰, Sandra Baldauf¹¹, Stuart Ballard¹², Jeffrey Boore¹³, Guy Cochrane⁶, James Cole², Claude dePamphilis¹⁴, Robert Edwards¹⁵, Nadeem Faruque⁶, Robert Feldman¹⁶, Frank Oliver Glöckner¹⁷, Dan Haft⁵, David Hancock^{3,4}, Henning Hermjakob⁶, Christiane Hertz-Fowler⁸, Phil Hugenholtz¹⁸, Ian Joint⁹, Matthew Kane¹⁹, Jessie Kennedy²⁰, George Kowalchuk¹¹, Renzo Kottmann¹⁶, Eugene Kolker^{22, 23}, Nikos Kypides²⁴, Jim Leebens-Mack²⁵, Suzanna E Lewis²⁶, Allyson Liste²⁷, Phillip Lord²⁷, Natalia Maltsev²⁸, Victor Markowitz²⁴, Jennifer Martiny²⁹, Barbara Methe⁵, Richard Moxon³⁰, Karen Nelson⁵, Julian Parkhill⁸, Susanna-Assunta Sansone⁶, Andrew Spiers¹, Robert Stevens³, Paul Swift¹, Chris Taylor⁶, Yoshio Tateno³¹, Adrian Tett¹, Sarah Turner¹, David Ussery³², Bob Vaughan⁶, Naomi Ward⁵, Trish Whetzel³³, Gareth Wilson¹, and Anil Wipat²⁷

¹ NERC Centre for Ecology and Hydrology, Oxford, OX1 3SR, UK.

² Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA.

³ School of Computer Science, University of Manchester, Manchester M13 9PL, UK

⁴ The NERC Environmental Bioinformatics Centre, Oxford Centre for Ecology and Hydrology, Oxford OX1 3SR, UK

⁵ The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

⁶ EMBL Outstation. The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

⁷ National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda MD 20894, USA.

⁸ Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

⁹ Plymouth Marine Laboratory, Prospect Place, Plymouth, PL1 3DH, UK.

¹⁰ Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK.

¹¹ Department of Biology, University of York, Box 373, York, YO10 5YW UK.

¹² NIEES, Department of Earth Sciences, University of Cambridge, Downing St, Cambridge, CB2 3EQ

¹³ DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

¹⁴ 208 Mueller Lab, University Park PA 16802, USA

¹⁵ Computational Science Research Center, San Diego State University, San Diego, CA 92182, USA

- ¹⁶ Symbio Corporation, 1455 Adams Dr. Menlo Park, CA 94025, USA.
- ¹⁷ Microbial Genome Research, International University Bremen, Max Planck Institute for Marine Microbiology, Bremen 28359 Germany
- ¹⁸ Phil Hugenholtz Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive Bldg 400-404, Walnut Creek, CA 94598, USA
- ¹⁹ The National Science Foundation, 4201 Wilson Boulevard, Arlington, VA 22230, USA
- ²⁰ School of Computing, Napier University, Merchiston Campus, Edinburgh, Scotland, UK.
- ²¹ Department of Plant-Microorganism Interactions, Netherlands Institute of Ecology, Centre for Terrestrial Ecology, Heteren. Netherlands
- ²² The BIATECH Institute, 19310 North Creek Pkwy S., Suite 115, Bothell, WA 98011, USA
- ²³ Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Information, University of Washington Seattle, WA 98195, USA
- ²⁴ Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94720, USA.
- ²⁵ Department of Plant Biology The University of Georgia Athens, GA 30602-7271, USA
- ²⁶ Department of Molecular and Cell Biology, University of California, 539 Life Sciences Addition, Berkeley, CA 94720-3200, USA.
- ²⁷ School of Computing Science, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
- ²⁸ Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA.
- ²⁹ Department of Ecology and Evolutionary Biology, University of California 455 Steinhaus Hall, Irvine, CA 92697, USA.
- ³⁰ Molecular Infectious Diseases Group, Weatherall Institute of Molecular Medicine and University of Oxford Department of Paediatrics, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK.
- ³¹ Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Mishima, Shizuoka 411-8540, Japan
- ³² Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, 2800, Denmark
- ³³ Center for Bioinformatics and Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

Abstract

By the end of next year there will be complete genome sequences of at least draft quality for more than 1,000 bacteria and archaea and 100 eukaryotes¹ as well as even larger numbers of viruses, organelles and plasmids. With the quantity of genomic information increasing at an exponential rate it is imperative these data be captured electronically, in a standard format. Standardization activities must proceed within the auspices of open-access and international working bodies. To tackle the issues surrounding the development of better descriptions of genomic investigations we have formed the Genomic Standards Consortium (GSC). Here, we introduce the “Minimum Information about a Genome Sequence” specification with the intent of promoting participation in its development and to discuss the resources that will be required to develop improved mechanisms of metadata capture and exchange. As part of its wider goals, the GSC also supports improving the ‘transparency’ of the information contained in existing genomic databases.

A wealth of genomic and metagenomic sequences

At the beginning of the genomic era, few could have imagined the wealth of data we have amassed in a single decade. With the rapid pace at which new genome sequences are appearing, the need to consider how best to ensure suitable stewardship of this data for the long-term has never been more pressing.

Our genome collection: More than the sum of its parts

The analysis of genomic information is having an impact on every area of the life sciences and beyond. A genome is the entire genetic complement of an individual and, as such, is a necessary pre-requisite to understanding the molecular basis of phenotype, how it evolves over time, and how we can manipulate it to provide new solutions to critical problems. Such solutions include potential cures for disease, drug therapies, industrial products including the biodegradation of xenobiotic compounds, and even renewable energy sources. With improvements in the technology of sequencing, the growing interest in metagenomic approaches, and the proven power of comparative analysis of groups of related genomes, the day can be envisioned when it will be commonplace to sequence tens to hundreds of genomes, or more, as part of a single study. At current rates of genome sequencing, it has been estimated that more than 4,000 bacterial genomes will be available soon after 2010¹.

Given the importance of the growing genome collection, the capital investment in its creation, and the benefits of leveraging its value through diverse comparative analyses, it seems obvious that we should make every effort to describe it as accurately and comprehensively as possible. There is an increasing interest from the community in doing so, for three main reasons. The first is the ‘grassroots’ interest of a growing number of isolated researchers in testing hypotheses about the features observed in genomes using comparative evo- and eco-genomic approaches². The second is the growing need to supplement the content of a variety of databases with high-level descriptions of genomes that allow useful grouping, sorting, and searching of the underlying data. The third is the growth in the number of genomes from environmental isolates and metagenomes – vast data sets of DNA fragments from environmental samples - that are being sequenced. The type of data generated by such studies will dwarf current stores of genomic information, making improved descriptions of genomes of utmost importance.

Currently, both top-level descriptors and genome descriptions are incomplete due to numerous factors. First and foremost, through hindsight we now know the minimum quality and quantity of information that is required to make each description precise, accurate and useful. For example, even for bacterial and archaeal species with validly published names, strain names were not routinely captured in genome annotation documents prior to the sequencing of large numbers of genomes from the same species³, but clearly such information is now considered essential. Through empirical observations, we are expanding our view of the types of information that are of critical importance for testing particular hypotheses⁴, exploring new patterns², and quantifying inherent sampling biases².

We are also being forced to re-think our concept of the minimum information required to adequately describe a genome sequence as the number of habitats and communities sampled using metagenomic approaches continues to grow. Without adequate description of the environmental context and experimental methods used to generate these data sets they will be of less value for researchers wishing to conduct subsequent comparative genomic studies or link genetic potential with the diversity and abundance of organisms. In fact, given the vast number of yet to be cultivated microbes, we are now questioning whether a DNA-centric approach, where the genes of microbes are linked to the habitats (locations) in which they are represented, is more useful than the species-centric view^{5,6}. Finally, sequencing technology is advancing rapidly and the new family of methods currently being unleashed^{2,7-9} will force the adoption of additional descriptors (e.g. the depth of sequence coverage, quality and if any 'finishing' was employed) in order to be able to distinguish amongst them.

Most often such metadata is found only in the primary literature (on a per-genome or per-sample basis) or in reference works, such as the widely recognized 'gold standard' for bacteria and archaea, Bergey's Manual¹⁰ (on a per-species basis). The distributed and patchy nature of this information and the difficulties of curating even a few pieces of information for what are now very large collections of genomes make the vision of a single definitive source of rich genomic descriptions highly desirable.

The need for co-ordinated efforts

Facilitating and accelerating the process of collecting relevant metadata would clearly reduce ongoing replication of effort and maximize the ability to share and integrate data within the genomics community. The obvious solution is to develop a consensus-based approach.

The Genomic Standards Consortium

The GSC is an open-membership working body which formed in September 2005¹¹. The goal of this international community is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data. The GSC community brings together (1) evolutionists, ecologists, molecular biologists, and other researchers analyzing collections of genomes, (2) bioinformaticians producing genomic databases, (3) those physically responsible for sequencing genomes and (4) computer scientists, ontology experts, and members of other standardization initiatives. These include key members of the International Nucleotide Sequence Database Collaboration (INSDC) who are responsible for the DDBJ/EMBL/GenBank databases (<http://www.insdc.org/>). The guidance of the DDBJ/EMBL/GenBank will be critical to the success of this initiative both because they are the official stewards of the public collection of genomes and because they make every effort to ensure their resources evolve in accordance with community needs.

Re-evaluating and extending the minimum information collected about genome sequences

The GSC is working to formalize a revised set of core descriptors for genomes through the generation of a “Minimum Information about a Genome Sequence” (MIGS) specification. MIGS provides an extension of the minimum information already captured by the INSDC. The MIGS checklist (v1.1) is available from the consortium’s website (<http://gensc.sf.net>) and is briefly overviewed here. The information required to comply with MIGS is routinely reported in primary genome publications (or is referenced therein). However, this information needs to be formalized and made available in electronic form to improve its accessibility ¹².

Since originally proposed ¹², the MIGS specification has been simplified and updated by the GSC through an iterative process of revision to contain: (1) only curated information that can not be calculated from raw genomic sequence, and (2) core descriptors specific to the major taxonomic groups (eukaryotes, bacteria/archaea ¹³, plasmids, viruses, organelles and metagenomes). MIGS is structured into an ‘*Investigation*’ composed of a ‘*Study*’ and an ‘*Assay*’, according to the Reporting Structures for Biological Investigations (RSBI) working group’s recommendation for the modularization of checklists ^{14, 15}. Under ‘*Study*’ sit the concepts Organism, Phenotype, Environment and under ‘*Assay*’ are Sample Processing and Data Processing. At its core, MIGS aims to support unencumbered access to genomic reagents (e.g. strains) ¹⁶, place the complete genome collection into geospatial and temporal context (latitude, longitude, altitude/depth, date and time of sampling), and provide essential details of the experimental method used (e.g. sequencing method). MIGS also provides a framework for the capture of additional information deemed ‘minimum’ to specific communities. For example, the “Minimum Information about a Metagenomic Sequence” (MIMS) specification has recently been introduced as an extension of MIGS ¹⁷.

The way in which genomes are described in our public databases has directly evolved from how we describe even the shortest and simplest pieces of DNA sequences without special attention to information such as the geographical origin of the sequence. Significant efforts are underway by the INSDC to adapt and extend the infrastructure for describing genomes through the Genome Project Metadata initiative ¹⁸. The INSDC efforts are open to evolution, albeit at a conservative pace ¹⁸, and aim to incorporate as much, if not all, of the MIGS specification covered by the Genome Project Metadata initiative. A mapping between INSDC features and MIGS has been developed for the purpose of placing MIGS information into INSDC documents and is available on our website. Any fields which are not already formally defined by the INSDC Feature Table Document (http://www.insdc.org/feature_table.html) will be represented within a structured comment block in INSDC records ¹⁸.

A Genome Catalogue – capturing input from the genomics community

The development of any checklist must be an open and iterative process that involves a balanced group of participants. Further, this development process must be supported by providing mechanisms for achieving compliance if a checklist is to be adopted as a tool for the standardization of a particular area

of knowledge. Such mechanisms involve an appropriate reporting structure for capturing and exchanging data (file formats), software, databases, and the development of appropriate controlled vocabularies and/or ontologies for defining the terms used in the annotations¹¹. The GSC is working towards these combined goals and has created an online system for capturing MIGS-compliant genome reports (gensc.sf.net).

In brief, we have implemented the MIGS checklist as an XML schema (migs.xsd) and built a freely available Genome Catalogue system (GCat) (gensc.sf.net). GCat is designed to generate forms automatically and ‘on-the-fly’ from this schema for the sake of data input. It also allows users to view and search genome descriptions as they accumulate during the process of refining the MIGS checklist. The GCat system is generic and could be applied to the capture of more expressive metadata for subsets of genomes. Indeed, it is flexible enough to support the implementation of any checklist that can be structured as an appropriate XML schema. The GSC is also working in the area of controlled vocabulary and ontology development through the collation of controlled vocabularies already in use in the community and through contributions to the Ontology for Biomedical Investigations (OBI, previously known as the Functional Genomics Investigation Ontology (FuGO)¹⁹). As a part of this process we have engineered GCat to make use of existing controlled vocabulary terms and to accept new terms that emerge from the community.

Increasing the transparency and value of information in genomic databases

By design, MIGS only contains *primary*, curated information. This is because *secondary*, or derived, information that can be calculated from a genome sequence is subject to frequent change, can be generated using multiple methods, and should be acquired directly from those producing the calculations. Still, access to computed information (e.g. in the simplest cases G+C content or the total number of predicted proteins) should be made as easily accessible as possible.

Genomic sequences and their initial annotations must be submitted to the INSDC (http://www.insdc.org/open_letter.txt) (and subsequent high quality, curated annotations derived from empirical observations to the Third Party Annotation dataset²⁰) but there are an ever-increasing number of genomic databases containing a wide range of additional computations. While it is not part of the goals of the GSC to endorse any particular method of analysis or database, it is in the interest of the genomic community to see the transparency of such resources increase for the sake of accurate downstream interpretation of the data and integration.

The first issue is that of exchanging calculated information. This could be facilitated in part, by the wide-spread adoption of a common exchange format, for example the Generic Feature Format Version 3 (GFF3) file format (<http://song.sourceforge.net/gff3.shtml>). There are numerous tools that support the reformatting of a variety of file types into GFF3, so generation of appropriate files by database providers would be straightforward. The availability of a wide suite of tools for downstream analyses of files in GFF3 format also means that users could combine the weight of evidence from many sources when examining a particular genome. This could reveal instances of systemic bias and therefore lead to better genomic annotations, as more composite features would be available and conflicting annotations could be highlighted for resolution.

Exchanging data, though, also relies on exploiting common standards for the generation of computational analyses and supporting data downloads is not enough, regardless of format. Data

resources should be further expected, within reason, to provide clear specifications for how the data are generated (e.g. computations like gene prediction, operon and ortholog computations, etc.). One example of this type of documentation is provided in *AboutIMG*, a web-based description of the Integrated Microbial Genomes (IMG) system²¹.

Hopefully, in the future it will be far simpler to combine various genomic features, exact details of how they were generated and enough information about the provenance (or ‘origin’) of the analyses to be able to transparently share data from multiple sources. Such interoperability, especially when provided by participating databases in a way that would enable automatic harvesting of the data (e.g. if available through web service technology), would multiply the individual value of these databases many times over and open up new opportunities to examine genome sequences in unprecedented detail.

The Future

The effort required to achieve the degree of transparency advocated here is considerable but offers significant, obvious, and immediate benefits. We argue that the cost of achieving such standardization is trivial compared to the sums spent generating the data. The capture of information in MIGS will not only facilitate comparative genomic analyses but also enhance the available descriptions of downstream ‘omic experiments based on these genomes. It will also enhance the much larger “halos” of 16S rRNA sequences that are now available for many sequenced genomes and metagenomes. For example, the genome sequence of the marine bacterium *Silicibacter pomeroyi*²² is “embedded” in a large number of environmental 16S rRNA sequences affiliated with the Roseobacter lineage that is accompanied by a fairly extensive literature describing the distribution, ecology, and other properties of this group²³.

With its ongoing efforts, the GSC hopes to stimulate interest in and provide a viable mechanism for the capture and analysis of additional stores of genomic metadata. The GSC has a standing call for community participation, is keen to modify MIGS in response to constructive suggestions from researchers worldwide, solicit MIGS compliant genome reports (including batch uploads) and collect relevant controlled vocabulary terms useful in the description of genome sand metagenomes. GCat identifiers have recently been implemented and are available for past or future projects¹⁷ and MIGS-compliant genome reports are starting to become available online (e.g.²⁴⁻²⁷). We expect a production version of MIGS (2.0) to be released by the end of 2007 with an appropriate set of terms formalized within OBI¹⁹. We would hope this milestone would be accompanied by recognition of MIGS by journals and implementation of MIGS by a variety of databases. Beyond this milestone, the ‘stable’ MIGS specification should still remain flexible enough to allow it to be updated in accordance with advances in technology and our biological knowledge of the natural world. The most up-to-date information about GSC activities can always be found at our website (gensc.sf.net).

Acknowledgements

We would like to thank NIEeS and the European Bioinformatics Institute (EBI) for hosting the first three GSC workshops and NERC for providing funds for co-ordination and infrastructure building activities.

Opinions, findings and conclusions or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

References

1. Overbeek, R. et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691-5702. Print 2005. (2005).
2. Zhang, K. et al. Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology* **24**, 680 - 686 (2006).
3. Coenye, T. & Vandamme, P. Bacterial whole-genome sequences: minimal information and strain availability. *Microbiology* **150**, 2017-2018 (2004).
4. Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N. & White, O. Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* **21**, 293-306 (2005).
5. Lombardot, T. et al. Megx.net--database resources for marine ecological genomics. *Nucleic Acids Res.* **34**, D390-393. (2006).
6. Tautz, D., Arctander, P., Minelli, A., Thomas, E. & Vogler, A.P. A plea for DNA taxonomy *Trends in Ecology and Evolution* **18**, 70-74 (2003).
7. Edwards, R.A. et al. Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics.* **7**, 57. (2006).
8. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* **437**, 376-380. Epub 2005 Jul 20 2005. (2005).
9. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat Rev Genet.* **5**, 335-344. (2004).
10. Garrity, G.M. (ed.) *Bergey's Manual of Systematic Bacteriology*, Vol. 1, Edn. 2nd. (Springer-Verlag, New York; 2001).
11. Field, D. et al. Meeting Report: eGenomics: Cataloguing Our Complete Genome Collection I. *Comparative and Functional Genomics* **6**, 357-362 (2006).
12. Field, D. & Hughes, J. Cataloguing our current genome collection. *Microbiology* **151**, 1016-1019 (2005).
13. Pace, N.R. Time for a change *Nature.* **441** 289. (2006).
14. Sansone, S.A. et al. A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. *Omic* **10**, 164-171 (2006).
15. Taylor, C., Field, D. & Sansone, S.-A. et. al. MICheck: A Minimum Information Checklist Resource. *Nat Biotechnol* (in review).
16. Ward, N., Eisen, J., Fraser, C. & Stackebrandt, E. Sequenced strains must be saved from extinction. *Nature* **414**, 148 (2001).
17. Field, D. et al. Meeting Report: eGenomics: Cataloguing our Complete Genome Collection III. *Comp Funct Genom* (2007).
18. Morrison, N. et al. Concept of sample in OMICS technology. *Omic* **10**, 127-137 (2006).
19. Whetzel, P.L. et al. Development of FuGO: an ontology for functional genomics investigations. *Omic* **10**, 199-204 (2006).
20. Cochrane, G. et al. Evidence standards in experimental and inferential INSDC Third Party Annotation data. *Omic* **10**, 105-113 (2006).
21. Markowitz, V.M. et al. The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34**, D344-348. (2006).
22. Moran, M.A. et al. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature.* **432**, 910-913. (2004).
23. Buchan, A., Gonzalez, J.M. & Moran, M.A. Overview of the marine roseobacter lineage. *Appl Environ Microbiol.* **71**, 5665-5677. (2005).
24. Angly, F.E. et al. The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368. (2006).

25. Bauer, M. et al. Whole genome analysis of the marine Bacteroidetes 'Gramella forsetii' reveals adaptations to degradation of polymeric organic matter. *Environ Microbiol.* (2006).
26. Glockner, F.O. et al. Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. *Proc Natl Acad Sci U S A* **100**, 8298-8303 (2003).
27. Rabus, R. et al. The genome of Desulfotalea psychrophila, a sulfate-reducing bacterium from permanently cold Arctic sediments. *Environ Microbiol.* **6**, 887-902. (2004).