An Open-Access, Standards-Supportive Publication that Rapidly Disseminates Concise Genome and Metagenome Reports in Compliance with MIGS/MIMS Standards

Scott H. Harrison[1]*, Sam V. Angiuoli[2], Patrick S. G. Chain[1,3,4,5], Dawn Field[6], Frank-Oliver Glöckner[7], Lynette Hirschman[8], Eugene Kolker[9,10], Nikos Kyrpides[4], Oranmiyan W. Nelson[1], Susanna-Assunta Sansone[11], Lynn M. Schriml[2], Peter Sterk[6,11], David W. Ussery[12], Owen White[2], **George M. Garrity[1,5]**

[1] Microbiology & Molecular Genetics, Michigan State University, East Lansing, Michigan; * = sharrison@standardsingenomics.org; [2] Institute for Genome Sciences and Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, Baltimore, Maryland; [3] Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, California; [4] U.S. Department of Energy Joint Genome Institute, Walnut Creek, California; [5] Center for Microbial Ecology, Michigan State University, East Lansing, Michigan; [6] NERC Centre for Ecology and Hydrology, Oxford, United Kingdom; [7] Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany; [8] Information Technology Center, The MITRE Corporation, Bedford, Massachusetts; [9] Seattle Children's Hospital Research Institute, Seattle, Washington; [10] Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Information, University of Washington, Seattle, Washington; [11] EMBL—European Bioinformatics Institute, Cambridge, United Kingdom; [12] Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark

Standards in Genomic Sciences

An Open Access Journal of the Genomic Standards Consortium

FOUNDING MEMBERS

Sam Angiuoli
Baltimore, MD, USA

Patrick Chain
Walnut Creek, CA, USA

Dawn Field
Oxford, UK

George M. Garrity
East Lansing, MI, USA

Frank Oliver Glöckner
Bremen, DE

Lynette Hirschman
Bedford, MA, USA

Eugene Kolker
Seattle, WA, USA

Nikos Kyrpides
Walnut Creek, CA, USA

Susanna-Assunta Sansone
Cambridge, UK

Lynn Schriml
Baltimore, MD, USA

Peter Sterk
Cambridge, UK

David Ussery
Lingby, DK

Owen White
Baltimore, MD, USA

EDITORIAL OFFICE

6160 Biomed Phys Sci
Michigan State Univ.
East Lansing, MI 48824

EDITORIAL STAFF

George M. Garrity
Editor-in-Chief

Scott H. Harrison
Managing Editor

Oranmiyan Nelson
Production Editor

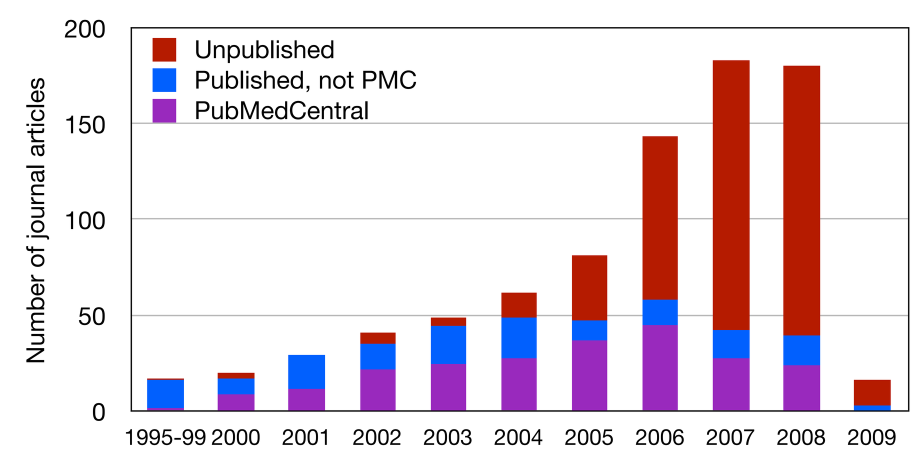
MICHIGAN STATE UNIVERSITY

Abstract

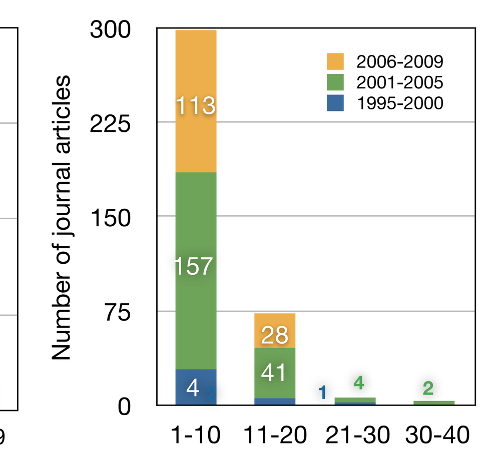
The Standards in Genomic Sciences eJournal (SIGS; http://standardsingenomics.org) is a newly established standards-supportive publication being developed to report on the exponentially increasing volume of genomic and metagenomic data[1]. SIGS will also include coverage of detailed standard operating procedures, meeting reports, reviews and commentaries, data policies, and white papers and other gray literature that are relevant to genomic sciences, but absent from the scholarly literature. To achieve standards compliance, SIGS has been designed to support the specification for the minimum information about a genome sequence (MIGS)[2]. MIGS is developed and maintained by the Genomic Standards Consortium (GSC), a group formed to promote the development of standardized annotations of genomic investigations. As a specification and checklist, MIGS has helped inspire formal models for gathering information routinely included in primary publications such as environmental context, biotic relationship, relationship to oxygen, source material identifier(s), nucleic acid sequence and sample metadata as well as the sequencing assay and genome annotation protocol. MIGS has been developed to address the genomic information of bacteria, archaea, eukaryotes, plasmids, viruses, and organelles. MIGS has been extended into a specification of minimum information about a metagenome sequence (MIMS). SIGS is the first standards-based journal in the life sciences and will provide open access to its content from the outset, on the principle that making research freely available supports a global exchange of knowledge to better accelerate scientific discovery.

Background, Material and Methods

Background: Starting with the publication of the complete genome sequence of *H. influenzae* in 2005, 821 bacterial and archaeal genomes are now publicly available with assigned RefSeq numbers. Since 2005, the number of genome sequences for which there is not a corresponding publication has grown, and the number of genome reports released under the open access model of PubMedCentral is declining. Page numbers for genome reports have been variable across the literature, ranging from two-page interpretative reports that exclude background data summaries to full descriptions that include detailed protocols. While genome reports in the literature will typically mention various details of genome sequencing, assembly, annotation and interpretation; coverage is non-uniform and often poor for basic aspects of sequence similarity, inferred protein structures, and organism phenotype and habitat; all of which have utility for further analysis and cross-comparison. The absence of associated journal articles for viral, plasmid and organelle replicons is even more pronounced. Although most eukaryotic genomes have an associated publication, issues of uniformity and coverage remain. In the case of metagenomics, we are finding that each aspect of an associated biosystem may not fully correspond to the simple content model of the standards specification, and our editorial workflow is being built to encourage comments by authors to address this during peer review and help stimulate further updates to the proposed MIGS/MIMS standards.



Number of published journal articles per year relating to the complete genome sequences of bacteria and archaea. Journal reference data is from those articles listed in gbk files for the primary chromosomes of each of the 821 bacterial and archaeal organisms (data retrieved January 2009).

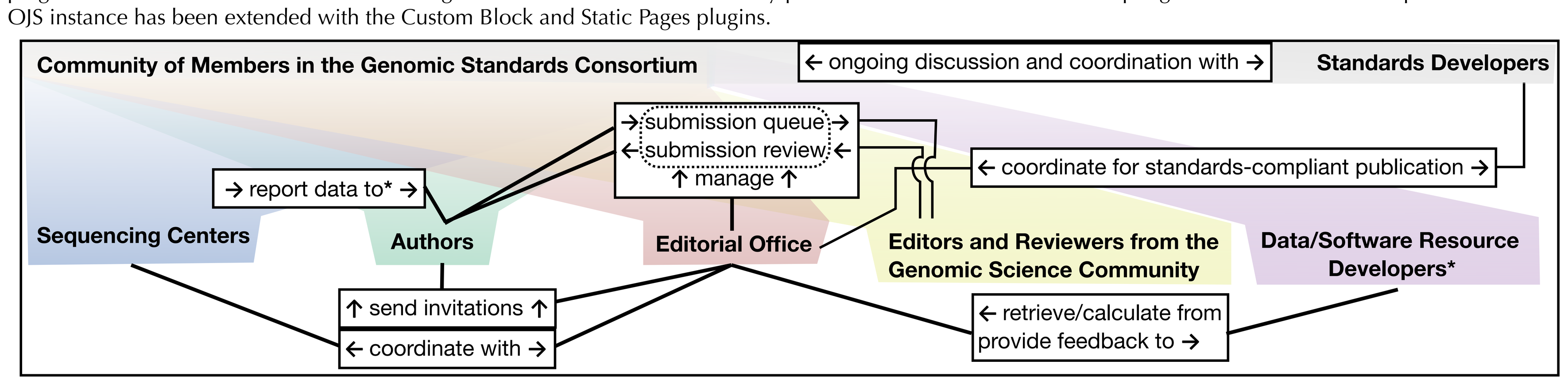


Length in pages of journal articles for bacterial and archaeal genomes for three different time periods. Based on page numberings indicated in gbk files.

Search word	# articles
protein	190
COG	77
paralog	90
phylogenetic tree	47
transmembrane	80
RNA	143
-philic	65
spor-	39
oxygen	86
electron receptor	`1
habitat	51
type strain	48

Number of genome sequence-related articles in PubMedCentral with specific search words. n=190 (data retrieved January 2009).

Material and Methods: Protocols for retrieval and calculation are based on a suite of 19 perl scripts that utilize GenBank data and metadata, NCBI COGs, EMBL calculation tools (SMART; for signal peptides and transmembrane domains), and an in-house paralog calculation and clustering approach on a 12-CPU Mac OS 10.4 XServe cluster powered by Sun Grid Engine 6.0 with a manually compiled version of NCBI C Toolkit version 6.1. Values retrieved and computed are numbers and percentages for: replicon sizes (bp), GC density (bp), coding density (bp), total genes, RNA genes, protein-coding genes, hypothetical proteins, protein coding genes in paralog clusters, conserved domain counts, and gene assignments to COGs. For genomes with replicons other than just a chromosome, percentages are counted both per replicon and per genome. Resources from the Ribosomal Database Project (http://rdp.cme.msu.edu) and the Taxonomic Outline of Bacteria and Archaea (http://taxonomicoutline.org) are used respectively for providing authors with a phylogenetic tree and taxonomic ranking. Basic comparisons of features for fully sequenced genomes within and across families, batch-retrieved and processed set of fully sequenced genomes having RefSeq accession numbers are also provided. The editorial workflow utilizes Open Journal Systems (OJS) web application, version 2.2.2.0 from the Public Knowledge Project (http://pkp.sfu.ca/ojs), because that application supports custom-specification of checklists that can be integrated into the submission and peer review process, and has an extensible plugin architecture with hook callback and registration features. We have already pilot-tested the software with a sampling of content. Our current production-level OJS instance has been extended with the Custom Block and Static Pages plugins.



Editorial workflow - the editorial office aids authors in the genesis, submission and review of their manuscript, coordinating efforts with sequencing centers, using existing data and software resources for infrastructure and generation of genome features, preliminary comparisons based on cladistic and pangenomic assays, and providing peer review from the genomic science community. * = Data resource centers receive data after it is sent to authors by the sequencing center.

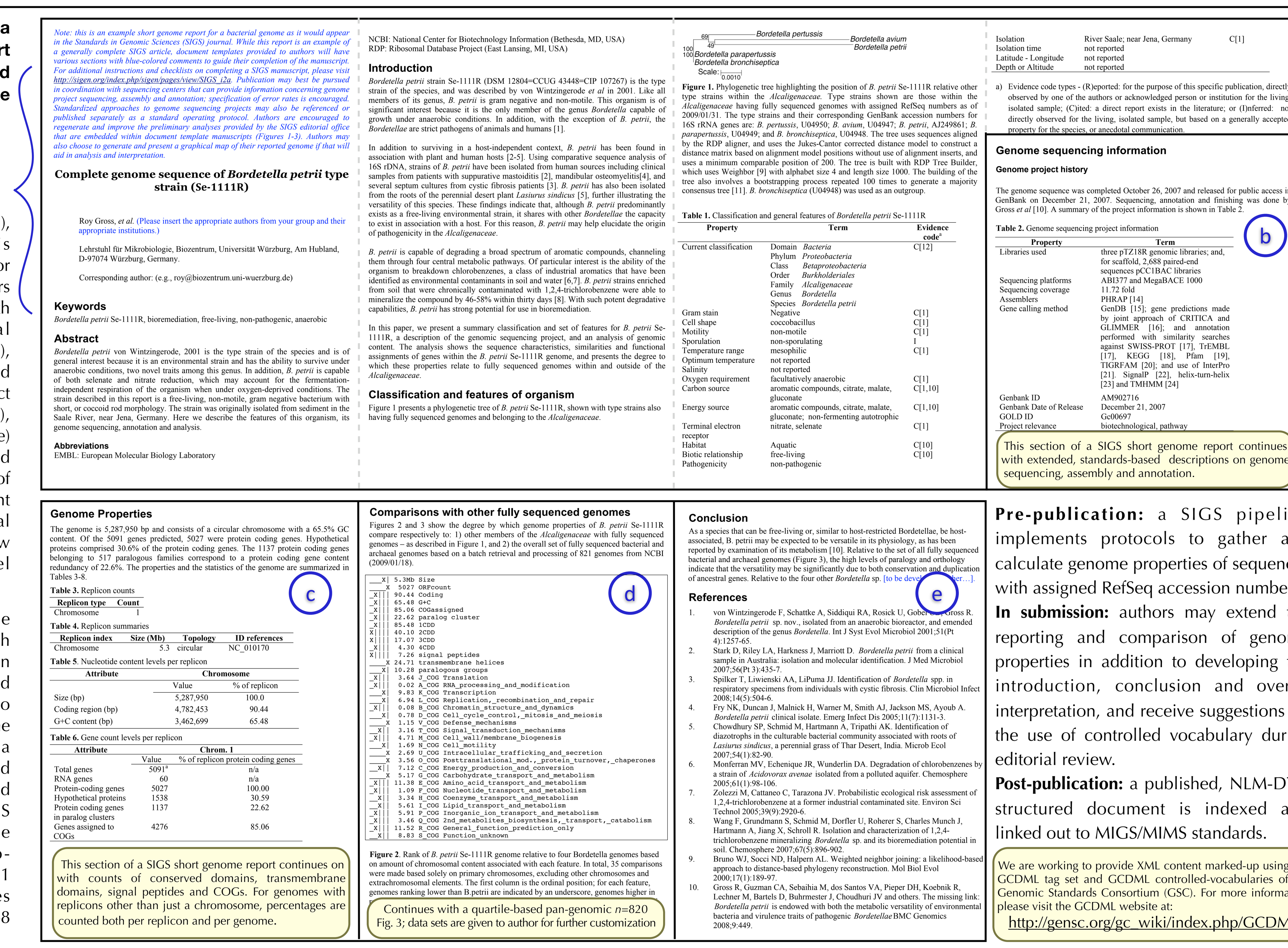
Results

While data sets grow, new research goals emerge, and database structures and applications change, SIGS provides static, archival records of genomic data and metadata enriched with interpretation and subject to peer review. Authors get a document template that they develop into a report with input from their sequencing center and aid from the SIGS editorial office.

The anatomy of a short genome report in SIGS (a-e), and genesis of a pipeline and workflow.

After gathering information for latter parts of the manuscript (b-d), much of which is provided to the author by sequencing centers (b) coordinating with the SIGS editorial office pipeline (c-d), the author is prepared to develop an abstract and introduction (a), and make claims (e) about their reported data set in terms of how it may present significant biological features, or be a new finding of a novel clade or niche.

SIGS editors have been working with sequencing centers on standards-based approaches to reporting of genome projects (b). With a series of automated and semi-automated protocols, the SIGS editorial office provides a phylogenetic tree in Fig 1 (a), and calculates values for Tables 3-8 (c) and Figs 2-3 (d).



Pre-publication: a SIGS pipeline implements protocols to gather and calculate genome properties of sequences with assigned RefSeq accession numbers.
In submission: authors may extend the reporting and comparison of genome properties in addition to developing the introduction, conclusion and overall interpretation, and receive suggestions on the use of controlled vocabulary during editorial review.
Post-publication: a published, NLM-DTD structured document is indexed and linked out to MIGS/MIMS standards.

We are working to provide XML content marked-up using the GCDML tag set and GCDML controlled-vocabularies of the Genomic Standards Consortium (GSC). For more information please visit the GCDML website at:
http://gensc.org/gc_wiki/index.php/GCDML

Conclusions and Discussion

Our community-based effort has led to archival snapshots that pursue consistency and uniformity for future and current cross-comparisons. Where the GSC and other standards initiatives have sought to develop interoperable approaches to ensure consistency in semantic and syntactic annotation of genomes and metagenomes[3,4], SIGS will apply and extend these standards to be published. Key steps for the immediate future include the downstream flow of content into XML tagged instances of articles, based on the NLM DTD, sending out of prepared invitations to authors in coordination with sequencing centers, drafting table of content manifests for subsequent journal issues of 2009, fine-tuning the peer-review process to ensure the use of nomenclature and terminology with standing in the literature, and for implementation in minimal and extended forms of XML-based tagging approaches such as can be done for MIGS/MIMS with the Genomic Contextual Data Markup Language (GCDML)[3].

References

1. Garrity, G. M., Field, D., Kyrpides, N., Hirschman, L., Sansone, S. A., Angiuoli, S., Cole, J. R., Glöckner, F. O., Kolker, E., Kowalchuk, G., Moran, M. A., Ussery, D., and White, O. 2008. Toward a standards-compliant genomic and metagenomic publication record. *OMICS* **12**:157-60.
2. Field, D. et al. 2008. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**:541-7.
3. Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., Glöckner, F. O., and the Genomic Standards Consortium. 2008. A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* **12**:115-121.
4. Hirschman, L., Clark, C., Cohen, K. B., Mardis, S., Luciano, J., Kottmann, R., Cole, J., Markowitz, V., Kyrpides, N., Morrison, N., Schriml, L. M., Field, D. and the EnvO Project. 2008. Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata. *OMICS* **12**:129-136.

Acknowledgements

We gratefully acknowledge the Office of the Vice President for Research and Graduate Studies at Michigan State University, the Michigan State University Foundation, and funding from the Department of Energy DE-FG02-08ER64707. We also thank Brian Baer and Julius Jackson for enabling access to a bioinformatics computer cluster in Dr. Jackson's laboratory. We would also like to gratefully acknowledge the support of many members of the Genomic Standards Consortium, and those who have indicated their willingness to serve as editors, reviewers and contributors.

Timeline

A — June 2008 — B — August 2008 — C — D — October 2008 — E — F — G — December 2008 — H — I — 2009 DOE GTL YOU ARE HERE — February 2009 — J — (March 2009)

OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2008.A2B2

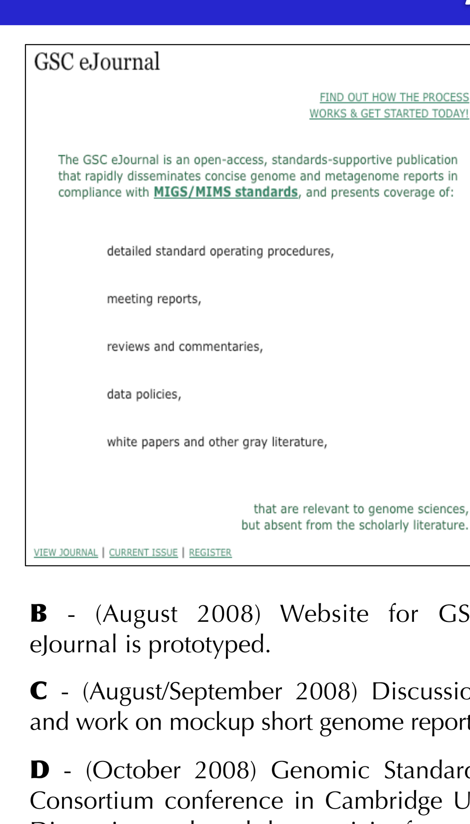
Toward a Standards-Compliant Genomic and Metagenomic Publication Record

George M. Garrity, Dawn Field, Nikos Kyrpides, Lynette Hirschman, Susanna-Assunta Sansone, Samuel Angiuoli, James R. Cole, Frank Oliver Glöckner, Eugene Kolker, George Kowalchuk, Mary Ann Moran, Dave Ussery, and Owen White

Abstract

Background

A - (June 2008) article in *OMICS* appears describing need for standards-compliant genome and metagenome reports[1].
http://dx.doi.org/10.1089/omi.2008.A2B2



GSC eJournal

FIND OUT HOW THE PROCESS WORKS & GET STARTED TODAY!

The GSC eJournal is an open-access, standards-supportive publication that rapidly disseminates concise genome and metagenome reports in compliance with **MIGS/MIMS standards**, and presents coverage of:

detailed standard operating procedures,

meeting reports,

reviews and commentaries,

data policies,

white papers and other gray literature,

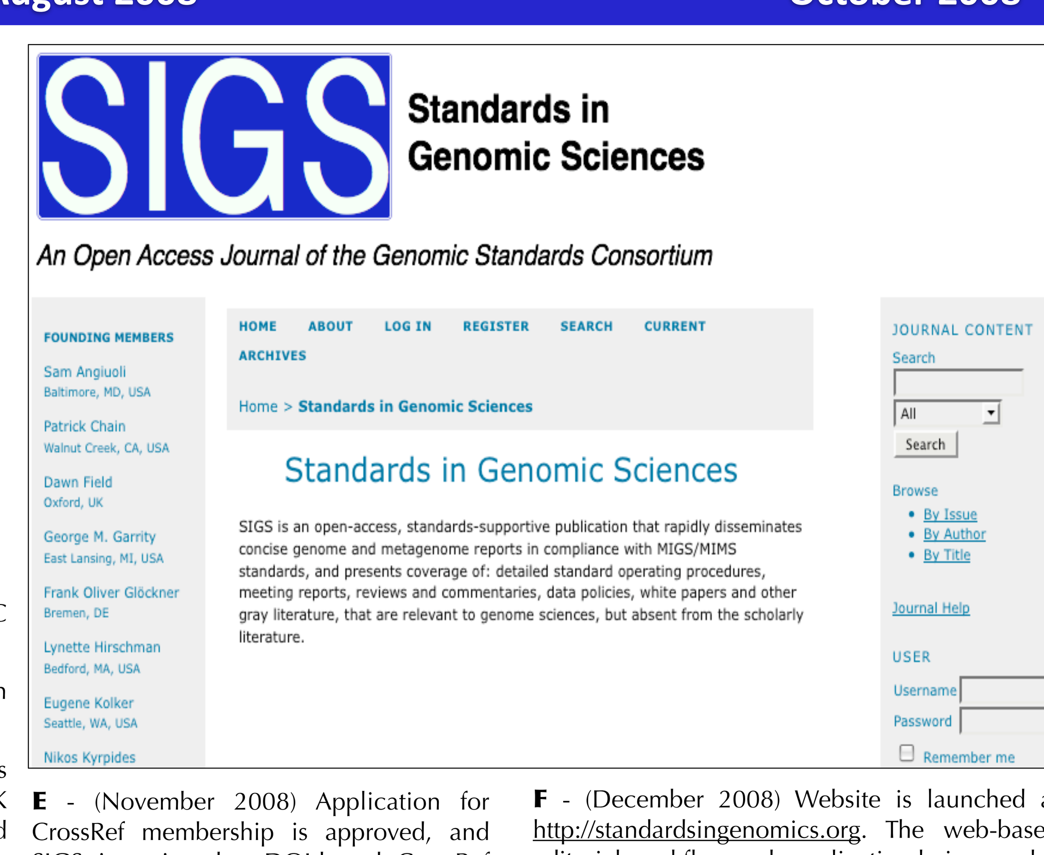
that are relevant to genome sciences, but absent from the scholarly literature.

VIEW JOURNAL | CURRENT ISSUE | REGISTER

B - (August 2008) Website for GSC eJournal is prototyped.

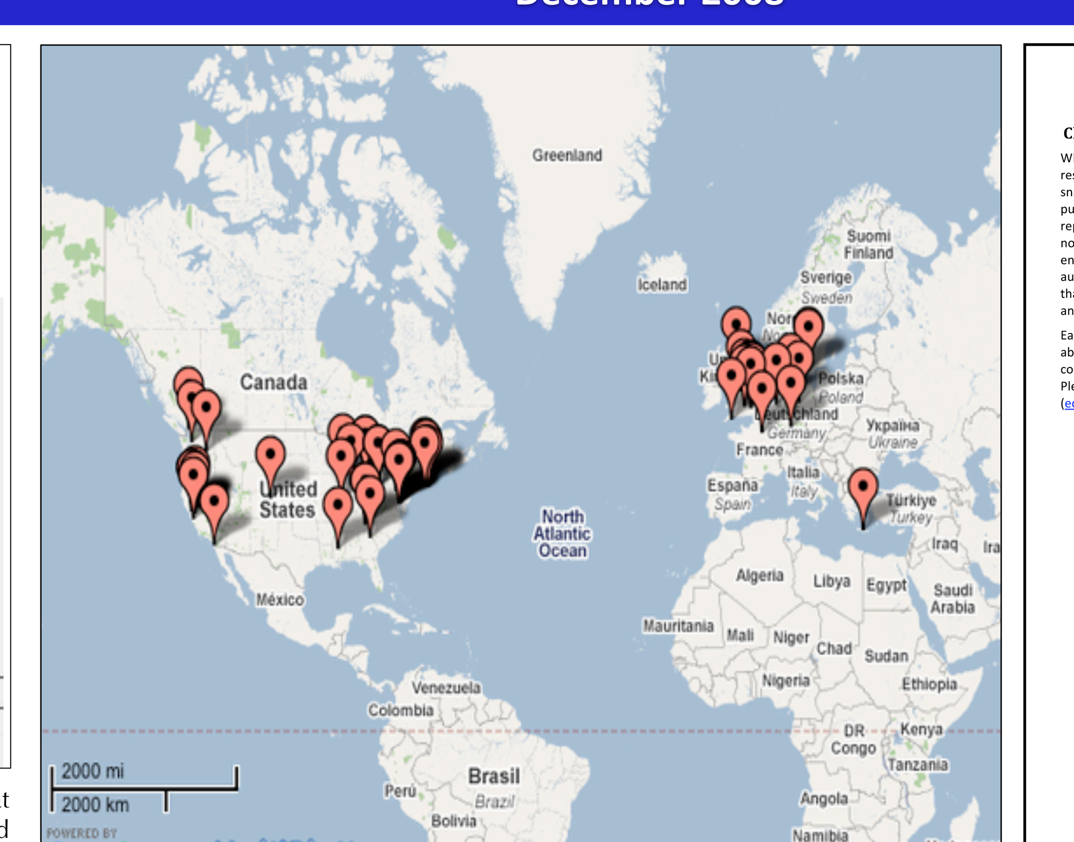
C - (August/September 2008) Discussion and work on mockup short genome report.

D - (October 2008) Genomic Standards Consortium conference in Cambridge UK Discussion and workshop activity focused on short genome reports begins, involving scientists from laboratories, sequencing and data resource centers. Process of recruiting editors begins.

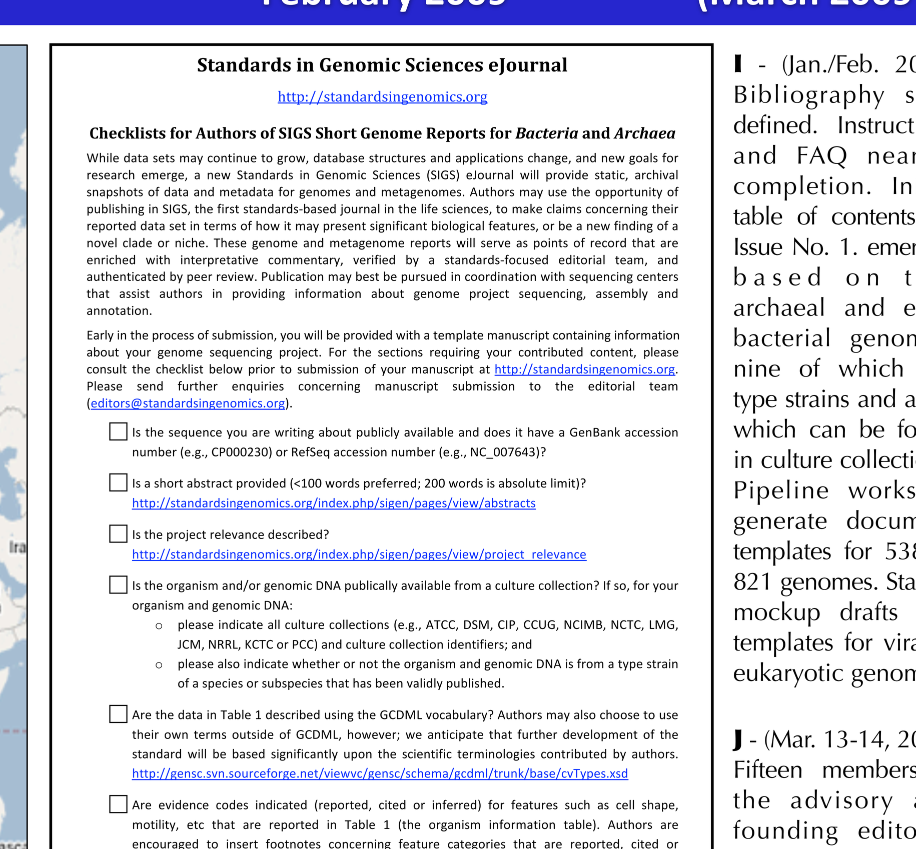


E - (November 2008) Application for CrossRef membership is approved, and SIGS is assigned a DOI-based CrossRef identifier. National Serials Data Program (NSDP) at the Library of Congress also makes an ISSN assignment: 1944-3277.

F - (December 2008) Website is launched at http://standardsingenomics.org. The web-based editorial workflow web application being used is the Public Knowledge Project's Open Journal Systems open source software. Work begins on instructions to authors as well as a journal FAQ.



G - (December 2008) The recruiting of experts from around the world to serve as editors reaches n=60. Their contributions will help SIGS be a major resource for the community.



Standards in Genomic Sciences eJournal

http://standardsingenomics.org

Checklists for Authors of SIGS Short Genome Reports for *Bacteria* and *Archaea*

While data sets may continue to grow, database structures and applications change, and new goals for research emerge, a new Standards in Genomic Sciences (SIGS) eJournal will provide static, archival snapshots of data and metadata for genomes and metagenomes. Authors may use the opportunity of publishing in SIGS, the first standards-based journal in the life sciences, to make claims concerning their reported data set in terms of how it may present significant biological features, or be a new finding of a novel clade or niche. These genome and metagenome reports will serve as points of record that are enriched with interpretative commentary, verified by a standards-focused editorial team, and authenticated by peer review. Publication may best be pursued in coordination with sequencing centers that assist authors in providing information about genome project sequencing, assembly and annotation.

Early in the process of submission, you will be provided with a template manuscript containing information about your genome sequencing project. For the sections requiring your contributed content, please consult the checklist below prior to submission of your manuscript at http://standardsingenomics.org. Please send further enquiries concerning manuscript submission to the editorial team (editors@standardsingenomics.org).

- [ ] Is the sequence you are writing about publicly available and does it have a GenBank accession number (e.g., CP000230) or RefSeq accession number (e.g., NC_007643)?
- [ ] Is a short abstract provided (<100 words preferred; 200 words is absolute limit)? http://standardsingenomics.org/index.php/sigen/pages/view/abstracts
- [ ] Is the project relevance described? http://standardsingenomics.org/index.php/sigen/pages/view/project_relevance
- [ ] Is the organism and/or genomic DNA publically available from a culture collection? If so, for your organism and genomic DNA:
  - please indicate all culture collections (e.g., ATCC, DSM, CIP, CCUG, NCIMB, NCTC, LMG, JCM, NRRL, KCTC or PCC) and culture collection identifiers; and
  - please also indicate whether or not the organism and genomic DNA is from a type strain of a species or subspecies that has been validly published.
- [ ] Are the data in Table 1 described using the GCDML vocabulary? Authors may also choose to use their own terms outside of GCDML, however; we anticipate that further development of the standard will be based significantly upon the scientific terminologies contributed by authors. http://gensc.svn.sourceforge.net/viewvc/gensc/schema/gcdml/trunk/base/cvTypes.xsd
- [ ] Are evidence codes indicated (reported, cited or inferred) for features such as cell shape, motility, etc that are reported in Table 1 (the organism information table). Authors are encouraged to insert footnotes concerning feature categories that are reported, cited or inferred as follows:

H - (January 2009) Submission checklists for genome reports and author invitation letter templates are finalized.

I - (Jan./Feb. 2009) Bibliography style defined. Instructions and FAQ nearing completion. Initial table of contents for Issue No. 1. emerges, based on two archaeal and eight bacterial genomes, nine of which are type strains and all of which can be found in culture collections. Pipeline works to generate document templates for 538 of 821 genomes. Start of mockup drafts and templates for viral & eukaryotic genomes.

J - (Mar. 13-14, 2009) Fifteen members of the advisory and founding editorial boards are scheduled for two-day workshop in East Lansing, Michigan at MSU.