Towards a standards-compliant genomic and metagenomic publication record

George M. Garrity[1], Dawn Field[2], Nikos Kyrpides[3], Lynette Hirschman[4], Susanna-Assunta Sansone[5] Samuel Anguiloi[6], James R. Cole[7], Frank Oliver Glöckner[8], Eugene Kolker[9.10.11], George Kowaluchuk [12], Mary Ann Moran[13], Dave Ussery[14], and Owen White[6]

[1]Microbiology & Molecular Genetics, Michigan State University, East Lansing, MI, garrity@msu.edu, (517) 432 2459

[2] NERC Centre for Ecology and Hydrology, Oxford, UK, dfield@ceh.ac.uk

[3]US. Department of Energy Joint Genome Institute, Walnut Creek, CA, nckyrpides@lbl.gov

[4]Information Technology Center, The MITRE Corporation, Bedford, MA, lynette@mitre.org

[5]EMBL – European Bioinformatics Institute, Cambridge, UK, sansone@ebi.ac.uk

Abstract

Increasingly we are aware as a community of the growing need to manage the avalanche of genomic and metagenomic data, in addition to related data types like ribosomal RNA and barcode sequences, in a way that tightly integrates contextual data with traditional literature in a machine-readable way. It is for this reason that the Genomic Standards Consortium (GSC) formed in 2005. Here we suggest that we move beyond the development of standards and tackle standards-compliance and improved data capture at the level of the scientific publication. We are supported in this goal by the fact that the scientific community is in the midst of a publishing revolution. This revolution is marked by a growing shift away from a traditional dichotomy between "journal articles" and "database entries" and an increasing adoption of hybrid models of collecting and disseminating scientific information. With respect to genomes and metagenomes and related data types, we feel the scientific community would be best served by the immediate launch of a central repository of short, highly structured "Genome Notes" that must be standards-compliant. This could be done in the context of an existing journal, but we also suggest the more radical solution of launching a new journal. Such a journal could be designed to cater to a wide range of standards-related content types that are not currently centralized in the published literature. It could also support the demand for centralizing aspects of the 'gray literature' (documents developed by institutions or communities) such as the call by the GSCl for a central repository of Standard Operating Procedures describing the genomic annotation pipelines of the major sequencing centers. We argue that such an "eJournal", published under the Open Access paradigm by the GSC, could be an attractive publishing forum for a broader range of standardization initiatives within, and beyond, the GSC and thereby fill an unoccupied yet increasingly important niche within the current research landscape.

Keywords: open-access publishing, standards, genomics, metagenomics, curation, metadata

Background

Modern biology is rapidly evolving into a data-driven discipline, much like physics and chemistry, and modern biologists are increasingly dependent on tools for analyzing and visualizing ever-larger data sets. Nowhere is this better seen than in the domain of genomics. Genomes and metagenomes are being sequenced at an ever-increasing pace and the emergence of ultra high throughput sequencing technologies will only accelerate the production of vast quantities of data. The Genomic Standards Consortium (GSC) formed in September 2005 to work as an international community towards solutions for improving the descriptions of our complete collection of genomes and metagenomes and mechanisms of data exchange and integration. As a first step, the GSC published the "Minimum Information about a Genome Sequence" (MIGS) specification, which describes the core information that should be reported with each new genome or metagenome publication (Field et al, 2008). Increasingly though, we are aware as a community that the publication of ever larger numbers of such data sets is either significantly delayed or are declined outright by both top-tier and specialist journals. Now that this field is reaching maturity, we suggest that the community could broadly adopt a 'short form' of a standardized genome/metagenome publication. We further argue that such genome notes could be centralized to maximize their value to the community. More radically, if this were done in the context of a new publishing forum it could open the door to a new range of possibilities for centralizing standards-supportive literature. We explore this option further, from a GSC context, in this article.

An explosion of genomes and metagenomes

There are already more than 650 completed genome sequences in the public domain, with hundreds more in various stages of completion around the world (Liolios et al, 2008). This number will increase dramatically as the cost of sequencing diminishes and the pace of sequencing accelerates. Ironically, as the field is ramping up, the number of peer-reviewed publications describing those genomes has declined. This is expected for several reasons. First, data can be generated far more quickly than it can be published. Second, both top-tier and specialist journals are only interested in the most exceptional full papers on genomes given the maturity of this field and ever-present competition from other types of papers, including those describing new technologies. This means an increasing number of genomes and metagenomes are currently only present in the public record as an INSDC entry. For example, at present, there is no journal publication for 20% of completed genomes in the Genomes Online Database (GOLD) (Liolios et al, 2008) compared to 6% just three years ago (Figure 1). This trend will only increase because genome sequencing is applied to a wider range of biological problems as just one part of the routine 'laboratory toolkit' available to the majority of researchers.

This is a fully expected scenario given the success of genome sequencing and follows the natural course of events for all data types generated with once 'new' technologies. Althought every sequences should be submitted to the International Nucleotide Sequence Databases Collaboration (INSDC) as part of the public record and can be retrieved and cited by its corresponding INSDC Genome Project Identifiers. Still, the value of those sequences is diminished, compared to published genome sequences, which are associated with a larger quantity and higher quality of metadata. Further they are not citable in the reference list of other papers in the same way as is a publication, making those sequences and the corresponding data less easily discovered. Overall,

the lack of an associated ‘context’ deeply erodes the value of the complete record of genomes and metagenomes.

A possible solution is to publish all scientifically worthy genomes in at least a highly-reduced but standardized ‘Note’ form. This was the recommendation of a recent EU/US strategy meeting on the future of cyberinfrastructure in microbial ecology that involved authors of this article (DF, JC, FOG, MAM). We highly commend the fact that this short-form solution has already been adopted by the Journal of Bacteriology (Foote, S.J. et al., 2008) and hope to see other journals adopt it. Now that the time for these ‘concise and practical’ version of genomes reports has come, we further argue that the centralization of all such ‘notes’ would further bring another level of utility and benefit. This could be done in a virtual (electronic) way, for example through the ‘hub’ mechanism of the journal PloS One for harvesting context from journals, or through the launch of this type of publication format by a journal taking genomes from all taxa.

Ideally, centralization of these Notes would lead to far greater expectations of standardization bringing maximum benefit to the community who might use them. In particular, such reports should contain a minimum of information describing the sequence, its origin (including details of the environment), how to obtain the biological material, the sequencing methodology, and the methods used to annotate the sequence(s). All sequences must be submitted to the INSDC and and include information on relevant protocols and standard operating procedures used in the generation and annotation of the data. Details of relevant electronic databases containing the genome/metagenome, beyond the INSDC, should also be documented.

This content directly mirrors what is required for compliance with the "Minimum Information about a Genome Sequence" (MIGS) specification developed by the GSC (Field et al 2008). Such structured genome/metagenome notes would therefore work to uphold a new community standard for richness of reporting. These brief descriptions would also greatly aid downstream computational analyses if they were highly structured such that they became machine-readable, in particular for the purpose of text-mining, the automatic extraction of information and exchange of data (cross reference to GCDML paper and the GRS paper in the special issue?). Additionally, common 'minimum information' features could also be stored in a tightly integrated databases such as the INSDC, GOLD (Liolios et al 2008), and the GSC's Genome Catalogue (Field et al 2008). Such papers, if successfully enriched with common features, would blur the boundaries between traditional publications and database entries and offer a more environmental and organism-centric entry point for the sequence data.

Since these 'notes' are so minimal they would not preclude applying this rich, condensed form of reporting to all published genomes to bring each one up to date and to bind their original publications with suitable contextual data. Such Notes are brief enough to serve, post major publication, as an extended and updated database entry form that would not conflict with the original report, post-publication. Whether applied to published or unpublished data, this model could serve as a new and badly needed mechanism for providing authorship credit to those who undertake the curation of these sequences. Such a publishing model would also enable high quality sequences with little downstream analysis to be generated and published quickly with the express aim of providing them to the wider community. This proposed model of publishing genomes and metagenomes will only gain in relevance over the next decade, especially with the advent of ultra

high-throughput sequencing technologies. This would be ideal, for example, for 'mega-sequencing' projects of the future generated with private or public funding to be delivered quickly to the wider community.

Centralized, citable 'standards-supportive' content: prospects for a dedicated journal

This special issue of OMICS contains a call for the establishment of a central repository of Standard Operating Procedures (SOPs) describing genomic annotation pipelines (Angiuoli, S.A. et al., 2008). Sequencing centers and other high-throughput data providers routinely use SOPs to standardize workflow and ensure quality control and yet these documents, critical to the interpretation of public annotations, have never been properly centralized despite the obvious benefits of doing so. Making these documents more accessible in a central location serves as a key step towards further standardizing genomic annotations (Angiuoli, S.A et al, 2008).

This call for an SOP repository, with buy in from major sequencing centers, prompts us to explore if there would be room in the currently crowed arena of academic journals for a *new* journal. Such a journal could be dedicated to supporting not only Genome and Metagenome Notes and related scientific papers, but also 'gray literature' such as SOPs. We can envision a range of 'standards-supportive' content emerging from the community in the future. With respect to scientific content, scope could extent to any type of article dealing with standards developments, analytical methods with a special emphasis on matters pertaining to curation and quality control of data, experiences and improvement of the use of ontologies and controlled vocabularies in

biological applications. It could also extend to large-scale computational analyses that consume contextual data associated with standards or which present large amount of novel curated data describing the genome or metagenome collection. With respect to aspects of the "g*ray literature*" (defined by being published by organizations and communities other than scholarly publishers (Adler et al, 2006, Mathews et al, 2004), like SOPs, a wide variety of content types could be imagined, for example, the data policies of the major funding agencies supporting 'omic research. Key documents in the gray literature are broadly useful to the community but are rarely centralized properly and usually are not indexed. These publications tend to be highly dispersed and rapidly lost (much like supplementary data), despite their overall importance to the scientific community.

It is certainly now possible, under the widely praised Open Access (OA) publishing model (Suber, P., 2007) for a society such as the GSC to launch a cost-efficient OA journal by distributing it only in electronic format and by using an off-the-shelf open source platform such as the Open Journal Systems (OJS) editorial environment (Willinsky, J., 2005). This model would further support integration with downstream resource and key databases and increase usage of the context by others.

Such a journal should explore further, the possibilities of integrating traditional publications with contextual data and supplementary information. The peer-reviewed literature is undergoing a radical transformation (Anonymous, 2003; Bourne, P., 2005; Suber, P., 2007; Ware, M., 2006, Ceol, A., 2008) and the future will be dominated by systems that can integrate the traditional concepts of journal articles and databases to create new, more user-friendly resources (Bourne,

P., 2005; Vastrik, I. et al., 2007) that best serve up the vast quantities of data that are accumulating.

Launching a new journal must be weighed up carefully given the recent explosion of OA journals. Such a journal could find the widest relevance in the community if launched as an OA publishing platform for the entire standardization community. The content of such a journal could extend to include contributions by the wide range of grass-roots standardization activities that now exist. The rapid accumulation of vast stores of 'omic data and the need for data integration has led to over 20 "Minimum Information" checklist projects being registered in the "Minimum Information about a Biomedical or Biological" (MIBBI) portal (http://mibbi.sf.net). Several groups participate in synergistic activities as part of the Functional Genomics (FuGE) project (Jones, A.R. et al., 2007), underpinning the XML-based formats (http://fuge.sf.org) they have developed. Other standards initiatives have sprung up from a growing number of communities that work collaboratively on a common tabular framework for presenting the experimental metadata (ISA-TAB, http://isa-tab.sf.net) and on biological ontologies. At present, over 60 groups participate under the OBO Foundry umbrella (http://www.obofoundry.org), with the objective of developing interoperable ontologies (Rubin, D.L. et al., 2006; Smith, B. et al., 2007).

Conclusions

Like the broad field of genomics, scholarly publishing is rapidly evolving. Here we advocate the tighter coupling of genomic and metagenomic datasets with standards and electronic databases be established at the publication stage. We further suggest that applying this principle to the description of genomes and metagenomes is just a beginning and that a wider range of 'standards-supportive' literature will be generated in the future. This could easily merit the creation of a dedicated OA journal that helped supported both scientific progress and the consensus-building activities of a wide-range of grass-roots standards bodies. Such a journal could be both standards compliant and standards enabled to a high level, and as such would provide a research asset for bioinformaticians and computer scientists interested in topics such as natural language processing, semantics, nomenclature, automated data-harvesting methods and in depth data integration.

References

ADLER, P. BROOKS, M.J. BLIXRUD, J. JOSEPH, H. SEGURA, S. GEORGE, L.A. (2006) To Stand the Test of Time. Long-term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe. National Science Foundation, Arlington VA, Septermber 26-27, 2007

ANDERSON, N. R., TARCZY-HORNOCH, P. and BUMGARNER, R. E. (2006). On the persistence of supplementary resources in biomedical publications. BMC Bioinformatics, 7, 260.

ANGIUOLI, S. A., GUSSMAN, A., KLIMKE, W., COCHRANE, G., FIELD, D., GARRITY,, et al. (2008). Towards an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. OMICS, in press.

ANONYMOUS. (2003). Economic analysis of scientific research publishing- A report commissioned by the Wellcome Trust, pp. 30. The Wellcome Trust, London.

BOURNE, P. (2005). Will a biological database be different from a biological journal? PLoS Comput Biol, 1(3), 179-181.

BROOKSBANK, C. and QUACKENBUSH, J. (2006). Data Standards: A Call to Action. OMICS: A Journal of Integrative Biology, 10(2), 94-99.

CEOL, A., CHATR-ARYAMONTRI, A., LICATA, L., and CESARNI, G. (2008) Linking entries in the protein interaction database to structured text: The FEBS Letters experiment. FEBS Letters 582

FIELD, D., GARRITY, G. M., GRAY, T., MORRISON, N., SELENGUT, J., STERK, P., et al. (2008). Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. Nature Biotechnology, in press.

FIELD, D., GARRITY, G. M., MORRISON, N., SELENGUT, J., STERK, P., TATUSOV, T., et al. (2005). eGenomics: Cataloguing Our Complete Genome Collection. Comp. Funct. Genom., 6, 363-368.

FIELD, D. and HUGHES, J. (2005). Cataloguing our current genome collection. Microbiology, 151(Pt 4), 1016-1019.

FIELD, D., MORRISON, N., SELENGUT, J. and STERK, P. (2006). Meeting report: eGenomics: Cataloguing our Complete Genome Collection II. OMICS, 10(2), 100-104.

FIELD, D. and SANSONE, S.-A. (2006). A Special Issue on Data Standards. OMICS: A Journal of Integrative Biology, 10(2), 84-93.

FIELD, D. and SANSONE, S. A. (2006). A special issue on data standards. OMICS, 10(2), 84-93.

FOOTE, S. J., BOSSE, J. T., BOUEVITCH, A. B., LANGFORD, P. R., YOUNG, N. M. and NASH, J. H. (2008). The complete genome sequence of Actinobacillus pleuropneumoniae L20 (serotype 5b). J Bacteriol, 190(4), 1495-1496.

JONES, A. R., MILLER, M., AEBERSOLD, R., APWEILER, R., BALL, C. A., BRAZMA, A., et al. (2007). The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol, 25(10), 1127-1133.

KYRPIDES, N. C. (1999). Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. Bioinformatics, 15(9), 773-774.

LIOLIOS, K., MAVROMATIS, K., TAVERNARAKIS, N. & KYRPIDES, N. (2008). The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research, 36((Database issue)), D475-479.

MATTHEWS, B.S. (2004). Gray literature: Resources for locating unpublished research College and Research Library News. 65:3

PENNISI, E. (2008). DNA data. Proposal to 'Wikify' GenBank meets stiff resistance. Science, 319(5870), 1598-1599.

QUACKENBUSH, J. (2004). Data standards for "omic. science. Nat Biotechnol, 22, 613.

RUBIN, D. L., LEWIS, S. E., MUNGALL, C. J., MISRA, S., WESTERFIELD, M., ASHBURNER, M., et al. (2006). National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS, 10(2), 185-198.

SMITH, B., ASHBURNER, M., ROSSE, C., BARD, J., BUG, W., CEUSTERS, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol, 25(11), 1251-1255.

SUBER, P. (2007). Open access in 2007. In SPARC Open Access Newsletter (ed. SUBER, P.), pp. 14. SPARC, the Scholarly Publishing and Academic Resources Coalition, Washington, DC.

VASTRIK, I., D'EUSTACHIO, P., SCHMIDT, E., JOSHI-TOPE, G., GOPINATH, G., CROFT, D., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biol, 8(3), R39.

WARE, M. (2006). Scientific publishing in transition: an overview of current developments, pp. 30. Mark Ware Consulting Ltd.

WILLINSKY, J. (2005). Open Journal Systems: A example of Open Source Software for journal management and publishing. Library Hi-Tech, 23(4), 504-519.

Garrity et al Figure 1.

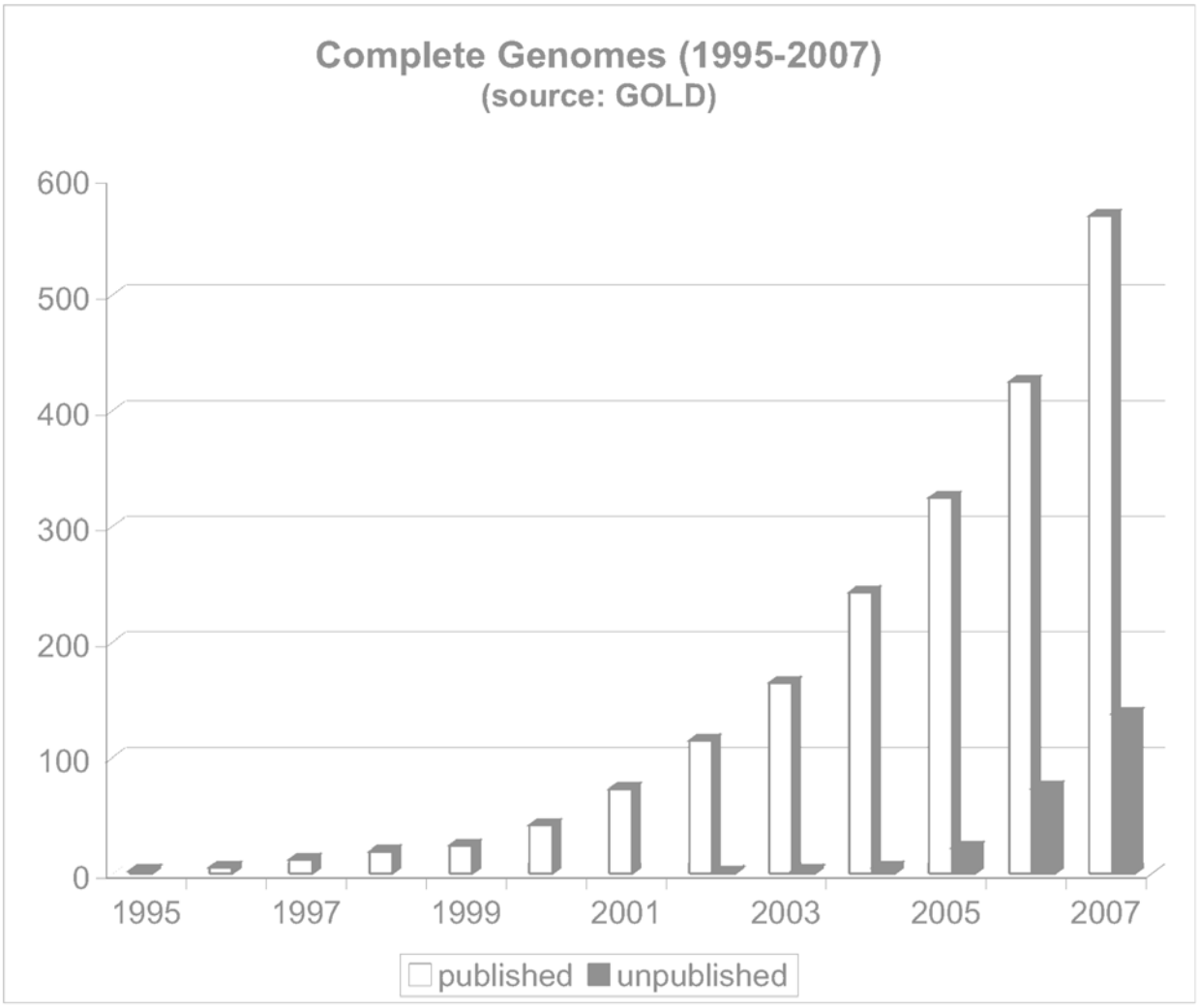

Figure 1. As sequencing of genomes and metagenomes becomes more commonplace, the practice of publishing a companion "genome paper" in the scientific literature has begun to decline. In 2007, approximately 20% of all completed genomes were without such a publication.