OMICS A Journal of Integrative Biology
Volume 10, Number 2, 2006
© Mary Ann Liebert, Inc.

Meeting Report: eGenomics: Cataloguing Our Complete Genome Collection II

DAWN FIELD,[1] NORMAN MORRISON,[2,3] JEREMY SELENGUT,[4] and PETER STERK[5]

ABSTRACT

This article summarizes the proceedings of the "eGenomics: Cataloguing our Complete Genome Collection II" workshop held November 10–11, 2005, at the European Bioinformatics Institute. This exploratory workshop, organized by members of the Genomic Standards Consortium (GSC), brought together researchers from the genomic, functional OMICS, and computational biology communities to discuss standardization activities across a range of projects. The workshop proceedings and outcomes are set to help guide the development of the GSC's Minimal Information about a Genome Sequence (MIGS) specification.

This paper is part of the special issue of OMICS on data standards.

INTRODUCTION

GENOMIC SEQUENCING TECHNOLOGIES continue to generate data at an exponential rate, and this is only set to increase as the application of ultra-high-throughput methods becomes more commonplace. This wealth of data poses both new opportunities and immense challenges. The traditional approach to genomic sequencing has been on a per "isolate" basis. However, an increasing number of "genomes" are now being sequenced that represent not only uncultivated organisms but also populations and communities from environmental samples (metagenomes). Clearly, for adequate interpretation of this type of data, simply recording only the most basic information is no longer sufficient.

Our workshop, entitled "eGenomics: Cataloguing our Complete Genome Collection II," explored the need to capture a richer set of information (meta-data) about our complete genome collection. The development of a new genomic standard would ensure that those generating genomes contribute to the quality and quantity of metadata available (Field and Hughes, 2005), and increase the amount of information available for the interpretation and analysis of collections of genomes, especially from an ecological and environmental perspective (Martinu and Field, 2005).

[1] Molecular Evolution & Bioinformatics Section, Oxford Centre for Ecology and Hydrology, Oxford, United Kingdom.

[2] School of Computer Science, University of Manchester, Manchester, United Kingdom.

[3] NERC Environmental Bioinformatics Centre, Oxford Centre for Ecology and Hydrology, Oxford, United Kingdom.

[4] The Institute for Genomic Research, Rockville, Maryland.

[5] EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

Community-driven standards have the best chance of success if developed within the auspices of international working groups, and the Genomic Standards Consortium (GSC) has recently formed to work towards creating the "Minimal Information about a Genome Sequence" (MIGS) specification (⟨gense.sf.net⟩). The GSC formed at the "eGenomics: Cataloguing our Complete Genome Collection" workshop held at the National Institute for Environmental E-Science in Cambridge, September 7–9, 2005 (Field et al., 2006). The outputs of this first exploratory workshop included an improved discussion document describing the core pieces of information to be collected in such a specification (the "checklist"), a list of newly formed working groups whose members will work together to refine the checklist, and an open call for the active involvement of the wider community in this standardization effort.

The second follow-up workshop, reviewed here, sought to explore integration and harmonization with existing projects at the The Wellcome Trust Genome Campus and raise general awareness of the GSC goals within the community. It was organized by Peter Sterk and Dawn Field, and took place at the European Bioinformatics Institute (EBI) in Cambridge, England, on 10–11 November, 2005. As with the first meeting, participants included those with experience of building community-based standards, computer scientists, those building genomic databases and conducting large-scale comparative genomic analyses, and biologists.

The outputs of this workshop included the identification of novel case study genomes and reference datasets, the expansion of the GSC, an article reviewing the concept of sample in OMICS standardization activities (Morrison et al., *this issue*), and the subsequent formation of an Implementation Working Group. This workshop also directly led to this special issue of the journal *OMICS* dedicated to data standardization activities, which contains contributed pieces by many of the speakers at the workshop as introduced below (Field and Sansone, *this issue*).

The workshop began with a welcome from Peter Sterk. To further set the context for the event, Dawn Field presented an introduction to the GSC largely by overviewing the first GSC meeting (Field et al., 2006). She thanked Peter Sterk for taking the lead in organizing this workshop as the oversight by representatives of the major bioinformatics centers, like the EBI, would be essential in guiding such a project. The invited speakers for the first and second workshops were brought together to help achieve three major goals, namely, to identify the science questions driving the need for more metadata, to look for ways to harmonize existing and future efforts at metadata capture, and develop a more detailed vision of the shape a new standard might take. Specifically, the speakers at this workshop had been brought together to (1) explore the way in which the GSC fits into functional OMICS data standardization activities, (2) investigate how it can work in harmony with the International Nucleotide Sequence Database Collaborators (INSDC) and other database and tool development projects, (3) examine how the community wants to describe genomes of different taxa, and (4) discuss the best ways to take forward the goals of the GSC. Further information about the GSC is available on its website (⟨gense.sf.net⟩).

OMICS STANDARDS INITIATIVES

This first session was designed to provide an overview of activities in OMICS standardization within the areas of transcriptomics, proteomics, metabolomics, and efforts at integration across these domains. To be of most use, speakers were asked to provide an update on the status of these activities and also a personal account of their experiences. These presentations are expanded upon in this special issue of *OMICS*, where Alvis Brazma (EBI) covers the MGED society's work in the area of transcriptomics (Ball and Brazma, *this issue*), Chris Taylor (EBI) covers HUPO's Proteomic Standards Initiative (PSI) (Taylor et al., 2006, *this issue*), Susanna-Assunta Sansone (EBI) contributes to pieces on metabolomics (Fiehn et al., 2006, *this issue*) and integrative activities in the form of the Reporting Structures for Biological Investigations (RSBI) (Sansone et al., *this issue*) and the Functional Genomics Ontology (FuGO) (Whetzel et al., *this issue*), and Norman Morrison (University of Manchester and the NERC Environmental Bioinformatics Centre), elaborates on the development of the "Env" specification for environmental transcriptomics (Morrison et al., *this issue*).

A few themes emerged from this session, the most important being the need for integration and the apparent role of the newly formed GSC in helping to define the concept of "sample" in collaboration with the wider functional genomics community (Morrison et al., *this issue*). Standardization activities have chosen to deal with the issue of describing "sample" or the source of the biological material for an OMICS experiment, in different ways, modelling it with differing levels of complexity, and assigning it different levels of priority. While the MIAME specification deals with "source" or "biomaterial" (and this has been extended by the Env specification [Morrison et al., *this issue*]), the PSI has decided to tackle the issue of "sample processing" but not descriptors of sample, and Chris Taylor openly welcomed the future activities of the GSC in this domain (Taylor et al., *this issue*). Other key messages included the need to define the scope of any standardization activity, an often long, but extremely important process, and the need to keep the integrity of a specification, an exchange format, and implementation activities as independent focused activities.

ALLIED PROJECTS

The second session included invited speakers from key projects at the EBI of special relevance to the future development of the GSC. Close interactions with, and guidance from, the International Nucleotide Databases, is central to the success of this project. Guy Cochrane (EBI), annotation coordinator of the EMBL Nucleotide Sequence Database, elaborated on two topics raised by Tatiana Tatusova, head of the NCBI's Entrez Genomes, at the first workshop (Field et al., 2006). The first is the way that International Nucleotide Sequence Database Collaborators (INSDC) have designed the features and qualifiers used in GenBank/EMBL/DDBJ to allow specialist communities to use and extend them to capture descriptions of genomes in greater detail (Morrison et al., *this issue*). The second is the establishment of a genome project database by the INSDC including the creation of genome project IDs and interfaces for the capture of an increased amount of relevant metadata describing genomes (Morrison et al., *this issue*). This means there is a clear mechanism by which the GSC can work with the INSDC towards improving our descriptions of genomes. Namely, through encouraging greater compliance with existing qualifiers, through the development of novel ones, and through work on ontologies that will allow these features to be described in a standardized way (Morrison et al., *this issue*).

Peter Sterk (EBI) overviewed the Integr8 and Genome Reviews databases, the rationale for their design, and the issues involved in third party annotation and standardization of annotation. He expands on the need to standardize the way in which we represent genome annotations further in his contribution to this issue (Sterk et al., *this issue*). Peter Rice (EBI) shifted the focus from standardization of data to standardization of processes by which data is analyzed, especially in the context of complex pipelines that tie together a large number of tools using E-Science solutions. He overviewed the collaborative EMBRACE, Compara-GRID, and myGrid projects and the general need to standardize approaches that make complex, repeatable, and transparent computational analyses possible.

DESCRIBING GENOMES

A key goal of the third session was the identification and description of future case study genomes from as wide a diversity of genomes (from a taxonomic and ecological perspective) as possible. Whereas the first meeting mainly focused on genomes from prokaryotes, viruses, and meta-genomes (Field et al., 2006), talks at this workshop covered plants, organelles, pathogenic eukaryotes, and prokaryotes.

This series of presentations successfully highlighted the large number of genomes that are being sequenced and the key biological features of different taxa. Jim Leebens-Mack (Penn State University) gave an excellent overview of the status of sequencing of plant genomes. More specifically, he described how phylogenomic approaches are especially useful in plant genomes which have been shaped by high number of duplication events. In this issue, he further discusses the issues involved in the representation of information from phylogenetic analyses (Leebens-Mack et al., *this issue*). Jeffrey L. Boore (DOE Joint Genome

Institute) gave an overview of organellar genome sequencing efforts with a specific focus on the interesting biological features of mitochondria. Such features repeatedly prove challenging to deal with from the standpoint of complying with standardized annotation and databasing solutions (Boore, *this issue*). Christiane Hertz-Fowler (Sanger Institute) complemented Julian Parkhill's presentation on the genome projects of the Sanger Institutes Pathogen Sequencing Unit (PSU) at the first meeting (Field et al., 2006) by discussing the PSU's eukaryotic pathogen sequencing efforts and some of the unusual features of the genomes of small eukaryotes. Barbara Methe (TIGR) overviewed the fascinating biology of Geobacter, a bacterial species capable of producing electricity and degrading uranium.

Finally, we had two talks about standardizing the ways in which we look at genomes and proteomes. Glenn Proctor (EBI) stood in for Ewan Birney, head of the Ensembl project, to overview the Ensembl system. Ensembl has made it possible to automatically annotate a large number of eukaryotic genomes and plays an important role in making this information widely accessible to the public (Birney et al., 2006). Eugene Kolker (The BIATECH Institute) discussed BIATECH's take on OMICS standards and work towards the standardization of experimental procedures (Hogan et al., *this issue*).

DISCUSSION

The discussions of the fourth session were led by the core members of the GSC present at the workshop (Dawn Field, and Norman Morrison, Jeremy Selegut [TIGR, by video link], and Peter Sterk). The most important issue to re-visit appeared to be the model under which the GSC would contribute to the existing efforts of the INSDC to capture metadata about genome projects. The INSDC has a long history of describing nucleotide sequences and is now dedicating substantial efforts to building custom solutions for managing genomic data (Morrison et al., *this issue*). Guy Cochrane elaborated on the process by which changes to INSDC file features and qualifiers are considered by the INSDC and agreed to represent the interests of the GSC at the collaborators meeting in May 2006.

The discussions also further highlighted the need for, and the benefits of, an early implementation of a future repository. A pilot implementation of the GSC's Genome Catalogue was demonstrated and received positive feedback from the group. In brief, this catalogue provides an interface for the MIGS specification that is generated "on-the-fly" from an underlying XML schema. This database is designed to have a low development overhead in the short term while the specification is expected to continuously change. The benefit of this early implementation in the short term is that the Genome Catalogue can serve as a "sandbox" to help the community discuss and illustrate the types of descriptors to be collected about different taxa.

There were also extended discussions on issues of implementation, and these have been followed up by the formation of a GSC Implementation Working Group. At its first meeting, this group discussed the current XML-based implementation of the Genome Catalogue, the concept of "sample" in OMICS, the need for ontologies to describe the proposed content of the Genome Catalogue, the issue of integration, and ideas for future user interfaces. This group has also since started working on the ontological needs of the GSC by working closely with the Environmental Genomics Working Group (EGWG) (Morrison et al., *this issue*) to contribute concepts and terms to the Functional Genomics Ontology (FuGO) (Whetzel et al., *this issue*).

CONCLUSION

Overall, this workshop produced a general consensus on the importance of having a resource for genome meta data that accurately describes the organisms to perform meaningful (comparative) analyses on these genomes. While the content and implementation of this project needs further discussions, the core members of the GSC felt that this workshop provided important insights and ideas that will help move this project forward.

The GSC continues to make its open call for new members, especially those interested in participation in the identified working groups dedicated to descriptions of different taxa (metagenomes, eukaryotes, prokaryotes, viruses, plasmids, and organelles) and concepts (organism, phenotype, environment, sample processing, data processing). Funding for future workshops has been secured in part in the form of a NERC International Opportunities Fund Award to D.F. (NE/3521773/1) and with the promise of two future workshops at NIEeS. Anyone interested in knowing more about or joining this effort is encouraged to contact us (⟨gense.sf.net⟩).

ACKNOWLEDGMENTS

We acknowledge the invaluable contributions of all of the participants who attended the workshop. We would also like to thank Margaret Shore-Nye, Alison Barker, Peggy Nunn, and John Livingston for organizational support. This workshop was funded by a NERC International Opportunities Fund Award (NE/3521773/1) to D.F.

REFERENCES

BALL, C.A., and BRAZMA, A. (2006). MGED standards: work in progress. OMICS (*this issue*).

BIRNEY, E., ANDREWS, D., CACCAMO, M., et al. (2006). Ensembl 2006. Nucleic Acids Res **34,** D556–561.

BOORE, J.L. (2006). Requirements and standards for organelle genome databases. OMICS (*this issue*).

FIEHN, O., KRISTAL, B., OMMEN, B.V., et al. (2006). Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. OMICS (*this issue*).

FIELD, D., and HUGHES, J. (2005). Cataloguing our current genome collection. Microbiology **151,** 1016–1019.

FIELD, D., and SANSONE, S.-A. (2006). A special issue on data standards. OMICS (*this issue*).

FIELD, D., GARRITY, G., MORRISON, N., et al. (2006). eGenomics: cataloging our complete genome collection I. Comp Funct Genomics **6,** 357–362.

HOGAN, J.M., HIGDON, R., and KOLKER, E. (2006). Experimental standards for high-throughput proteomics. OMICS (*this issue*).

LEEBENS-MACK, J.E.A. (2006). Taking the first steps towards a standard for reporting on phylogenies: Minimal Information about a Phylogenetic Analysis (MIAPA). OMICS (*this issue*).

MARTINU, J.B.H., and FIELD, D. (2005). Ecological perspectives on our complete genome collection. Ecol Lett **8,** 1334–1345.

MORRISON, N., COCHRAN, G., FARUQUE, N., et al. (2006a). Concept of sample in OMICS technology. OMICS (*this issue*).

MORRISON, N., WOOD, A.J., HANCOCK, D., et al. (2006b). Annotation of environmental OMICS data: application to the transcriptomics domain. OMICS (*this issue*).

SANSONE, S.-A., ROCCA-SERRA, P., TONG, W., et al. (2006). A strategy capitalizing on synergies: The Reporting Structure for Biological Investigation (RSBI) working group. OMICS (*this issue*).

STERK, P., KERSEY, P.J., and APWEILER, R. (2006). Genome reviews: standardizing content and representation of information about complete genomes. OMICS (*this issue*).

TAYLOR, C.F., HERMJAKOB, H., JULIAN, JR., R.K., et al. (2006). The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). OMICS (*this issue*).

WHETZEL, P.L., RYAN, R.R., BRINKMAN, H.C., et al. (2006). Development of FuGO: an ontology for functional genomics investigations. OMICS (*this issue*).

Address reprint requests to:
Dr. Dawn Field
Molecular Evolution & Bioinformatics Section
Mansfield Road
Oxford Centre for Ecology and Hydrology
Oxford, OX1 3SR UK

E-mail: dfield@ceh.ac.uk

This article has been cited by:

1. Puja Ravikumar, R K SinghGenome Enabled Technologies in Green Chemistry 611-626. [CrossRef]
2. George M. Garrity. 2011. The State of Standards in Genomic Sciences. *Standards in Genomic Sciences* **5**:3, 262-268. [CrossRef]
3. Lars Vogt, Thomas Bartolomaeus, Gonzalo Giribet. 2010. The linguistic problem of morphology: structure versus homology and the standardization of morphological data. *Cladistics* **26**:3, 301-325. [CrossRef]
4. Inigo San Gil, Wade Sheldon, Tom Schmidt, Mark Servilla, Raul Aguilar, Corinna Gries, Tanya Gray, Dawn Field, James Cole, Jerry Yun Pan, Giri Palanisamy, Donald Henshaw, Margaret O'Brien, Linda Kinkel, Katherine McMahon, Renzo Kottmann, Linda Amaral-Zettler, John Hobbie, Philip Goldstein, Robert P. Guralnick, James Brunt, William K. Michener. 2008. Defining Linkages between the GSC and NSF's LTER Program: How the Ecological Metadata Language (EML) Relates to GCDML and Other Outcomes. *OMICS: A Journal of Integrative Biology* **12**:2, 151-156. [Abstract] [Full Text PDF] [Full Text PDF with Links]
5. Jeroen Raes, Konrad Ulrich Foerstner, Peer Bork. 2007. Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology* **10**:5, 490-498. [CrossRef]
6. Julian L. Griffin, Andrew W. Nicholls, Clare A. Daykin, Sarah Heald, Hector C. Keun, Ina Schuppe-Koistinen, John R. Griffiths, Leo L. Cheng, Philippe Rocca-Serra, Denis V. Rubtsov, Donald Robertson. 2007. Standard reporting requirements for biological samples in metabolomics experiments: mammalian/in vivo experiments. *Metabolomics* **3**:3, 179-188. [CrossRef]
7. Norman Morrison, Dan Bearden, Jacob G. Bundy, Tim Collette, Felicity Currie, Matthew P. Davey, Nathan S. Haigh, David Hancock, Oliver A. H. Jones, Simone Rochfort, Susanna-Assunta Sansone, Dalibor Štys, Quincy Teng, Dawn Field, Mark R. Viant. 2007. Standard reporting requirements for biological samples in metabolomics experiments: environmental context. *Metabolomics* **3**:3, 203-210. [CrossRef]
8. Mariët J. van der Werf, Ralf Takors, Jørn Smedsgaard, Jens Nielsen, Tom Ferenci, Jean Charles Portais, Christoph Wittmann, Mark Hooks, Alberta Tomassini, Marco Oldiges, Jennifer Fostel, Uwe Sauer. 2007. Standard reporting requirements for biological samples in metabolomics experiments: microbial and in vitro biology experiments. *Metabolomics* **3**:3, 189-194. [CrossRef]
9. Dawn Field, George Garrity, Tanya Gray, Jeremy Selengut, Peter Sterk, Nick Thomson, Tatiana Tatusova, Guy Cochrane, Frank Oliver Glöckner, Renzo Kottmann, Allyson L. Lister, Yoshio Tateno, Robert Vaughan. 2007. eGenomics: Cataloguing Our Complete Genome Collection III. *Comparative and Functional Genomics* **2007**, 1-7. [CrossRef]
10. Dawn Field, Gareth Wilson, Christopher van der Gast. 2006. How do we compare hundreds of bacterial genomes?. *Current Opinion in Microbiology* **9**:5, 499-504. [CrossRef]