OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2008.0014

Meeting Report: The Fourth Genomic Standards Consortium (GSC) Workshop

Dawn Field,[1*] Frank Oliver Glöckner,[2] George M. Garrity,[3] Tanya Gray,[1] Peter Sterk,[4] Guy Cochrane,[4] Robert Vaughan,[4] Eugene Kolker,[5–7] Renzo Kottmann,[2] Nikos Kyrpides,[8] Sam Angiuoli,[9] Peter Dawyndt,[10] Robert Guralnick,[11] Philip Goldstein,[11] Neil Hall,[12] Lynette Hirschman,[13] Saul Kravitz,[9] Allyson L. Lister,[14] Victor Markowitz,[15] Nick Thomson,[16] and Trish Whetzel[17]

Abstract

This meeting report summarizes the proceedings of the "eGenomics: Cataloguing our Complete Genome Collection IV" workshop held June 6–8, 2007, at the National Institute for Environmental *e*Science (NIE*e*S), Cambridge, United Kingdom. This fourth workshop of the Genomic Standards Consortium (GSC) was a mix of short presentations, strategy discussions, and technical sessions. Speakers provided progress reports on the development of the "Minimum Information about a Genome Sequence" (MIGS) specification and the closely integrated "Minimum Information about a Metagenome Sequence" (MIMS) specification. The key outcome of the workshop was consensus on the next version of the MIGS/MIMS specification (v1.2). This drove further definition and restructuring of the MIGS/MIMS XML schema (syntax). With respect to semantics, a term vetting group was established to ensure that terms are properly defined and submitted to the appropriate ontology projects. Perhaps the single most important outcome of the workshop was a proposal to move beyond the concept of "minimum" to create a far richer XML schema that would define a "Genomic Contextual Data Markup Language" (GCDML) suitable for wider semantic integration across databases. GCDML will contain not only curated information (*e.g.*, compliant with MIGS/MIMS), but also be extended to include a variety of data processing and calculations. Further information about the Genomic Standards Consortium and its range of activities can be found at http://gensc.org

[1] NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford, OX1 3SR United Kingdom.
[2] Microbial Genomics Group, Max Planck Institute for Marine Microbiology & Jacobs University Bremen, Bremen, Germany.
[3] Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan.
[4] EMBL Outstation—The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom.
[5] The BIATECH Institute, Bothell, Washington.
[6] Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Information, University of Washington Seattle, Washington.
[7] Seattle Children's Hospital Research Institute, Seattle, Washington.
[8] Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, Calfiornia.
[9] J. Craig Venter Institute, Rockville, Maryland.
[10] Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium.
[11] Department of Ecology and Evolutionary Biology and University of Colorado Natural History Museum, University of Colorado, Boulder, Colorado.
[12] The University of Liverpool, School of Biological Sciences, Liverpool, United Kingdom.
[13] Information Technology Center, The MITRE Corporation, Bedford, Massachusetts.
[14] CISBAN & School of Computing Science, Newcastle University, Newcastle upon Tyne, United Kingdom.
[15] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California.
[16] The Pathogen Sequencing Unit, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom.
[17] Center for Bioinformatics and Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania.

Introduction

The Genomic Standards Consortium (GSC) is an initiative working toward richer descriptions of our collection of genomes and metagenomes. Established in September 2005, this international community includes representatives from NCBI, EMBL, DDBJ, JCVI, JGI, EBI, Sanger, CAMERA, and a range of research institutions. The goal of the GSC is to promote mechanisms of standardizing the description of (meta) genomes and the exchange and integration of (meta)genomic data. The rapid pace of genomic and metagenomic sequencing projects will only increase, as the use of ultrahigh throughput methods becomes commonplace, and standards are vital to scientific progress and data sharing.

Specifically, the GSC is developing the "Minimum Information about a Genome Sequence" (MIGS) specification, which has recently been extended to create the closely integrated "Minimum Information about a Metagenome Sequence" (MIMS) specification. Capture of information compliant with MIGS/MIMS is possible using the GSC Genome Catalog (or GCat) portal (http://gensc.org). The primary aim of developing such new standards is to ensure that researchers generating (meta) genome sequence data contribute to, and increase, the quality and quantity of contextual (meta) data available. This ensures that meaningful comparisons can be performed across the entire public (meta) genome collection without hindrance, especially from an ecological and environmental perspective. More information about the GSC can be found at its Web site (http://gensc.org).

The fourth workshop was organized by **Dawn Field** (NERC Centre for Ecology and Hydrology), **Frank Oliver Glöckner** (Max Planck Institute for Marine Microbiology), **George Garrity** (Michigan State University), and **Tanya Gray** (NERC Centre for Ecology and Hydrology). This workshop took place at the National Institute for Environmental *e*Science (NIE*e*S) in Cambridge, England on June 6–8, 2007.

Workshop proceedings

Dawn Field (NERC Centre for Ecology and Hydrology) introduced the GSC workshop briefly. She welcomed returning and new participants and reviewed the agenda and goals of the workshop. The purpose of this workshop was to review the "big picture" surrounding MIGS development and seek new linkages, for example, through the culture collection and text mining communities.

An introductory presentation was then given by **Stuart Ballard**, the Deputy Director of NIEeS (The National Institute for Environmental eScience), who introduced the work of NIEeS. NIEeS has funding to host workshops on behalf of the environmental *e*Science community and has provided funding and logistical support for three GSC workshops thus far (1, 3, and 4). Stuart was pleased to see familiar faces and new people in the group. There are many new and exciting projects ongoing at NIEeS (http://www.niees.ac.uk/). One such project is a new collaborative development project called SciSpace, that allows collaborators to exchange data and communication using wikis, instant messaging, and e-mail. NIEeS has been successful in establishing new working groups in addition to hosting meetings, and encourages groups like the GSC to submit ideas for support.

Session I: updates on GSC activities

In the past three workshops, the first session was always designed to provide an overview of the current and future state of genome and metagenome collection. In contrast, this workshop assumed all the participants understood the urgency of standardizing the description of our ever-growing collection of genomes and metagenomes and began directly with updates on GSC activities. To start the session, **Dawn Field** explained the strategy behind the workshop agenda and gave an overall update on progress toward defining the scope, syntax, and semantics of MIGS/MIMS.

The completion of MIGS 2.0 by the end of 2007 was well underway. Significant help has come from several members of the CAMERA project (http://camera.calit2.net/), the first official adopter of the MIGS/MIMS specification. Working with other GSC members, CAMERA has proposed a set of changes to allow MIGS/MIMS to deal with the top-level description of complex metagenomic experiments. The GSC continues to contribute to the Ontology for Biomedical Investigations (OBI) and "Minimum Information about a Biological and Biomedical Investigation" (MIBBI) projects, and has most recently helped launch the EnvO (Environmental Ontology) project to aid in the description of habitat. The manuscript introducing MIGS/MIMS has been accepted by *Nature Biotechnology,* following extensive community consultation (Jan 1 to April 19, 2007), and has been published under Open Access terms as part of a theme on data standardization.

A key outcome of the third GSC workshop was the agreement by the EBI (**Guy Cochrane** and **Bob Vaughan**) and NCBI (**Tatiana Tatusova**) to provide their genome collections for the sake of read-only import into GCat (Field et al., 2007). Population of the Genome Catalog, it was hoped, would increase the chance of obtaining submissions of MIGS/MIMS-compliant reports from the community and would also help quantify the amount of optional data already available (e.g., geographic location coordinates in the optional "/lat_long" qualifier). Lists of genomes from Genomes Online Database (GOLD) (http://www.genomesonline.org/) and the NCBI are now included in the GCat. However, deposition of EMBL genomes is dependent upon the as-yet incomplete mapping of EMBL genomes to INSDC Genome Project Identifiers (PIDs). The issue of a growing need to map identifiers across databases has led to the establishment of a new GSC project called the "Genomic Rosetta Stone."

MIGS/MIMS implementation plans by CAMERA have generated detailed discussions about how to extend the core MIGS structure to allow the capture of more complex metagenomic experiments (e.g., those with pooled data or multiple samples). An initial mapping experiment from the CAMERA schema to MIGS/MIMS showed an extensive amount of data loss, which would be expected in the case of a "minimum" specification. Interest in capturing not only the minimum but also a richer set of data sparked the idea of extending the MIGS/MIMS XML and schema into a full markup language, a concept that would be introduced by **Renzo Kottmann** (MPI-Bremen).

With respect to the 10-point roadmap developed at the last workshop (Field et al., 2007), five items have been completed (numbers in parentheses denote items in the original roadmap). MIGS was updated to version 1.1, GCat identifiers have been implemented, a production version of GCat has been released, guidelines (instructions) for the submission of genome reports were developed, and the GSC has returned to NIEeS as planned for a fourth workshop. The remaining points of the Roadmap are still ongoing activities: the GSC's solicitation of MIGS/MIMS compliant genome reports needs to rise in priority as the checklist stabilizes; the batch upload facility in GCat for importing data from GOLD needs to be applied more widely; the formal policy on ownership of the contents of the genome report needs to be written; and discussions of funding opportunities are ongoing.

At the end of the overview, **Dawn Field** handed over to **Guy Cochrane** (EMBL), who provided an overview of discussions about the GSC at the INSDC annual meeting in May 2007. The INSDC is a collaboration between the DDBJ/EMBL/Genbank (http://www.insdc.org/) that the GSC recognizes as the primary international authority for the stewardship of genomes and metagenomes (Field et al., 2008). As a result of Guy's update, the INSDC offered five main items of guidance:

1. The INSDC takes a passive interest in the MIGS standard (both its development and its uptake by the community);
2. The INSDC encourages curated content at GCat as a means for GSC to further community acceptance;
3. The INSDC recommends that fields common to both INSDC and GCat records be made read-only in GCat, such that the INSDC remains the authoritative source;
4. The INSDC will not direct its submitters routinely toward GCat, but will provide access to dumps of read-only information for upload;
5. The INSDC considers that gene nomenclature should not be covered by GSC activities.

The GSC responded as follows:

1. The GSC welcomed the fact that the INSDC gave considered guidance that will help shape the evolution of the GSC;
2. In particular, the GSC agreed that to deserve full support from INSDC, high-quality MIGS/MIMS-compliant, curated data must be produced and made available through Gcat;
3. The GSC underscored that it recognizes the INSDC as the authority in the capture of metadata; read-only import is ideal. There could be further discussion of optional fields in INSDC record that are mandatory in MIGS;
4. The GSC realizes it must gain further ground in the community before it can expect the INSDC to actively direct submitters to Gcat for the submission of extended metadata.

5 The GSC reiterates that it is interested in extending the quantity and quality of metadata captured about genomes, and is pleased to see others lead on the area of improving gene nomenclature.

Tatiana Tatusova, the head of NCBI's Entrez Genomes, sent apologies so instead of an update on the INSDC's Genome project databases and genome project IDs, **Bob Vaughan** from EMBL stepped in to discuss the results of trying to "dump" all EMBL genomes for the sake of importing them into GCat. This project was a result of discussions at the last workshop, but implementation proved more complex than anticipated due to vocabulary issues. For example, some organism names were not identical—in some cases, even for two chromosomes from the same genome. EMBL is now simplifying their relational database by reducing the number of tables. The changes allowed EMBL curators to fix internal mistakes and inconsistencies. EMBL curated genomes have a very high quality of information, which improves the data dump procedure. About 50% of the genomes submitted through EMBL to the INSDC are curated by EMBL staff and the rest are received via automated batch uploads. The latter, external entries, do not receive the same level of attention. Another aid to the EMBL genome dump procedure is that future entries will have INSDC PIDs, theoretically making this exercise straightforward. EMBL is considering internal IDs for entries that do not get INSDC PIDs.

Peter Sterk (EBI) gave an update on the GSC involvement in MIBBI ("Minimum Information about Biomedical and Biological Investigations," http://mibbi.sf.org). Initiated by the Proteomics Standards Initiative (PSI), the Reporting Standards for Biological Investigations (RSBI) and the GSC in response to the proliferation of checklist communities, MIBBI now has committed 20 checklist projects. In addition to a portal designed to serve as a one-stop shop for checklists and associated information, the MIBBI Foundry has also been created. Groups subscribing to the Foundry are, as in the OBO Foundry (http://obofoundry.org/), expected to not only make their checklists more transparent to the community, but to also participate in modularization of checklists and even potential revisions to make their checklists more orthogonal where possible. **Chris Taylor** of the EBI, who sent his apologies to the meeting, has undertaken the first modularization exercise across these checklists. The results are available as an excel spreadsheet on the MIBBI Web site (http://mibbi.sf.net).

Allyson Lister (Newcastle University) then addressed the issue of a common syntax for data exchange and interoperability that extends beyond the (meta)genomic community. While the GSC already has an XML schema for MIGS/MIMS, it can further address syntax harmonization with the wider standards community by evaluating the use of the Functional Genomics Experiment Model (FuGE) (http://fuge.sf.net) to create a MIGS/MIMS Object Model (Jones et al., 2008). FuGE has been developed to be a core object model from which community-based extensions can be created to describe all aspects of an experiment, including protocols, software, and equipment. By implementing an extension of FuGE for the MIGS/MIMS community, the GSC can provide an implementation- and operating system-independent Unified Modeling Language, or UML, based format. Such UMLs can be easily and automatically used to generate implementations such as XML schema, relational databases, and Java APIs. FuGE is more generic than the existing GSC strategy, and Allyson reviewed the first draft Object Model she developed of the FuGE project, in collaboration with **Andy Jones** (University of Manchester).

Tanya Gray overviewed her progress on developing the Genome Catalog (GCat) for the GSC. Since the last workshop, a generic version of this XML catalog application has been developed. The significantly improved software autogenerates input forms from uploaded XML schema files and can capture XML schema-compliant reports. Adding more MIGS/MIMS compliant content was identified as a priority. Methods of adding content to the Genome Catalogue (GCat) and the requirement for a genome report batch upload function were described. The introduction of a framework to support collaborative software development was also proposed.

Session II: MIGS/MIMS—extending MIGS to capture complex metagenomic studies

The second session on day 1 was chaired by **Frank Oliver Glöckner**, the originator of the MIMS ("Minimum Information about a Metagenome Sequence") extension of MIGS for describing contextual data (measurements) that help define habitat. At the third GSC workshop, he introduced MIMS, stressing the need to place sequences into their properenvironmental context (e.g., marine, terrestrial, symbiotic). At that time, he emphasized the need to know the exact geographic location (*x, y*), depth/altitude (*z*) and time (*t*) of any sample be taken in any molecular field study. Such geospatial information can then be used as a universal anchor by allowing sequence data to be described in the context of prevailing biodiversity and habitat parameters. It also allows the supplementation of onsite information with dynamic data layers from global monitoring systems, leading to an integrated ecosystem assessment.

While *x, y, z,* and *t* are core parts of MIGS, environmental descriptors like salinity and pH are part of MIMS to help define habitat. MIMS should capture a more extensive list of habitat parameters that provide a rich set of contextual metadata for the sake of hypothesis generation and testing as well as ecosystems biology.

Renzo Kottmann (Max Planck Institute for Marine Microbiology) from **Frank Oliver Glöckner's** group and a developer of the Genomes Mapserver (http://www.megx.net/mm/genomes_mapserver/mapserver/) reviewed the technical implementation of the MIGS/MIMS checklist as an XML Schema, especially in light of the many discussions with the CAMERA and Alpine Microbial Observatory (AMO) teams. The ongoing work to make MIGS/MIMS suitable for the capture of complex metagenomic studies had raised several issues: (a) the need to clarify the relationship between MIGS and MIMS (something that was largely done as a result of this workshop and postworkshop discussions), (b) the need for a way to allow capture of richer data sets while still developing a minimum standard (MIGS/MIMS). To overcome these issues, a general solution was proposed—the extension of the current schema into a far richer Genomic Contextual Data Markup Language (GCDML). GCDML would serve as a common XML vocabulary encoding for modeling, transporting and storing (meta)genomic contextual data. This new approach would allow implementation of a rich, open, and extensible main schema from which minimal checklists can be derived. Moreover, this solution combines the need for a rich set of metadata with "minimum" reporting requirements that are easier to comply with and more likely to be adopted by users.

In addition, Renzo suggested a series of modifications to the current XML schema. While it had been initially developed to simplify rapid prototyping of shared and unshared descriptors across the six report types, it was now essential to modify it to allow validation for each report type as well as the inclusion of different controlled vocabularies in the same element for different report types. These are timely improvements, as the checklist is now relatively stable. The strict hierarchical structure of the current MIGS XML Schema should be replaced by a "flatter structure" that both shortens and takes advantage of recycling of elements. The "multiview" nature of the current schemas makes validation problematic; a single element could be mandatory for one type of report, optional for another and not applicable for a third. This could be solved by making elements for each report type explicit.

Saul Kravitz, the program manager for CAMERA at the JCVI, then described CAMERA's mission and specific requirements for MIGS/MIMS. CAMERA has a 5-year grant of $24.5m from the Moore Foundation to build the computational infrastructure required for large-scale analyses of metagenomic data sets, with special emphasis on the Global Ocean Survey (GOS) samples from the Sorcerer II voyage. Saul described the portal and its datasets, and then overviewed the development of MIGS version 1.2 to capture the complexity of describing metagenomic datasets.

Rob Guralnick and **Phil Goldstein** (University of Colorado) discussed the principles behind and downstream uses of the Alpine Microbial Observatories (AMO) database. A key goal of this database, funded by the National Science Foundation within its Microbial Observatories program, is to link *x, y, z,* and *t* data with sequence and biogeochemical data. AMO has developed a working data model (http://amo.colorado.edu/data_model.html) and a relational database (http://amo.colorado.edu/db_nav.html) for storing such data. This database and others with similar goals are targets for MIGS/MIMS compliance given their focus on sequences, metagenomes, genomes, and environmental and geographic data. Phil discussed a preliminary mapping of the AMO data model to MIGS/MIMS showing a good overlap of concepts. Both Rob and Phil stressed that community adoption of MIGS/MIMS will be valued for data aggregation not only for end users but also for applications that use such data to facilitate research missions in understanding global microbial diversity and processes.

Victor M. Markowitz (JGI) discussed the metadata collection carried out with JGI's IMG-M metagenome data management and analysis system (http://img.jgi.doe.gov/m). Both isolated genomes and microbiome samples are characterized by a variety of *metadata* attributes that are used in IMG/M for metadata driven search and browsing. Some metadata, such as phenotype, ecotype, disease, relevance, temperature, and pH, are collected via GOLD, with additional metadata collected directly from scientists or publications. The IMG/M is using MIGS/MIMS metadata attributes for its metagenome data submission site. Scientists have found the MIGS/MIMS standard easy to follow, and metadata coverage has increased for the new IMG/M submissions.

A "retrofitting" exercise was carried out for an older metagenome dataset from an agricultural soil sample. This exercise shows the importance of collecting metadata at the time of dataset generation: while some metadata were retrieved from literature and the scientists who analyzed this dataset, other metadata were not recorded, and therefore lost. In order to increase metadata coverage the IMG/M, jointly with GOLD, regularly undertakes surveys of other metadata resources (such as NCBI Projects and NCBI Metagenomics resources). Metadata collection is hindered not only by lack of crossreferences, the problem GSC's Genomic Rosetta Stone effort hopes to address, but also by semantic discrepancies in the way metagenomic datasets are defined. For example, a metagenome project may identify an individual metagenome dataset or may refer to a study involving multiple metagenome datasets that are not identified separately. IMG/M and GOLD are currently addressing such semantic discrepancies by reviewing the way public metagenome studies and data sets are defined, with a revised classification of these projects due to be published by GOLD.

Roundtable discussion

Frank Oliver Glöckner led a roundtable discussion about the scope and syntax of MIGS/MIMS. This discussion built on a technical session held at the beginning of the day and would lead to similar sessions on days 2 and 3. The final outcomes of these continuing discussions are summarized at the end of the report and visually in Figure 1.

Session III: Controlled Vocabularies and Ontologies

The workshop then addressed the urgent need for appropriate semantics resources (controlled vocabularies, ontologies, and ontology-aware tools) for the description of (meta)genomes. **Trish Whetzel** (University of Pennsylvania) chaired the first session of day 2 on the need for controlled vocabularies and ontologies for the description of genomes and metagenomes. Trish had described the development of the Ontology for Biomedical Investigations (OBI) at the previous GSC workshop.

Neil Hall (University of Liverpool) opened the session with an overview of the use of CVs and ontologies in the Genomic Metadata for Infectious Agents (GEMINA) project (http://gemina.tigr.org), a Web-based system to retrieve epidemiological and genomic data associated with microbial pathogens. Currently, the metadata required to track infection in detail, such as disease symptoms, geographic location of isolation, and host age, are missing from genomic records. Such data is a necessity for studying the genetic basis of medically meaningful traits. This project uses a number of CVs, including the infectious disease branch of the disease ontology. It also uses new ontologies to describe symptoms, and sites of infection.

Tanya Gray described the CV and ontology requirements of the GCat and the specific requirements of GCat to support the completion of MIGS/MIMS compliant reports. A process proposed by Trish Whetzel for term capture and submission to external ontology and CV projects was presented. Submission of terms to ontologies and controlled vocabularies available through the EBI's Ontology Lookup Service (OLS) was proposed, with term retrieval via the OLS's Web Services. The talk finished by highlighting some issues for future discussion: namely, the requirement to establish a MIGS/MIMS term approval board and how to validate reports that contain terms defined outside of the MIGS/MIMS XML schema, for instance, in an ontology or controlled vocabulary.

Lynette Hirshman (MITRE) talked about prospects for using text mining in the context of MIGS/MIMS. She discussed

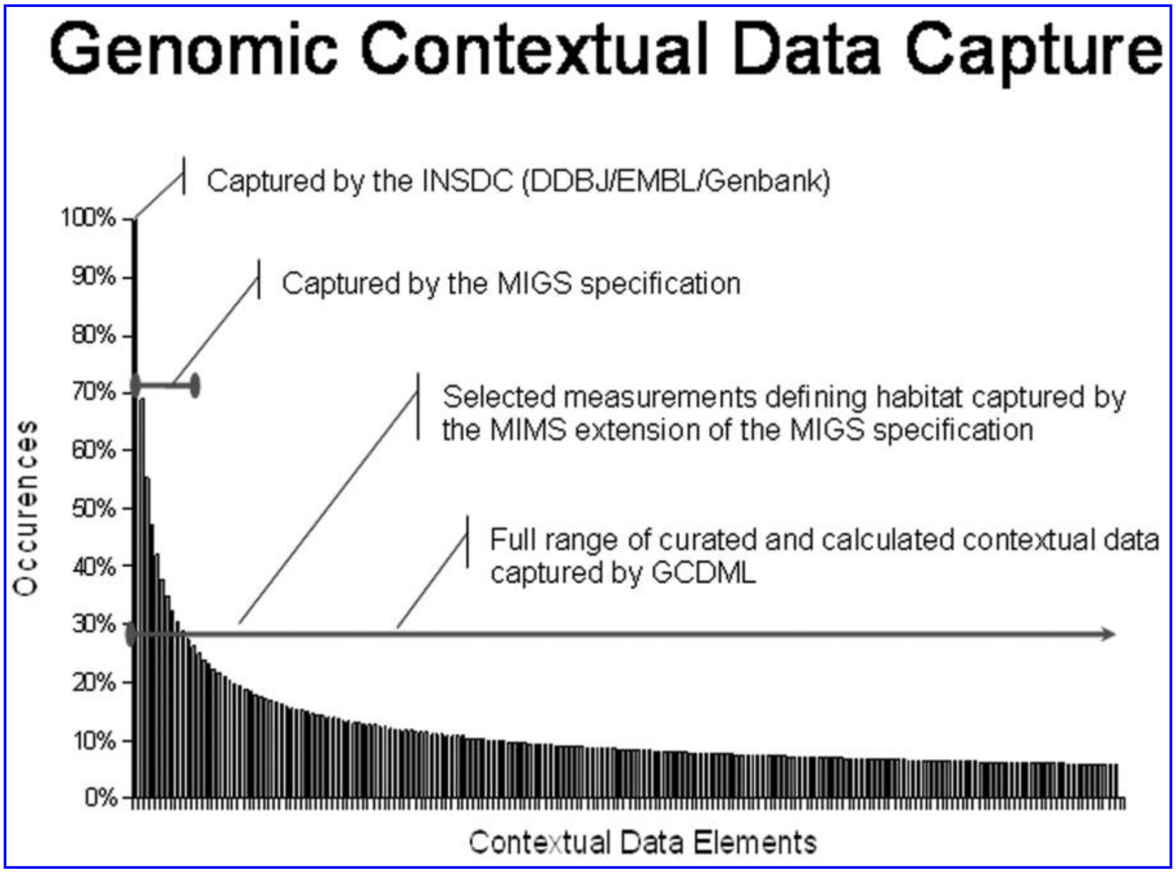

FIG. 1. The frequency of descriptors across experiments. Only a few descriptors are common across all experiments, and these are captured by different levels of standardization. GCDML could capture both common and rare descriptors in a single uniform way.

the capture of environmental metadata, including the geographic location of sampling sites, as well as environmental conditions. This task is complex because the metadata are scattered throughout the full text articles, including experimental methods sections, tables, and supplementary materials (often available only in PDF format). She cited work by **Renzo Kottmann** (MPI-Bremen), who found that, of 77 metagenomics articles, only 77% contained mention of the geospatial metadata, and that text mining techniques were able to identify sentences containing that information at about 70% balanced precision/recall.

Hirschman also reported on results from the recent BioCreative Workshop (Critical Assessment for Information Extraction in Biology, http://biocreative.sourceforge.net) on the evaluation of text mining tools applied to genomics and proteomics. Recent results are promising: automated systems can extract text mentions of genes or proteins and map them to unique identifiers from EntrezGene or SwissProt; however, capture of complex relations, such as protein–protein interaction, is more challenging. Text mining tools may also be useful to the GSC for the identification of terms and concepts for ontology building. She also proposed the creation of a "metadata checker," which would be similar to a spellchecker, that could be used to interact with authors, and to aid in data capture and mapping of metadata to controlled vocabularies.

Discussion during this session was led by **Trish Whetzel**. Appropriate ontology projects have been identified for the descriptors in MIGS/MIMS and this makes it possible to start more effectively managing term capture from MIGS/MIMS genome reports. Trish and Tanya have developed a spreadsheet of terms currently enumerated in the schema and are in the process of identifying how many are already in these ontology projects. As part of the discussion, Trish reviewed and updated this list. At the end, **Dawn Field** called for a vetting group to be established and volunteers were identified. This committee will teleconference once a month to consider new terms. A list of terms will be automatically generated from GCat submissions for inclusion in the spreadsheet.

Session IV: Toward a single, global list of genomes, and metagenomes

The second session of day 2 focused on actions required to start filling the content of the GSC's Genome Catalogue. A key step toward enabling the import of "read-only" information from other sources is the establishment of a single global list of genomes and metagenomes as outlined at the third GSC workshop. Persistent identifiers for genomes and genes are part of the essential infrastructure for the future organization of the complete genome and metagenome collection. Now that PIDs are available from the INSDC, it is possible to use them as "anchors" to generate a mapping of genomic identifiers across a range of databases. This mapping, named the Genomic Rosetta Stone, will offer a new mechanism for the practical integration of a range of source of information.

Using a flip chart, **Dawn Field** gave an overview of the GSC's Genomic Rosetta Stone. The starting point is the "gene cards" **Nikos Kyrpides** (Joint Genome Institute) has implemented in his Genomes Online Database (http://www.genomesonline.org/). There is one page per genome that shows all stored metadata and a list of hyperlinks to identifiers held in a number of other databases. Nikos agreed to add links to GCat to raise awareness of the project, and it was then agreed that it would be easiest if he would also assign (autoincrement) a GCat identifier to each new genome/metagenome when it was entered. Just as GOLD maps goldstamps against INSDC PID, it would be possible (and expected) for other databases to do the same. This would form the mechanism for generating a far larger map of identifiers. Progress toward the Genomic Rosetta Stone and a list of participating databases is available in the GSC wiki along with a pilot ID Resolver (http://gensc.sf.net).

Rob Edwards (San Diego State University), who sent apologies, has already agreed to overview the mapped identifiers made available through the SEED database, one of the many databases involved in the GRS project This left more time for **Peter Dawyndt** (University of Gent) to describe the rationale and technical aspects of the straininfo.net portal (http://www.StrainInfo.net). This portal serves as a one-stop shop for navigating the microbial information landscape found in a large range of physical culture collections (Biological Resource Centres) located around the world. The portal is the result of an intensive mapping project to combine strain-level identifiers/names to create permanent and unique IDs local to the StrainInfo portal that can be used by the wider community in a predictable fashion. For example, mapping these IDs to genome projects in the GOLD database allows tracking of all organisms with complete genome sequences that are catalogued by this portal (Fig. 2).

Session V: data capture and exchange: Web services

This session was chaired by **Anil Wipat** (Newcastle University) and focused on the development of appropriate Web services for the exchange of MIGS/MIMS metadata between different databases and end users. **Matthew Pocock** (Newcastle University) kicked off the session with a general introduction to Web services. Wherever possible, automation should be used to analyze experimental results and understand biological systems in order to properly access and manipulate the large amounts of available biological data. Web services provide a technological solution to exposing data to software. By identifying both data and software with URIs, we can weave bioinformatics into the fabric of the Internet in a way that enables massively high throughput and systematic analysis with minimal human intervention. By embracing semantic Web technologies such as the Web Ontology Language (OWL), computers are able to do some of the conceptual modeling on behalf of the researcher. He then described the power of workflows in combination with tools like Taverna (Oinn et al., 2004) and provided an overview of ComparaGRID, the comparative genomics semantic data integration project of which he is a part (http://metagenome.ncl.ac.uk/comparagrid/).

Tanya Gray described how external Web Services such as NCBI Entrez have been used in the Genome Catalog to retrieve data for reports as well as on-the-fly genomic data integration. Plans for the incorporation of further genomic data Web Services including CBS's Genome Atlas, EBI's Genome Reviews, and the Ribosomal Database Project were described. The current provisions for Web Services in the Genome Catalogue were described. Future work includes: col-

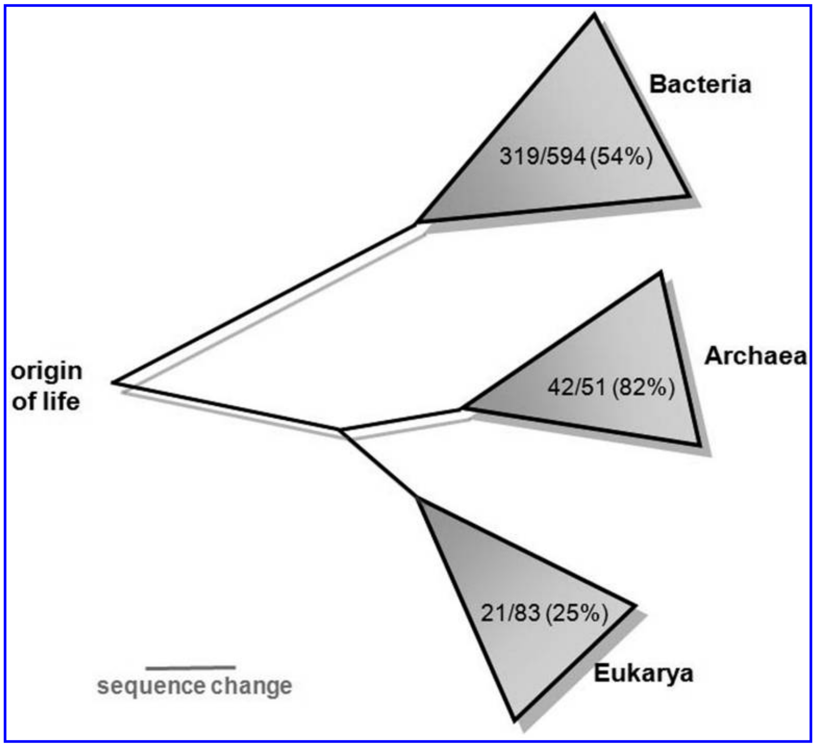

FIG. 2. Organisms targeted by completed whole-genome sequencing projects that are available through the StrainInfo.net bioportal. The low availability of eukaryotic organisms is caused by the current restriction of the bioportal to culture collections specializing in bacteria, archaea, filamentous fungi, and yeasts. This information is kept up to date through automated orchestration with the Genomes Online Database (GOLD) (Liolios et al., 2008).

laboration with the Taverna team for the Genomic Rosetta Stone Web Service; gathering requirements from the GSC community; identifying infrastructure requirements to support Web Service provision, and the addition of RSS feeds for genome reports and other resources in the Genome Catalogue as recommended by Rod Page. Finally, some issues were highlighted for future discussion: the possible implementation of Web services using SOAP; the usefulness of on-the-fly genomic data integration beyond increasing search engine-related traffic; and community requirements for the Web services provided in GCat.

Session VI: The GSC roadmap

All of the workshop discussions were spurred on by a series of pivotal breakout sessions. These focused group discussions proved critical to the formulation of the GSC Roadmap, and therefore are presented collectively here. The technical breakout sessions were initiated by the CAMERA team to spend time with the wider GSC group to address key issues associated with extending the XML schema. The *ad hoc* "big picture" sessions were sparked by the issue that the technical discussions could not advance if the scope of the checklist was not sufficiently defined.

The first morning's technical discussions were shaped by the CAMERA project's need to express a richer set of fields in MIGS/MIMS for the sake of data exchange. This could either explode the number of descriptors in MIGS/MIMS, or lead to an XML schema that implemented MIGS/MIMS plus provided a range of optional fields for much large representation of data. Renzo Kottman formalized this latter idea with GCDML (discussed above). The group agreed, in this context, on a useful concept of defining standards based on the frequency of particular descriptors across a set of experiments (Fig. 1).

In reality, there are only a few descriptors shared by any type of experiment. For genomes and metagenomes, most of these are captured by the INSDC. There are further descriptors that are shared among all of a type of experiments (i.e., a broad taxonomic group), and this is what MIGS aims to capture. Beyond this, there could be absolutely essential descriptors that might only occur in a handful of projects—or even just one. This low frequency does not diminish their importance to the interpretation of that experiment, just provides a rationale for leaving them out of a "blanket" specification to be widely applied. The measurements that define habitat within MIMS fall into lower frequency bins when all genomes and metagenomes are considered, but into higher frequency bins when only environmental samples are considered and into the highest when only metagenomes are considered. This is therefore grounds for considering the descriptions of metagenomes as a separate but closely integrated specification—hence, the future use of MIGS/MIMS. **Sam Angiuoli** generated a graph of this relationship and presented in it the MIMS roundtable discussion. With the general agreement on the construction of GCDML, this graph was further elaborated into Figure 1. Consensus on beginning work on GCDML was a critical outcome of the workshop, and one that will shape the work of the GSC for the next year and beyond.

The *ad hoc* "checklist" session on the morning of day 2 occurred in response to emergent difficulties in the technical sessions. People had difficulty separating checklist from implementation. Without a clear definition of the checklist, implementation would proceed far more slowly. Discussions of the checklist and its technical implementation need to be widely separated. When the groups rejoined, **George Garrity** summarized the work of the group by presenting three top level categories. He made the group put aside the checklist and work from scratch to identify the most important concepts from which to build the schema. After much debate, this concept resulted in three concepts: that of nucleic acid (to replace "organism," which could not include metagenomes), location, and sequencing method. Reassuringly, these matched the checklist. George Garrity also underscored that the whole purpose of MIGS/MIMS was to create a searchable source of information to enable comparative genomic research.

Frank Oliver Glöckner reported that the technical session participants largely came to the same conclusions and worked during the subsequent session to code these concepts into XML. This set the stage for a revised version of the XML schema to better meet everyone's needs. This provided the foundation for the final technical discussions on day 3 and the first draft version of GDCML. This division of labor among participants with different interests produced an improved syntax (the technical sessions worked directly in XML) while not losing sight of the all important factor of a rigorous definition of scope.

On day 3, the workshop split again into those wanting to continue technical discussions and those wanting to discuss the high-level strategy of the GSC. Both groups came to very similar conclusions about the nature of the GSC Roadmap and priorities were clear. The adoption of MIGS would require a stable version of MIGS by early 2008, and the commitment of early adopters ready to the use of the XML implementation of MIGS to collect and exchange data, George also advocated writing an executive summary of the GSC and its activities as a tool for engaging the wider community. To fulfill this last action, a concise overview has been posted to Wikipedia and will be kept up to date (http://en.wikipedia.org/wiki/Genomic_Standards_Consortium).

Themes and the way forward

A clear theme that emerged repeatedly throughout the workshop was that the immediate goal of the GSC must be the generation of content (especially through batch uploading mechanisms). This depends heavily on the creation of a stable version of the checklist and schema, as well as the ability to batch upload a series of promised data sets. It was agreed that the Genome Catalog would serve as the portal for the capture of MIGS/MIMS-compliant reports, although further implementations were discussed and strongly encouraged. The proposed Genomic Rosetta Stone project was further formalized as a necessary step in producing a single, global list of (meta) genomes. The final wrap-up session allowed for a general review of actions. Housekeeping activities addressed at this point included a review of funding options, final revisions to the MIGS paper, planning for the workshop report, and further planning for the next workshop.

Perhaps the single most important outcome was the enthusiasm of the group for another GSC workshop in Dec 2007. This workshop will have very specific goals, and will be a final collaborative discussion on MIGS/MIMS before the MIGS checklist is presented as a stable version ready for implementation. The workshop will present the first draft of GCDML, implementation of the Genomic Rosetta Stone and will feature a report from the Term Vetting Group. Reports from MIGS/MIMS implementation case studies, especially CAMERA, will also be featured, along with a demonstration of data exchange (e.g., between CAMERA and GCat) and a review of all "early adopters" willing to accept working toward MIGS/MIMS compliance following the workshop.

Acknowledgments

The authors acknowledge the invaluable contributions of all of the workshop participants and contributors. We express special thanks to Caroline Dombrowski for reading and editing this report. We greatly appreciate the work of NIE*e*S in supporting and hosting the workshop, The GSC is supported by a NERC funds (NE/3521773/1 and NE/E007325/1) to D.F. Additional sponsorship generously provided by Paul Gilna through the CAMERA project.

Author Disclosure Statement

The authors declare that no competing financial interests exist.

References

Field, D., Garrity, G.M., Gray, T., Selengut, J., Sterk, P., Thomson, N.R., et al. (2007). Meeting report: eGenomics: cataloguing our complete genome collection III. Comp Funct Genom. Article ID 47304, 7 pgs.

Field, D., Garrity, G.M., Gray, T., Morrison, N., Selengut, J.D., Sterk, P., et al. (2008). Towards a richer description of our complete collection of genomes and metagenomes: the "Minimal Information about a Genome Sequence." Nat Biotechnol **26,** 541–547.

Jones, A.R., Miller, M., Aebersold, R., Apweiler, R., Ball, C.A., Brazma, A., et al. (2008). The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol **25,** 1127–1133.

Liolios, K., Mavormatis, K., Tavernarakis, N., and Kyrpides, N. (2008). The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36 (Database issue), D475–D479.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., et al. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics **20,** 3045–3054.

Address reprint requests to:
Dawn Field
NERC Center for Ecology and Hydrology
Mansfield Road
Oxford, OX1 3SR United Kingdom.

E-mail: dfield@ceh.ac.uk

This article has been cited by:

1. Christina Holmes, Fiona McDonald, Mavis Jones, Vural Ozdemir, Janice E. Graham. 2010. Standardization and Omics Science: Technical and Social Dimensions Are Inseparable and Demand Symmetrical Study. *OMICS: A Journal of Integrative Biology* **14**:3, 327-332. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]
2. Dawn Field, Iddo Friedberg, Peter Sterk, Renzo Kottmann, Frank Oliver Glöckner, Lynette Hirschman, George M. Garrity, Guy Cochrane, John Wooley, Jack Gilbert. 2009. Meeting Report: Metagenomics, Metadata and Meta-analysis; (M3) Special Interest Group at ISMB 2009. *Standards in Genomic Sciences* **1**:3, 278-282. [CrossRef]