Standards in Genomic Sciences (2012) **6:**276-286 DOI:10.4056/sigs.2876184

Report of the 13th Genomic Standards Consortium Meeting, Shenzhen, China, March 4–7, 2012

Jack A. Gilbert[1,2], Yiming Bao[3], Hui Wang[4], Susanna-Assunta Sansone[5], Scott C. Edmunds[6], Norman Morrison[7], Folker Meyer[1], Lynn M. Schriml[8], Neil Davies[9], Peter Sterk[5], Jared Wilkening[1], George M. Garrity[10], Dawn Field[4], Robert Robbins[11], Daniel P. Smith[1], Ilene Mizrachi[12], Corrie Moreau[13]

[1]Argonne National Laboratory, Argonne, IL, USA
[2] Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA
[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[4]Centre for Ecology & Hydrology, Wallingford, Oxfordshire, UK
[5]University of Oxford e-Research Centre, Oxford, UK
[6]GigaScience, BGI Hong Kong Ltd., Hong Kong
[7]School of Computer Science, University of Manchester, Manchester, UK
[8]University of Maryland School of Medicine, Baltimore MD, USA
[9]Gump South Pacific Research Station, University of California Berkeley, French Polynesia
[10] Michigan State University, Department of Microbiology and Molecular Genetics, East Lansing, MI, USA
[11] University of California at San Diego, La Jolla, CA, USA
[12] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[13]Field Museum, Chicago, IL, USA

Corresponding author: Jack Gilbert (gilbertjack@anl.gov)

Keywords: Genomic Standards Consortium, microbiome, microbial metagenomics, fungal genomics, viral genomics, Genomic Observatories Network

This report details the outcome of the 13th Meeting of the Genomic Standards Consortium. The three-day conference was held at the Kingkey Palace Hotel, Shenzhen, China, on March 5–7, 2012, and was hosted by the Beijing Genomics Institute. The meeting, titled *From Genomes to Interactions to Communities to Models*, highlighted the role of data standards associated with genomic, metagenomic, and amplicon sequence data and the contextual information associated with the sample. To this end the meeting focused on genomic projects for animals, plants, fungi, and viruses; metagenomic studies in host-microbe interactions; and the dynamics of microbial communities. In addition, the meeting hosted a Genomic Observatories Network session, a Genomic Standards Consortium biodiversity working group session, and a *Microbiology of the Built Environment* session sponsored by the Alfred P. Sloan Foundation.

Introduction

The Genomic Standards Consortium (GSC) held its 13th GSC workshop, *From Genomes to Interactions to Communities to Models* in Shenzhen, China, on March 5–7, 2012. The meeting, hosted by the Beijing Genomics Institute (BGI), included over 100 attendees from more than 20 countries. This was the first GSC meeting held in Asia and represented an opportunity to provide outreach to researchers working in China. The meeting format focused on science enabled by standards, highlighting the breadth of scientific endeavor supported by the work of the GSC community.

The GSC was formed in 2005 with the aim of bringing together the genomics community to improve contextual data quality for genomic sequence data [1]. The GSC community works to build community consensus and promotes community interaction and consultation through meetings, working groups, workshops, and publications. The GSC is an open-member international community consisting of over 200 biologists, bioinformaticians, and computer scientists and includes representatives from the International Nucleotide Sequence Database Collaboration

(DDBJ/ENA/GenBank) and major sequencing centers including Argonne National Laboratory (ANL), the J. Craig Venter Institute (JCVI), Joint Genome Institute (JGI), Institute for Genome Sciences (IGS), and Wellcome Trust Sanger Institute (WTSI).

The GSC creates, maintains, and adopts a range of genomic metadata standards and collaborative projects. The GSC has developed three well-described, minimal information checklists that cover genomes and metagenomes (MIGS and MIMS [2];) and marker gene sequences (MIMARKS [3,4]); that are combined under the "Minimal Information about any Sequence" (MIxS) specification [4]. These checklists are now accompanied by detailed environmental metadata packages that provide standard formats for recording the myriad of environmental parameters (e.g., ammonia concentration, conductivity, wind speed, patient health).

The GSC is constantly striving to facilitate easy adoption of its minimal information standards, including the launch of the GSC journal, *Standards in Genomic Sciences* (SIGS) [5]. Implementation and adoption projects include the Genomic Contextual Data Markup Language (GCDML, an XML data format to support GSC minimal standards) [6], the Genomic Rosetta Stone (GRS, a resolving service for top-level genome and metagenome project information from different resources) [7], Habitat-Lite (a lightweight vocabulary for the environment of any organism or biological sample) [8], and M5 [9,10] which aims to describe tools and infrastructure to cope with the large quantities of metagenomic data generated by projects such as the Earth Microbiome Project [11-14].

GSC13 was structured like a traditional scientific meeting, with keynote presentations and 11 plenary sessions, and a parallel workshop session for two GSC working groups. The workshop was recorded on video by BGI; all talks are accessible on SciVee [15].

Day 1

The theme for Day 1 was genomics enabled by standards, focusing on animal, plant, fungi and viruses.

Session I: Keynote and introduction to the GSC

Day 1 started with a keynote address and welcome session introduced and chaired by Jack Gilbert (Argonne National Laboratory and University of Chicago, USA). The keynote address was provided by Rita Colwell (University of Maryland, USA), who highlighted the power of genomics and metagenomics in uncovering human disease, through comparative genomics of *Vibrio* species, and the use of the metagenomics to determine the environmental etiology of persistent diseases in developing countries. She emphasized the need to use the myriad tools available to us to explore the world in which we live, including the use of satellite mapping to explore remote sensing of microbial dynamics and infection potential via environmental events. Colwell also discussed the need to standardize the way in which this information was collected and disseminated, in order to enable wide-reaching implementation and use of the data. Second to speak in this session was Dawn Field, chair of the Board of the Genomic Standards Consortium, who introduced the GSC, providing historical perspective for the organization and a discussion of the various projects and initiatives being implemented by the GSC within the broader community. Field highlighted the role that data standards play in making sense of the data bonanza that biology is currently experiencing. The session concluded with a direct talk (no slides) from Jun Wang (executive director of BGI), who as the representative of the local host organization welcomed the conference participants to Shenzhen. Wang talked about the power of genomics and metagenomics to help interrogate biology for the benefit of mankind and the role that BGI plays in this activity. He highlighted the exceptional capability of BGIs sequencing and informatics service and identified several key projects as examples of the forward thinking nature of BGI. These included (1) a genome sequencing project that aims to sequence the genome of a million species/varieties, specifically targeted at economically and scientifically important plants and animals and model organisms (e.g., giant panda, potato, *Macaca* genome); (2) the Million Human Genomes Project, which is focusing on large-scale population studies and association studies, using whole genome or whole exome sequencing strategies; and (3) the Million Eco-System Genomes Project, which aims to sequence the metagenome and cultured microbiome of all kinds of environment, including the microenvironment of the human body.

Session II: Megagenome projects I: animal and plant genomics

Session II was chaired by Linda Amaral-Zettler (Marine Biological Laboratory, USA). The session focus was on animal and plant genome projects that were aided by or included data standards formatting. The first speaker was Xiaodong Fang (BGI, China), who discussed the recently sequenced oyster genome and the ongoing effort to resequence the many commercially important species and strains of oysters. Fang discussed the need for transcriptomic sequencing to contextualize the environmental response of specific genomes. The next speaker, Takeshi Itoh (National Institute of Agrobiological Sciences, Japan), highlighted the rice genome, specifically the recent developments for the Rice Annotation Project Database [16], which has provided manually curated functional annotation and other genomics information on the genome assembly of *Oryza sativa* (*japonica*, cv. Nipponbare) since 2005. Itoh emphasized the need for interdatabase standard descriptions of annotations and gene calling, to aid improved strain annotation, as well as better descriptions of environmental contexts for plant cultivars to explore the link between genomic variation and potential phenotypic environmental interaction. Xun Xu (BGI, China) next talked about the BGI Plant Reference Genome database and the need to sequence many different species, and cultivars of every commercially important crop to better understand genomes. Xu highlighted the need for transcriptomic sequencing to better understand plant genomics, including the use of transcriptomic sequencing from different stages of plant development and different plant tissues (e.g., in the potato). Xu also highlighted the use of BAC cloning and subsequent Illumina sequencing of BAC inserts to overcome difficulties associated with heterogeneity, polyploidy, and highly repetitive sequences that make plant genomes difficult to finish.

Session III: Megagenome projects II: viral genomics

After lunch, the third session was cochaired by Hui Wang (Center for Ecology and Hydrology, UK) and Yiming Bao (National Center for Biotechnology Information, USA). The GSC has yet to fully address the needs of the viral community in terms of minimal information checklists to describe viral environments or the quirks of viral genomics. Therefore this session was designed to explore the ongoing work of the viral community in creating their own initiatives and to explore the opportunities to create a viral working group within the GSC in order to determine the appropriate language standards needed by the community. After the chairs' brief discussion of virus prevalence, the knowledge gap, and current technology advances, the first speaker was Charles Chiu (University of California, San Francisco, USA), who discussed the role of clinical metagenomics on the diagnosis and discovery of viral pathogens. Chiu demonstrated the application of his group's automated, cloud-based computational pipeline for identifying viruses in metagenomic microarray and deep sequencing data, which relies heavily on the existence of annotated and well-characterized viral genomes. The pipeline solves significant issues relating to bioinformatic analyses of the data bonanza, such as data storage, parallel processing, portability, and scalability. Opportunities for addressing viral quasi-species by using deep sequencing data were discussed in response to a question.

Next, Ulrich Melcher (Oklahoma State University, USA) discussed issues and problems associated with appropriately sampling and sequencing the virome of different plant species in the Tallgrass Prairie Preserve of Osage County, Oklahoma, USA. Melcher emphasized the need to have contextual information in viral metagenomics, showing that the metagenomic samples collected from a known plant species, with a known life stage, at a known time and location were easier to scientifically interpret to derive biological rationale for the viral community as opposed to blanket sampling of large areas of environment containing many different and potentially unknown plant species. Melcher also stressed the need to differentiate between sequence homology and phenotype, suggesting that "pathogen-like" is not always a relevant statement. When answering questions, Melcher noted that fungal and bacterial viruses can also be detected from the plant materials.

Richard Scheuermann (University of Texas, Southwestern Medical Center, USA) highlighted existing efforts to standardize the recording of pathogenic virus sequence data and metadata. He also described a new U.S. National Institute of Allergy and Infectious Diseases (NIAID) initiative that is helping the scientific community deal with sequencing data volumes in pathogenic virus databases through the support of two resource programs: the Genome Sequence Centers for Infec-

tious Diseases (GSCID), which provides sequencing and analysis services for sample sets of pathogenic microorganisms and invertebrate vectors of disease, and the Bioinformatics Resource Centers (BRC), which integrates genome sequence data with related relevant information to the pathogen research communities.

The fourth speaker in the session was Zhengli Shi (Wuhan Institute of Virology, China), who discussed efforts to describe the transmission of viral communities within bat populations and through freshwater systems, which led to the discovery of many novel viral genetic types enabled through Illumina sequencing followed by bioinformatics based on assembly of short reads. Timothy Stockwell (J. Craig Venter Institute, USA) then talked about the efforts at JCVI to sequence environmental virus communities in a high-throughput pipeline, discussing the need for standard descriptions of the standard operating procedures for exploring these communities. Stockwell described several new techniques for low-cost molecular barcoding to allow high multiplex sequencing using hybrid next generation technologies.

Gane Ka-Shu Wong (University of Alberta, Canada) concluded the session talks with a presentation of efforts to improve viral discovery and tracking viral pathogens in the clinical setting. He emphasized the need to focus on both acute and chronic infections in order to identify viral pathogens that were responsible for different disease states.

Session IV: Megagenome projects III: fungal genomics

The fourth session of the meeting was chaired by Linda Amaral-Zettler (Marine Biological Laboratory, Woods Hole, USA) and was focused on the bioinformatic challenges facing those working with fungal genomics. The first speaker, Jaeyoung Choi (Seoul National University, S. Korea), described the application of all existing and future fungal genomes to a standardized database to enable comparative fungal genomic analysis in one place. The tool, Comparative Fungal Genomics Platform [17], was released in 2007 aiming for a comprehensive bioinformatics workbench with the standard data warehouse. Second, Patrick Chain (Los Alamos National Laboratory, USA) discussed the development of a baseline to describe fungal diversity and new tools for microbial genomic comparisons. Jason Stajich (University of California, Riverside, USA) then discussed the creation of a fungal database for improved annotation and characterization of novel fungal genes. The database, FungiDB [18], is a functional genomic database and website tool for fungal genomes to enable data mining and analyses of the pan-fungal genomic resources. Stajich highlighted the need for improved functional gene annotation in fungal genomics and more reference strain genome sequencing. Kessy Abarenkov (University of Tartu, Estonia) gave the final talk in this session, discussing the implementation of the UNITE database for improved identification of fungal communities in metagenomic data. UNITE [19] is a fungal rDNA ITS sequence database hosted by the PlutoF cloud [20].

Evening session

In the evening Dawn Field (Center for Ecology and Hyology, UK) introduced Jim Tiedje (Michigan State University, USA), who gave an evening lecture on the application of genomics and metagenomics to understanding microbial communities, especially focused on soil systems. He highlighted the data challenges and described a suite of systems capable of answering those challenges. He made a plea to the community to adopt stricter protocols for acquiring and implementing data standards for genomic and metagenomic data, suggesting that no one solution was good enough but that something should be done.

Day 2

The morning session started with a keynote lecture by Mitch Sogin (Woods Hole, USA), who was introduced by Frank Oliver Glöckner (Max Planck Institute for Marine Microbiology and Jacobs University Bremen, Germany). Sogin discussed the implication of the rare biosphere and demonstrated analysis of several projects taken from the International Census of Marine Microbes that, thanks to the extensive metadata recorded for each study during that ICoMM analysis, were exceptionally easy to analyze and explore. He highlighted the need for consistent and updated metadata standard formats and suggested that the MIxS format from GSC [4] was an exceptionally powerful example and hence had been adopted for describing the metadata associated with studies in the ICoMM database. He also discussed the importance of dealing with error rates in new sequencing technologies.

Session I: Interactions: host-associated microbiome projects

Session I of Day 2 was chaired by Ilene Karsch-Mizrachi (National Center for Biotechnology Information, USA), who started by highlighting the development of project descriptions in NCBI Genbank and the efforts required to describe host-associated metadata for these types of microbiome studies. Granger Sutton (JCVI, USA) gave the first presentation highlighting the development of *PanOCT*, a pan-genome ortholog clustering tool for pan-genomic analysis of closely related prokaryotic species or strains. *PanOCT* has been applied to determining the pan-genomic structure of communities associated with human gut studies from the Human Microbiome Project. Second, Junjie Qin (BGI, China) discussed the MetaHIT initiative to discover human gut microbiome community structure in Chinese and European people. He discussed the importance of improving the experimental design to include more representative groups, as well as adding greater longitudinal breadth to improve the detection of population-scale microbiome variation in human communities. Jack Gilbert (Argonne National Laboratory and University of Chicago, USA) then presented some initial work on the Merlot Microbiome Project, an initiative to explore the plant-associated microbiota and its influence in vine health and wine quality in Merlot vineyards. The final talk of the morning was by Corrie Moreau (Field Museum, USA), who presented her work exploring the microbiome associated with different ant species, in different plant-associated relationships. Moreau discussed differences in the core microbiome of predatory and herbivorous ants and notable similarities among distantly related herbivorous ants. She highlighted her work associated with the Earth Microbiome Project [21] and the need to explore more standardization between datasets generated from insect microbiota.

Session II: Microbial metagenomics projects

The second session of the day focused on communities of organisms. The session was chaired by Folker Meyer (Argonne National Laboratory, University of Chicago, USA). The first speaker was Greg Caporaso (University of Northern Arizona, USA), who talked about recent developments in the application of ultra-high-throughput sequencing of microbial communities associated with many different environments, including human, animal, and soil ecosystems. He noted that a single study can generate more than 80 GB of sequence data and that development of tools for the efficient computation of such studies was essential. Caporaso discussed the development of QIIME [22], which was designed to deal with massive amplicon metagenomic datasets. He also presented recent work from extensive time series analyses of microbial communities [23].

Second, Jack Gilbert (Argonne National Laboratory and University of Chicago, USA) discussed the Earth Microbiome Project and detailed some exciting initial results suggesting that the initial exploration of the global microbiome has yielded a considerable amount of diversity, including capturing more than 85% of the known microbial diversity, with 16S rRNA sequencing of 5,000 environmental samples collected along environmental gradients from around the world [21].

Third, James Tiedje (Michigan State University, USA) discussed recent work on the soil metagenomic of various sites around the United States. He also highlighted a recent U.S. National Science Foundation award for a Research Coordination Network, known as the "Terragenome – the Soil Metagenome Network." Its purpose is to facilitate the analysis of soil metagenomes by holding periodic meetings to plan strategies and share information, coordinating sequencing and bioinformatics activities, hosting workshops to train students and scientists in metagenomic analysis, and generally enhancing communication and information sharing.

Fourth, Jacob Parnell (National Ecological Observatory Network, USA) presented NEON Inc [24], a 30-year NSF initiative to provide infrastructure and data from 25 sites around the United States. The mission is to enable understanding and forecasting of climate change, land use change, and invasive species on continental-scale ecology by providing infrastructure and consistent methodologies to support research, education, and policy in these areas. Parnell showed some exciting data collected from 400 samples in four ecological regions, which indicated that spatial and temporal variations are correlated with pH, total carbon and nitrogen, and microbial biomass.

Fifth, Patrick Wincker (Genoscope, Institut de Genomique du CEA, Evry France) presented the TARA Oceans project, which aims to sample the major oceanic systems for small planktonic organisms, viruses, and fish, with exceptionally detailed metadata. Wincker discussed the need for detailed

recording of the data and for standard language that would make sharing of the metadata easier. The final speaker in this session was Bharat Patel (Griffith University, Australia), who discussed ongoing work to characterize the microbial communities thriving in the hot subsurface of Australian aquifers. Patel highlighted the need for extending the environmental package descriptions provided by the GSC to include environments such as the hot aquifer systems with unique chemistry and environmental descriptions.

Session III: Toward a genomic observatories network

The third session looked to the future of genomic research and the benefits to intense studies of specific sites, or "genomic observatories." The session was introduced by Dawn Field (Centre for Ecology and Hyology, UK), who described efforts to date aimed at organizing a set of leading sites championing "omics" science into a global "Genomic Observatories Network" [25]. This introduction was followed by presentations from two Genomic Observatories (GOs; Moorea and L4), two presentations from researchers pioneering site-based research including GOs, and an introduction to the work of the GEO BON community, an international effort to undertake biodiversity observations at the international scale. Crucial to all these projects is access to uniform methods of sampling and describing data, a key reason for GOs to engage heavily with the GSC, now and in the future.

The first speaker, Neil Davies (UC Berkeley - Moorea, USA), described work on the Moorea Genomic Observatory, which is home to the GBMF-funded Moorea Biocode Project that is DNA barcoding all organisms larger than 1 mm on the island of Moorea in French Polynesia. Jack Gilbert (Argonne National Laboratory and University of Chicago/, USA) then described efforts to characterize microbial communities at the L4 Genomic Observatory in the Western Channel Observatory using a combination of metagenomic analysis and modeling. Linda Amaral-Zettler (Marine Biological Laboratory, Woods Hole, USA) spoke about her work studying microbial diversity across the aquatic Long Term Ecological Research (LTER) sites that are funded by the U.S. National Science Foundation, including the Moorea Coral Reef LTER. Frank Oliver Glöckner (Max Planck Institute for Marine Microbiology and Jacobs University Bremen, Germany) introduced the newly funded Micro B3 Project: Biodiversity, Bioinformatics, Biotechnology, which includes an ambitious global Ocean Sampling Day (OSD) planned for the solstice of 2014 (June 21). Work at the L4 GO has observed a dip in microbial diversity on the longest day of the year in the northern hemisphere. The OSD megasequencing project already has more than 30 subscribed participants, including several GOs (Moorea, L4, etc.). Makiko Mimura (Kyushu University, Japan) then discussed the topic of building linkages between Genomic Observatories and GEO BON. The formal talks were followed by a short panel discussion involving all the speakers about how to best build this network.

Session IV: Policies and standards for reproducible research: from theory to practice

The second themed session of the afternoon brought together a diverse group of speakers with different roles in the production, dissemination, and use of data, to discuss the role of policies and standards enabling reproducible research and data sharing [26]. The session was introduced by Susanna-Assunta Sansone (University of Oxford, UK) and Scott Edmunds (GigaScience, BGI, China). Sansone presented the BioSharing initiative [27] that—building on the effort of the widely known Minimum Information for Biological and Biomedical Investigations effort (MIBBI [28];)—works to strengthen collaborations between researchers, funders, industry, and journals and to discourage redundant (if unintentional) competition between standards-generating groups.

The second introductory talk, from Edmunds, focused on the issues and additional incentives needed to enable data dissemination. Covering work that BGI's *GigaScience* journal and database has done to release datasets with citable DOIs, Edmunds demonstrated the utility of releasing genomes before publication, citing the subsequent crowd-sourcing of the deadly 2011 *E. coli* O104:H4 outbreak genome sequenced by the BGI [29].

The perspective of funders—being key gatekeepers able to enforce and influence data policies and standards—was then covered. Rita Colwell providing her wealth of experience as former director of the NSF. Paula Olsiewski (Alfred P. Sloan Foundation) presented challenges and opportunities of the Microbiology of the Built Environment Program, focusing on one of its objectives to improve the cohesiveness of the community and its ability to communicate internally and externally

by developing data visualization and imaging techniques, and repositories.

The next group of talks covered "Breaching the Bio-Domain," providing a more hands-on point of view from data producers, curators, and database managers. Philippe Rocca-Serra (University of Oxford, UK) introduced the ISA Commons [30], a growing community —including GSC members— that uses a common metadata tracking framework to facilitate standards-compliant collection, curation, management, and reuse of multi-omics datasets in an increasingly diverse set of life science domains, including genomics and metagenomics [31-33].

Srikrishna Subramanian (Institute of Microbial Technology, India) shared the lessons learned from his experience in developing a community-regulated collaborative knowledge environment that has enabled researchers in the field of structural genomics to annotate and extend the structural data to discover functional insights [34]. Folker Meyer (Argonne National Laboratory, USA) talked of his experience running MG-RAST, noting that of the 41,000 datasets in the database, only a minority are publicly accessible, and appealing for funders to insist that data from projects they have funded be publicly available. Yong Zhang (BGI Shenzhen, China) gave a "data-producer" and BGI perspective, outlining the scale of the challenges ahead and previewing some of the work under way to build biobanks and data centers that will become the China National Genebank.

The session ended with final perspectives from journal editors, with Clare Garvey (*Genome Biology*, BioMed Central) and Craig Mak (*Nature Biotechnology*, Nature PG) giving overviews of their journal policies and examples of their publishers' efforts in aiding data sharing and standardization.

Day 3

Session I: The Alfred P. Sloan Foundation Microbiology of the built environment session

The first session was a special one embedded within GSC13 and funded by the Alfred P Sloan Foundation's (APSF) Microbiology of the Built Environment program. The session was organized by Jack Gilbert (Argonne National Laboratory and University of Chicago, USA), and introduced by the APSF Program Officer Paula Olsiewski. Olsiewski highlighted the difficulty she had experienced in convincing microbiologists to come into the indoor environment to explore the microbial world. She also highlighted the need for improved standards in describing the physical, chemical, and biological parameters of the indoor world, highlighting several key talks in this session.

The first speaker was Jeffrey Siegel (University of Texas at Austin, USA), who discussed the environmental factors that should be measured in buildings in order to understand the drivers of microbial diversity indoors. Siegel presented work exploring the impact of mechanical systems, ventilation, occupant demographics and history, and abiotic and inhibitor contaminant concentrations. He showed some exciting work on characterizing the metal contaminants in dust samples, showing new capability for exploring the "health" of indoor air.

Second, Lynn Schriml (University of Maryland School of Medicine, USA) discussed the development of an environmental package to complement MIMS and MIMARKS standard metadata reports. This new initiative brought together researchers, architects, and the Microbiology of the Built Environment Data Analysis Core (MoBeDAC, mobedac.org). The new standard provides information on samples collected, sequenced, and annotated with MIxS-BE metadata (Built Environment) for wastewater, air filters, air, and surfaces of indoor spaces.

Third, Daniel Smith (Gilbert Laboratory - Argonne National Laboratory, USA) presented the Home Microbiome Project, which aims to categorize the rate and intensity of human skin microbiome and house surface microbiome interactions. This study is looking at how quickly and in what direction microbial life moves between human and house when a house is newly occupied. Evidence presented by Smith suggested that in certain instances, the floor of people's houses becomes inoculated with the dominant microbial group on the soles of the new occupant's feet within six days of occupancy. This "citizen science" study is ongoing.

Fourth, Scott Kelley (San Diego State University, USA) presented work funded by Sloan on the indoor virome, exploring the viral diversity that exists within homes, hospitals, and work places. He also presented data on the microbiology of bathrooms, neonatal hospital environments, and offices and highlighted the need for recording as many environmental parameters about each environment as possible if the drivers of diversity are to be determined.

Mitch Sogin (Marine Biological Laboratory, Woods Hole, USA) presented the first of three talks outlining the MoBeDAC initiative. Sogin discussed the role of the VAMPS package (Visual Analysis of Microbial Population Structures [33]) in analyzing data generated by research in this program. The VAMPS component is linked to QIIME in that it generates QIIME compatible output from the VAMPS analysis and interpretation of the 16S sequencing data. Jason Stajich (University of California, Riverside, USA) presented the MoBeDAC fungal database study. This database is being rolled out to the MoBeDAC partners, QIIME, VAMPS, and MG-RAST, so that fungal ITS or rRNA data can be better characterized using these systems. Jack Gilbert (Argonne National Laboratory and University of Chicago, USA), as a last-minute replacement for Folker Meyer (Argonne National Laboratory and University of Chicago, USA), presented Meyer's overview of MoBeDAC and the interface language between VAMPS, QIIME, and MG-RAST.

Session II: The RCN4GSC GSC biodiversity working group session

The second session of Day 3 was sponsored by the NSF-RCN4GSC and brought together speakers with a specific interest in molecular biodiversity. The session was organized by Robert Robbins (UCSD, USA) and Norman Morrison (NERC Environmental Bioinformatics Centre & The University of Manchester, UK). Robbins began the session by introducing the audience to the aims of the GSC Biodiversity Working Group (GBWG), including recent and future milestones in the activities program. Highlights included the outputs from a workshop held in Oxford sponsored by the Global Biodiversity Information Facility (GBIF) that took some important steps toward harmonizing the Darwin Core and MiXS reporting standards [35]. Robbins also noted that upcoming biodiversity-GSC events will include a "semantics of biodiversity" ontology workshop, to be held May 2012 at the University of Kansas, and continued engagement with the Asian biodiversity community at the TDWG Annual Meeting, to be held in Beijing in the fall of 2012.

Emphasizing the importance of conceptual standards, Robbins observed that calls for nomenclatural standards can be found as far back as the Analects of Confucius (13:3):

名不正 則言不順 言不順 則事不成.

If names be not correct, language is not in accordance with the truth of things. If language be not in accordance with the truth of things, affairs cannot be carried on to success.

The rectification of names (正名)—the establishment and harmonization of standards—is a critical first step in any endeavor and is especially important when two groups with differing prior standards begin to interact, as has been occurring be the genomics-metagenomics communities and the traditional biodiversity community. Much of the GBWG work to date has focused on harmonizing standards between communities. Analyzing the conceptual underpinning of that harmonization will be a key goal of the forthcoming semantics-of-biodiversity workshop.

Norman Morrison introduced the BioVeL project [36], a virtual e-laboratory that supports research on biodiversity issues using large amounts of data from cross-disciplinary sources. Morrison demonstrated how researchers can use the virtual laboratory to build their own workflows by selecting and applying successive "services" (data-processing techniques) or by reusing existing workflows available from BioVeL's online library. BioVeL is a consortium of fifteen partners from nine countries and is funded through the European Community 7th Framework Programme. Frank Oliver Glöckner (Max Planck Institute for Marine Microbiology and Jacobs University Bremen, Germany) gave an introduction to the newly started European MicroB3 project [37] and described how the project proposes to connect the B's in the project (Biodiversity, Bioinformatics, and Biotechnology), emphasizing that capturing contextual data such as geolocation is integral to the function of the project. Hiroshi Mori (Tokyo Institute of Technology, Japan) introduced environmental contextualization in MicroDB [38], including important work to extend a number of ontologies to enable contextualization within a semantic integration framework. Neil Davies (UC Berkeley, USA) discussed how we can layer information from the same place to build a full picture of interactions in model ecosystems. Davies argued that this was a fundamental building block for ecosystem services, because having this full picture would help us to plan. The final speaker of the session was Linda Amaral Zettler (Marine Biology Laboratory, USA), who discussed the Life in a Changing Ocean Project, a next-generation science program with the goal of using biodiversity discovery

and knowledge to support healthy and sustainable ocean ecosystems. Amaral Zettler again stressed how the understanding of baselines will help inform decision making at a global level and how biodiversity is intimately related with all aspects of our livelihood and can impact it in many ways.

Session III: GSC projects and activities

The aim of the third session was to provide the audience with an overview of GSC projects and activities. The session was organized by Peter Sterk (University of Oxford, UK) and Lynn Schriml (University of Maryland, USA). Sterk started the session with a brief history of the GSC. He then described the Minimum Information about any Sequence (MIxS) family of standards (MIGS/MIMS + MIMARKS [4] in some detail and gave examples of organizations and projects that have adopted the MIxS standards. Next, George Garrity (Michigan State University, USA), editor-in-chief of the GSC's eJournal *Standards in Genomic Sciences* (SIGS [39];) explained the focus and scope of the journal and gave a brief overview of its almost three-year history. The journal has published over 200 articles, most of which are short genome reports. SIGS is currently the third largest publication of short genome reports, and it is anticipated that SIGS will be the largest in the near future. Papers are available for download from the SIGS website as well as through PubMed Central. Sterk then continued with his overview of the GSC projects already briefly mentioned on the first day of this meeting, including the Genomic Contextual Data Markup Language (GCDML), an XML schema, which is a reference implementation of the MIxS family of standards (reference); the Genomic Rosetta Stone (GRS), which provides a mapping of genomic identifiers across a wide range of databases; and Habitat-Lite or EnvO-Lite, a small set of terms describing diverse environments. Jared Wilkening (Argonne National Laboratory, USA) presented the M5 project [2]. Its aim is to build tools and standards to enable sharing of high-volume data and computation using a consensus-driven approach. Wilkening briefly described M5nr, an effort to join a number of different data sources into a single nonredundant database [9,40] and the Metagenome Transport Format (MTF), a format for sharing sequences and computation.

Sterk then introduced the audience to the GSC's MIxS Compliance and Database Interoperability Working Group, a group of volunteers working on standards and tools development; and he encouraged members of the audience to join. He also reminded the audience of the existence of the GSC's Biodiversity Working Group introduced during the previous session. He concluded the session by pointing out that GSC membership is open and free and asked the audience to consider joining.

Session IV: Parallel sessions – GSC working groups

MIxS working group

The MIxS Working Group session discussed the content and structure of the MIxS standard. the suggestion was made to split the MIxS core data into two subsets: sample and analysis metadata. The steps to include the new MIxS-Built Environment package into NCBI's BioSample submissions were reviewed. Through participant suggestions, the group discussed the process to coordinate mapping of the MIxS to newer standards initiatives (NIAID/GRC, FDA/CDC).

Virus standards working group

The Virus Standards Working Group session had an intense discussion on metadata standards for viruses. The participants recognized the importance of dividing metadata into different categories or modules, for example, clinical and environmental. While symptoms are critical pieces of metadata, some suggested that a well-defined list of symptoms should be established for viral diseases so that submitters can focus on those rather than an extensive list. Sample preparation and treatment methods should be included. Passage history also is important for viral samples and therefore should be added as a source qualifier in sequence records (a recommendation to INSDC). Geographical location can be sensitive sometimes, and thus granularity should be allowed. It would also be beneficial to have controlled vocabularies for isolation source. The Working Group session also looked at a sequence generated by NGS in the NCBI Sequence Viewer. They appreciated the alignment view and statistics of SNPs and suggested that it would be useful to be able to see quality scores of the reads and to filter the SNP statistics by quality scores. In addition, the Working Group briefly discussed what qualified as a virus sequence, especially when there is very low or no similarity to known viral sequences, and how such a sequence should be named in sequence databases. Based on the interest in this topic, the participants proposed forming a Virus Working Group within GSC.

GSC13-GSC14 handover and meeting close

The meeting was formally closed with a handover from GSC13's chief organizer, Jack Gilbert, to GSC14's chief organizer, Dawn Field. Field discussed the principal topics that of discussion planned for GSC14 (to be held at Oxford University, September 17–19, 2012). Of special interest is the Genomic Observatories Network, which aims to provide a coordinated approach to the generation and recording of genomic sequence data from long-term environmental monitoring sites around the world [41].

The GSC13 meeting raised questions regarding how to improve adoption of GSC standards within the community. This ongoing problem will require the GSC community to lower the barrier to compliance by enabling researchers to easily adopt the standard relevant to their research initiatives.

Acknowledgments

This work was supported in part by the US Department of Energy under Contract DE-AC02-06CH11357 and in part by the US National Science Foundation through the research coordination network award RCN4GSC, DBI-0840989. We thank Eppendorf, MoBio, BGI, Lucigen, and Hua Yue Enterprise Holdings Ltd. for their sponsorship of the meeting. We also thank the Gordon and Betty Moore Foundation for support and the U.S. Department of Energy for supporting the attendance of Greg Caporaso, Daniel Smith, Andreas Wilkening, Austin Davis-Richardson, and Patrick Chain through a young investigators travel award.

References

1. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi, *et al*. The Genomic Standards Consortium. *PLoS Biol* 2011; **9**:e1001088. http://dx.doi.org/10.1371/journal.pbio.1001088 PubMed

2. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, *et al*. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008; **26**:541-547. http://dx.doi.org/10.1038/nbt1360 PubMed

3. Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G, Nakamura Y, Sansone SA, Glöckner FO, Field D. The genomic standards consortium: bringing standards to life for microbial ecology. *ISME J* 2011; **5**:1565-1567. http://dx.doi.org/10.1038/ismej.2011.39 PubMed

4. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 2011; **29**:415-420. http://dx.doi.org/10.1038/nbt.1823 PubMed

5. Garrity GM, Field D, Kyrpides N, Hirschman L, Sansone SA, Angiuoli S, Cole JR, Glöckner FO, Kolker E, Kowalchuk G, *et al*. Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 2008; **12**:157-160. http://dx.doi.org/10.1089/omi.2008.A2B2 PubMed

6. Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, Lombardot T, Field D, Glöckner FO. A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* 2008; **12**:115-121. http://dx.doi.org/10.1089/omi.2008.0A10 PubMed

7. Van Brabant B, Gray T, Verslyppe B, Kyrpides N, Dietrich K, Glöckner FO, Cole J, Farris R, Schriml LM, De Vos P, *et al*. Laying the foundation for a Genomic Rosetta Stone: creating information hubs through the use of consensus identifiers. *OMICS* 2008; **12**:123-127. http://dx.doi.org/10.1089/omi.2008.0020 PubMed

8. Hirschman L, Clark C, Cohen KB, Mardis S, Luciano J, Kottmann R, Cole J, Markowitz V, Kyrpides N, Morrison N, *et al*. Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *OMICS* 2008; **12**:129-136. http://dx.doi.org/10.1089/omi.2008.0016 PubMed

9. Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, Desai N, Eisen JA, Evers D, Field D, *et al*. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand Genomic Sci* 2010; **3**:243-248. http://dx.doi.org/10.4056/sigs.1433550 PubMed

10. Metagenomics, Metadata, MetaAnalysis, Models and MetaInfrastructure. http://gensc.org/gc_wiki/index.php/M5.

11. Gilbert JA, Bailey M, Field D, Fierer N, Fuhrman J, Hu B, Jansson J, Knight R, Kowalchuk G, Kyrpides NC, *et al*. The Earth Microbiome Project: The Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13th-15th 2011. *Stand Genomic Sci* 2011; **5**:243-247. http://dx.doi.org/10.4056/sigs.2134923

12. Gilbert JA, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J, Ley R, Fierer N, Field D, Kyrpides N, *et al*. The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. *Stand Genomic Sci* 2010; **3**:249-253. http://dx.doi.org/10.4056/aigs.1443528 PubMed

13. Gilbert JA, Meyer F, Knight R, Field D, Kyrpides N, Yilmaz P, Wooley J. Meeting report: GSC M5 roundtable at the 13th International Society for Microbial Ecology meeting in Seattle, WA, USA August 22-27, 2010. *Stand Genomic Sci* 2010; **3**:235-239. http://dx.doi.org/10.4056/sigs.1333437 PubMed

14. Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman J, Hugenholtz P, Meyer F, Stevens R, Bailey M, et al. Designing Better Metagenomic Surveys: The role of experimental design and metadata capture in making useful metagenomic datasets for ecology and biotechnology. *Nat Biotechnol* 2012; (In press).

15. SciVee. http://www.scivee.tv/node/46384

16. Rice Annotation Project Database. http://www.rapdb.dna.affrc.go.jp

17. Comparative Fungal Genomics Platform. http://cfgp.snu.ac.kr

18. Fungi DB. http://FungiDB.org

19. A molecular database for the identification of fungi. http://unite.ut.ee

20. PlutoF cloud. http://plutof.ut.ee

21. Earth Microbiome Project. www.earthmicrobiome.org

22. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; **7**:335-336. http://dx.doi.org/10.1038/nmeth.f.303 PubMed

23. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, *et al*. Moving pictures of the human microbiome. *Genome Biol* 2011; **12**:R50. http://dx.doi.org/10.1186/gb-2011-12-5-r50 PubMed

24. National Ecological Observatory Network. http://www.neoninc.org

25. Davies N, Field D. Sequencing data: A genomic network to monitor Earth. *Nature* 2012; **481**:145. http://dx.doi.org/10.1038/481145a PubMed

26. Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, *et al*. Megascience. 'Omics data sharing. *Science* 2009; **326**:234-236. http://dx.doi.org/10.1126/science.1180598 PubMed

27. BioSharing. www.biosharing.org

28. MIBBI. www.mibbi.org

29. Rohde H, Qin J, Cui Y, Li D, Loman NJ, Hentschke M, Chen W, Pu F, Peng Y, Li J, *et al*. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N Engl J Med* 2011; **365**:718-724. http://dx.doi.org/10.1056/NEJMoa1107643 PubMed

30. Commons ISA. www.isacommons.org

31. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, *et al*. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 2010; **26**:2354-2356. http://dx.doi.org/10.1093/bioinformatics/btq415 PubMed

32. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, *et al*. Toward interoperable bioscience data. *Nat Genet* 2012; **44**:121-126. http://dx.doi.org/10.1038/ng.1054 PubMed

33. Analysis of Microbial Population Structures. http://vamps.mbl.edu

34. Krishna SS, Weekes D, Bakolitsa C, Elsliger MA, Wilson IA, Godzik A, Wooley J. TOPSAN: use of a collaborative environment for annotating, analyzing and disseminating data on JCSG and PSI structures. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2010; **66**:1143-1147. http://dx.doi.org/10.1107/S1744309110035736 PubMed

35. MiXS. http://www.gbif.org/communications/news-and-events/showsingle/article/genomic-data-in-gbif-moves-a-step-closer

36. BioVeL. www.biovel.eu

37. European MicroB3 project. www.microb3.eu

38. Micro DB. www.microdb.jp

39. Standards in Genomic Sciences. http://sigen.org

40. M5nr. http://tools.metagenomics.anl.gov/m5nr

41. Genomic Observatories Network. www.genomicobservatories.org