

A Strategy Capitalizing on Synergies: The Reporting Structure for Biological Investigation (RSBI) Working Group

SUSANNA-ASSUNTA SANSONE,¹ PHILIPPE ROCCA-SERRA,¹ WEIDA TONG,²
JENNIFER FOSTEL,³ NORMAN MORRISON,⁴ ANDREW R. JONES,⁵
and RSBI Members⁶

ABSTRACT

In this article we present the Reporting Structure for Biological Investigation (RSBI), a working group under the Microarray Gene Expression Data (MGED) Society umbrella. RSBI brings together several communities to tackle the challenges associated with integrating data and representing complex biological investigations, employing multiple OMICS technologies. Currently, RSBI includes environmental genomics, nutrigenomics and toxicogenomics communities, where independent activities are underway to develop databases and establish data communication standards within their respective domains. The RSBI working group has been conceived as a “single point of focus” for these communities, conforming to general accepted view that duplication and incompatibility should be avoided where possible. This endeavour has aimed to synergize insular solutions into one common terminology between biologically driven standardisation efforts and has also resulted in strong collaborations and shared understanding between those in the technological domain. Through extensive liaisons with many standards efforts, several threads have been woven with the hope that ultimately technology-centered standards and their specific extensions into biological domains of interest will not only stand alone, but will also be able to function together, as interchangeable modules.

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

INTRODUCTION

WHEN THE FIRST MICROARRAY EXPERIMENTS were published, it became apparent that the lack of capture of adequate biological metadata impeded the interpretation of the experiments across the scientific

¹EMBL-EBI, European Nutrigenomics Organisation (NuGO), Hinxton, Cambridge, United Kingdom.

²FDA's National Center for Toxicological Research (NCTR), Center for Toxicoinformatics, Jefferson, Arkansas.

³NIEHS National Center for Toxicogenomics, Research Triangle Park, North Carolina.

⁴NERC Bioinformatics Centre, Oxford Centre for Ecology and Hydrology, Oxford, and School of Computer Science, University of Manchester, Manchester, United Kingdom.

⁵School of Computer Science, University of Manchester, Manchester, United Kingdom.

⁶The listed authors moderate the Environmental Genomics, Nutrigenomics, and Toxicogenomics communities; see the RSBI website, under the MGED Society umbrella, for a complete list of contributing groups: (www.mged.org/Workgroups/rsbi).

community. The Minimum Information About a Microarray Experiment (MIAME) (Brazma et al., 2001) is a reporting structure written in response to this lack, by a group of biologists, computer scientists, and data analysts. MIAME aims to define the minimum information that needs to be reported (in some detail and in a structured way) to interpret unambiguously and potentially reproduce and verify a microarray experiment. This group then went on to organize its composition by founding the Microarray Gene Expression Data (MGED) Society, and develop a standard object model and an ontology to represent microarray experiments (Ball and Brazma, *this issue*). The response from the scientific community has been extremely positive, and currently most of the major scientific journals and funding agencies require publications describing microarray experiments to comply with MIAME.

Undoubtedly, the success of this reporting structure also relies on the exploitation of a simple but clear formulation, in the form of a checklist, which has favored its adoption by a large number of communities (Quackenbush, 2004). For example, as microarrays are incorporated into complex biological investigations such as environmental genomics, nutrigenomics, and toxicogenomics, it has become apparent that analogous minimal descriptors should be identified for these application fields. There have been three extensions to MIAME to date, describing further this complex biological context that accompanies the microarray data and conclusions. MIAME/Tox provides a structured framework that includes the conventional toxicology component of an experiment (e.g., clinical chemistry, histopathology) (Mattes et al., 2004). Developed by the MGED Toxicogenomics working group, MIAME//Tox has been extended by a larger consortium of stakeholders from the private and public sectors into a data dictionary containing terms, definitions, and relationships (Fostel et al., 2005). MIAME/Env fulfills the diverse needs of those working on the functional genomics of environmentally relevant organisms, which are not covered by the model organism community (Morrison et al., *this issue*). MIAME/Nut seeks to provide a structured annotation that includes descriptors for the nutritional component of the experiments (Garosi et al., 2005).

A STRATEGY CAPITALIZING ON SYNERGIES

Discipline-specific initiatives are important because they target “real world” reporting requirements. A consequence of this, however, is that knowledge can become fragmented, resulting in an unnecessary duplication of descriptors across these reporting structures or problems when third parties try to use these. For example, when reporting an eco-toxicogenomics investigation, should one comply with the MIAME/Env or MIAME/Tox requirements, or combine the two? It has also become evident that when other OMICS technologies, such as proteomics and metabolomics, are used in combination with microarrays, these MIAME-based checklists will soon be insufficient to serve the scope of an experimenter’s needs. Although reporting structures, using similar principles to MIAME, are being developed for proteomics, metabolomics, flow cytometry, *in situ* hybridization, and immunohistochemistry experiments, these are designed to be technology centered (Taylor et al., *this issue*; Fiehn et al., *this issue*; Spidlen et al., *this issue*; Deutsch et al., *this issue*). If taken individually, these reporting structures will not be sufficient to fully describe environmental genomics, nutrigenomics, and toxicogenomics investigations. This is true also for ontologies and exchange formats, the other two pillars of data communication standards. The reuse of the same semantics and syntax will benefit the entire scientific community by simplifying the job of data integration, but it will also ease the task of software vendors and equipment manufacturers by reducing costs and time for implementing standards-compliant products. It is necessary therefore that technology-centered standards and their specific extensions into biological domains of interest are developed not only to stand alone but also to function together as interchangeable modules that fulfil the needs of experimenters who are using “multi-omics” approaches to their biological research. To achieve this aim, from a technical perspective, it will be necessary to remove redundancies and fill the gaps between the domains covered by these reporting structures, ontologies, and exchange formats. This is a difficult but not insurmountable task. By contrast, the sociological barriers could be quite challenging, and extensive liaison would be necessary among communities that generally are only loosely connected.

Fortunately, investigators in many different areas of biology have stumbled over this same problem, and a shared interest in overcoming such obstacles is often a good catalyst for producing synergy. Conforming

TABLE 1. COMMUNITIES COLLABORATING IN REPORTING STRUCTURE FOR BIOLOGICAL INVESTIGATION (RSBI)

<i>Participating organizations</i>	<i>URLs</i>
EMBL—European Bioinformatics Institute (EBI)	www.ebi.ac.uk/microarray/Projects/tox-nutri
European Nutrigenomics Organization (NuGO)	www.nugo.org
FDA National Center for Toxicological Research (NCTR), Center for Toxicoinformatics	www.fda.gov/nctr/science/centers/toxicoinformatics/
ILSI Health and Environmental Sciences Institute (HESI) “Genomics Committee”	http://hesi.ilsil.org/index.cfm?pubentityid=120
Natural Environmental Research Council (NERC) Environmental Bioinformatic Centre (NEBC)	http://envgen.nox.ac.uk
NIEHS National Center for Toxicogenomics (NCT)	www.niehs.nih.gov/nct

RSBI Working Group webpage: www.mged.org/Workgroups/rsbi.

to this prediction, for example, the MGED Toxicogenomics working group soon broadened its scope to include the nutrigenomic and environmental genomic communities. The resulting Reporting Structure for Biological Investigations (RSBI; www.mged.org/Workgroups/rsbi/rsbi.html) working group, established in 2004, represents these communities operating in the functional genomics field, where efforts are underway to establish data communication standards and to develop software and databases to support their large user base (Table 1).

In this article, the authors, who are coordinators of the RSBI communities, will (a) present the work that has led to the proposed framework for reporting biological investigations using a multi-OMICS approach, (b) describe the threads woven through extensive liaisons with several standards efforts, and (c) explain how this initial work contributes to the development of an ontology and model for functional genomics.

BUILDING A REPORTING STRUCTURE FOR BIOLOGICAL INVESTIGATION

As addressed previously, neither MIAME-based checklists nor those developed around specific technologies will be sufficient to report environmental genomics, nutrigenomics and toxicogenomics investigations, if taken individually. These investigations are complex in nature and the information is highly nested; a series of events occur over time and one or more omics technologies are employed in combination with other more conventional methods. Defining a full-blown reporting structure for these investigations, however, has never been seen as a trivial task. The RSBI working group is also aware that what could be a minimum and sufficient requirement in one case (or domain) may be inappropriate or insufficient for another. For this reason, the working group has decided to focus on identifying and describing the “high-level concepts” of this reporting structure that encompass any biological or technical domain of application and leave to each biological community the task of extending it into their domains. These “high-level concepts” propose a way to avoid redundancies between individual reporting structures, allowing core biological descriptors to be shared, as well as descriptors relating to the design of investigations, sample generation, treatments and also instruments. In the light of this experience, we believe this to have been a wise decision, as undertaking such work, even at a general level, has proved very challenging.

The representation of the information in these “high-level concepts” should also be consistent with the words that experimentalists use. As a first step, therefore, the RSBI working group has surveyed its communities and explored the flow of information that occurs during the experimental process, using a top-down approach. Each coordinator gathered a wide array of real-life investigations’ designs by a combination of direct interviews with the domain experts (experimentalists and bioinformaticians), reading their research grants, extrapolating concepts from their existing data models, and reviewing the literature. Our scenario also involved geographically distributed domain experts. To enhance the representation of the information, during the knowledge elicitation phase, the domain experts were asked to present their experiments graphically in order to allow clarification and semantic characterisation of the concepts (Garcia et

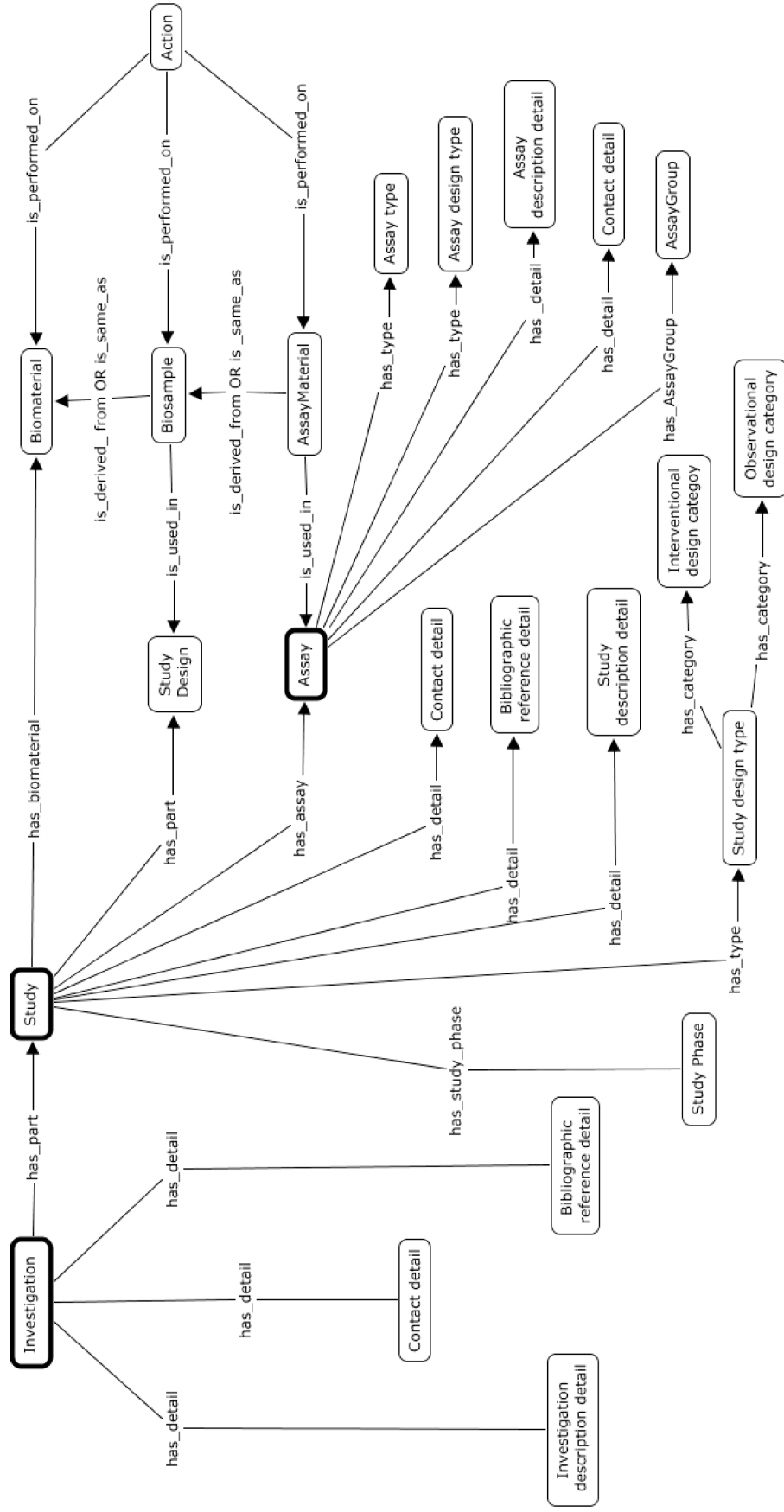


FIG. 1. Some of the Reporting Structure for Biological Investigation (RSBI) 'top-level concepts' and their relationships, illustrated using a freely available concept mapping tool, Cmap (<http://cmap.ihmc.us>).

al., 2006). As a second step, in early 2005, the coordinators of the RSBI communities met up and during a one-week workshop compiled the use cases, and broke down the information contained therein into elementary concepts. Identifying a core set of common concepts across the RSBI communities has been carried out as an iterative process with the following founding principles in mind: (a) concepts should represent information relevant to the understanding of the experimental process; (b) concepts should be meaningful also in the context of another biological community. A result, a minimum set of core concepts has been consistently identified across the different groups of experimentalists, irrespective of their domain of research.

In order to describe the interaction of different technologies during the course of a scientific endeavour the RSBI working group concluded that there was a need for three high-level abstractions where to place the information relevant to the biology as well as that relevant to the different technologies employed. An *Investigation* is a self-contained contained unit of scientific enquiry, with a holistic hypothesis or objective, and a *Design* is defined by the relationships between *Study(s)* and *Assay(s)*. While the first is a “container” for the description of and steps performed on the *Subject(s)*, the latter contains the test(s)—and its produced data—performed on the *Subject(s)*. A *Study* is an experiment with distinct periods or *Phase(s)*, having experimental procedure, or *Action(s)*, applied to *Subject(s)* or *Group(s)* of these. A *Subject* is the biological material under investigation (e.g., a population, an individual, a tissue slice, cells). A *Group* consists of biological replicates (i.e., *Subjects* exposed to similar conditions under investigation). An *Assay* is an experiment carried out on whole or part of the *Subject(s)*, employing OMICS based or other technologies, producing data for computational purpose. Figure 1 shows some of these “top-level concepts” and their relationships.

In current reporting structures, the above concepts are duplicated across different standards but leave little space for the description of complex series of events occurring during the *Study*. Additionally, the current reporting structures are designed around one technology (or type of *Assay*) for example microarray for MIAME. According to the proposed organization by the RSBI working group an *Assay* would be a microarray or a mass-spectrometry experiment (shared by proteomics and metabolomics domains), a histopathology examination, a set of biometrics or *in situ* hybridization. The proposed framework has already been adopted and extended into a toxicogenomics data dictionary containing terms, definitions and relationships (Fostel et al., 2005) and currently it forms the basis of environmental module (Morrison et al., *this issue*).

The MIAME trend is likely to progress and more reporting structures are yet to come. It is essential, however, that these are not created in isolation. Their formulation should attempt to anticipate the need of functional genomics and systems biology and their design should allow the different reporting structures to function together as interchangeable modules. The RSBI working group sees that this would only be possible if coordination were to occur within and between technology-centered and biologically driven standardization efforts, in two main areas: (1) description of biological material (e.g., provenance, storage, treatments), here defined as *Study*, and (2) description of analysis techniques, here defined as *Assay*. This framework proposes a way to facilitate this coordination by restructuring the formulation of this reporting structure. With this proposal plan it is the intention of the RSBI working group to spark a discussion between the initiatives and call for some concrete actions.

WEAVING THE THREADS

Although the RSBI currently is under the MGED Society umbrella, it has always been the intent of this working group to work in the wider functional genomics context. The working group has established a strong alliance with the HUPO Proteomics Standards Initiative (PSI) (Taylor et al., *this issue*), which currently tracks the work of RSBI, both to guide HUPO’s internal development and also to assist in promoting the development of a modular reporting structure for biological investigation. The RSBI community coordinators also serve as moderators of the Biological environment, nutrition and toxicology subgroups and the Ontology working group with the nascent Metabolomics Standardization Initiative (MSI, <http://msi-workgroups.sourceforge.net> under the Metabolomics Society (Fiehn et al., *this issue*). The RSBI environmental group is also working in collaboration with the Genomic Standards Consortium (GSC;

(www.genomics.ceh.ac.uk/genomecatalogue/gsc.php) (Morrison et al., *this issue*) to formalise the description of genomic sequence metadata under the Minimum Information about a Genome Sequence (MIGS) project (Field et al., *this issue*). RSBI aims to be fully open and inclusive and the community coordinators will continue outreaching to other efforts (Bruskiewich et al., *this issue*).

Data standardization is now considered beyond the research application of high throughput technologies and several initiatives have been set up also to address management issues (format and reporting structures for the exchange of information) and database interoperability. Regulatory-driven efforts, aiming for a broader understanding and use of OMICS data, are working to define data models for data submission to regulators. As these technologies are used in industry and are considered by regulatory agencies, the methodology itself comes under scrutiny, and validation programs and production of standard materials and methods are now the focus of many initiatives. Global organizations have also initiated a dialogue between technological experts, regulators, and the principal validation bodies to draw road maps for development, validation and regulatory use of OMICS-based technologies, particularly in chemical assessment. Others are liaising with different life science disciplines offering support, mediation, and consultancy to speed up the standards development process. The RSBI community coordinators and members are also directly participating in these initiatives. In 2004, the working group published a review of these different ongoing standardization efforts, with particular focus on toxicological and environmental applications (Sansone et al., 2004), highlighting their key objectives and target audiences and also identifying opportunities for synergies. The value of this extensive involvement and liaisons with these other efforts is also visible in the discussions initiated around MIAME and the formulation of the MIAME/Tox and MIAME/Env reporting structures. MIAME has been included in the guidance on genomics data submissions released by the Food and Drug Administration (FDA) (www.fda.gov/cder/guidance/5900dft.doc). MIAME/Tox and MIAME-Env have been considered as good groundwork for tox and eco-toxicogenomics investigations and received positive comments in expert workshops in the industrial and regulatory arenas (OECD, 2004; Corvi et al., 2006).

ADDING SEMANTICS AND SYNTAX TO THE REPORTING STRUCTURE

As previously mentioned, the RSBI working group made a decision to focus on the identification and description of the “high-level concepts” of a reporting structure that encompasses any biological or technical domain of application. The discussion therefore has focused on how to “combine” the information currently formulated into (separate) technologically or biologically focused reporting structures. More difficult issues such as how to capture descriptions at a high level of granularity, and how to model and exchange the information have been deferred to two collaborative informatics projects, developing an ontology and an object model for functional genomics.

Often the description of complex investigations, such as environmental genomics, nutrigenomics and toxicogenomics, is captured in diverse formats, mostly as free text, and is commonly subject to typographical errors. The increased cost of interpreting the experimental procedures and exploring data has encouraged several scientific communities to develop and adopt ontology-based knowledge representations to extend power of their computational approaches (Blake, 2004). Incorporation of ontologies into annotations has enabled “semantic integration” of complex data, making explicit the knowledge within a certain domain. The Functional Genomics Ontology (FuGO) (Whetzel et al., *this issue*) is a large collaborative group aiming to build a common ontological framework to annotate functional genomics experiments. The project is driven by a coordinating committee that brings together the representatives of several communities, including RSBI, MGED, PSI, and MSI Ontology working groups. In a first phase the participating communities will collect descriptors for the design of the study, protocols and instrumentation used, the data generated and the types of analysis performed on the data. In a second phase these will be positioned in a common ontological framework, ensuring reuse and interaction with existing bio-ontologies (Rubin et al., *this issue*). The RSBI communities will contribute to the FuGO project in three ways; firstly, by contributing common “high-level concepts” (*Investigation, Study and Assay*); secondly, by the terms specific to environment, nutrition and toxicology domains; and thirdly, by participating in the process of agreeing and defining those terms that are shared between many domains.

The use cases and definitions released by the RSBI communities have directly contributed to the development of the Functional Genomics Experiment Object Model (FuGE) (Jones et al., *this issue*), a collaborative effort by a consortium of researchers from academia and industry. FuGE provides a model of common components in functional genomics investigations, such as materials, data, protocols, equipment and software. FuGE can be extended to develop modular data formats with consistent structure. FuGE also provides a framework for capturing complete lab workflows, in which pre-existing data formats can be integrated. FuGE can be used, in conjunction with ontologies such as FuGO, for describing any kind of investigation with complex designs (such as environmental, genomics, nutrigenomics, toxicogenomics and others in biomedical science) enabling the integration of pre-existing data formats. FuGE is being used as the basis for the next version of the microarray and proteomics standard (Ball and Brazma, *this issue*; Taylor et al., *this issue*), and it is intended that data models for other OMICS domains will use it, thereby improving facilities for data comparison across different techniques and allowing systems biology approaches within a single framework. The RSBI use cases have provided FuGE developers with real examples and terminology that bench researchers believe should be reported in a data model. These use cases have also highlighted several components that required changes in FuGE to report annotation in a format that could be queried. One example is the concept of a *Study Phase*, defined as “distinct period defined by the investigator for convenience of understanding the timeline of the study.” For more details on how FuGE represents a *Phase* and other RSBI “high-level” concepts, see (www.mged.org/Workgroups/rsbi/rsbi.html).

CONCLUSION

Since 2004, the RSBI working group has brought together environmental genomics, nutrigenomics, and toxicogenomics communities, where independent activities are underway to develop databases and establish data communication standards within their respective domains. By capitalizing on synergies, clearly substantial progress has been made in the past 2 years, and through extensive liaisons with many standards efforts, several links have been established. The RSBI working group invites other biological domains to join this endeavour (www.mged.org/Workgroups/rsbi/rsbi.html).

In the era of functional genomics and system biology data communication standards cannot be developed in isolation. It is evident that standardization initiatives need to retain their own identity because they serve the needs of a specific community. A shared interest in data communication standards may be a catalyst for synergies. These initiatives, however, include different stakeholders from all the segments of a certain domain and their different backgrounds, objectives, and scientific perspectives are a potential cause of conflicts. Despite this, it is still possible to remain cognizant of concepts shared with other domains, and synergise with relevant standardization efforts. The RSBI working group, FuGO and FuGE projects, the GSC initiatives and the others listed in this special issue of the OMICS journal provide strong evidence that such a balance can be reached. It is important to note that such endeavour requires each community to invest time, effort and funds; the latter is often a limiting factor (Brooksbank and Quackenbush, *this issue*). Standards of some depth also require a considerable amount of time to develop fully. Such collaborative investments may also be limited by standards reaching a mature and stable development stage over different time periods. The consequence of failing to deliver compatible data communication standards is a steep escalation in the burden and cost of data management tasks.

ACKNOWLEDGMENTS

We would like to thank Cath Brooksbank, Daniel Schober, Joe Wood, and Andy Jones for reading the paper critically. We gratefully acknowledge and the contributions of Alexander Garcia Castro, Karim Nashar, and Robert Stevens in the knowledge elicitation phase. We thank the environmental genomics, nutrigenomics and toxicogenomics communities for their expert contributions. We also acknowledge Mike Waters, Ben van Ommen, and Dawn Field for their support of the RSBI activities. The RSBI working groups are supported in part by the Intramural Research Program of the NIH and NIEHS (contract 273-02-C-0027), EU Network of Excellence NuGO (NoE 503630), BBSRC (grant BB/D524283/1), and EMBL.

REFERENCES

- BALL, C.A., and BRAZMA, A. (2006). MGED standards: work in progress. *OMICS (this issue)*.
- BLAKE, J. (2004). Bio-ontologies—fast and furious. *Nat Biotechnol* **22**, 773–774.
- BRAZMA, A., HINGAMP, P., QUACKENBUSH, J., et al. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* **29**, 365–371.
- BRUSKIEWICH, R., DAVENPORT, G., HAZEKAMP, T., et al. (2006). Generation Challenge Programme (GCP): standards for crop data. *OMICS (this issue)*.
- CORVI, R., AHR, H., ALBERTINI, S., et al. (2006). Validation of toxicogenomics-based test systems: ECVAM-ICCVAM/NICEATM considerations for regulatory use. *Environ Health Perspect* **114**, 420–429.
- DEUTSCH, E.W., BALL, C.A., BOVA, S.G., et al. (2006). Minimum information specification for *in situ* hybridization and immunohistochemistry experiments (MISFISHIE). *OMICS (this issue)*.
- FIEHN, O., KRISTAL, B., VAN OMMEN, B., et al. (2006). Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *OMICS (this issue)*.
- FIELD, D., MORRISON, N., STERK, P., et al. eGenomics: cataloging our complete genome collection. *OMICS (this issue)*.
- FOSTEL, J., CHOI, D., ZWICKL, C., et al. (2005). Chemical effects in biological systems—data dictionary (CEBS-DD): a compendium of terms for the capture and integration of biological study design description, conventional phenotypes, and ‘omics data. *Toxicol Sci* **88**, 585–601.
- GARCIA, A.G., ROCCA-SERRA, P., STEVENS, R., et al. (2006). The use of concept maps during knowledge elicitation in ontology development processes—the nutrigenomics use case. *BMC Bioinformatics* **7**, 267.
- GAROSI, P., DE FILIPPO, C., VAN ERK, M., et al. (2005). Defining best practice for microarray analyses in nutrigenomic studies. *Br J Nutr* **93**, 425–432.
- JONES, A.R., PIZARRO, A., SPELLMAN, P., et al. (2006). FuGE: Functional Genomics Experiment object model. *OMICS (this issue)*.
- MATTES, W.B., PETTIT, S.D., SANSONE, S.A., et al. (2004). Database development in toxicogenomics: issues and efforts. *Environ Health Perspect* **112**, 495–505.
- MORRISON, N., WOOD, A.J., HANCOCK, D., et al. (2006). Standard annotation of environmental OMICS data: application to the transcriptomics domain. *OMICS (this issue)*.
- OECD. (2004). OECD chemical testing guideline: activities to explore and evaluate regulatory application of genomic methods; toxicogenomic. The OECD/IPCS Workshop, Kyoto. Available at: www.oecd.org/document/29/0,2340,en_2649_34377_34704669_1_1_1_1,00.html.
- QUACKENBUSH, J. (2004). Data standards for “omic” science. *Nat Biotechnol* **22**, 613–614.
- SANSONE, S.A., MORRISON, N., ROCCA-SERRA, P., et al. (2004). Standardization initiatives in the (eco)toxicogenomics domain: a review. *Comp Funct Genomics* **8**, 633–641.
- SPIDLEN, J., GENTLEMAN, R.C., HAALAND, P.D., et al. (2006). Data standards for flow cytometry. *OMICS (this issue)*.

Address reprint requests to:
Dr. Susanna-Assunta Sansone
EMBL-EBI
NuGO
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, UK

E-mail: sansone@ebi.ac.uk