Data Standards for Omics Data: The Basis of Data Sharing and Reuse

Stephen A. Chervitz, **Eric W. Deutsch**, **Dawn Field**, **Helen Parkinson**, **John Quackenbush**, **Phillipe Rocca-Serra**, **Susanna-Assunta Sansone**, **Christian J. Stoeckert Jr.**, **Chris F. Taylor**, **Ronald Taylor**, and **Catherine A. Ball**

Abstract

To facilitate sharing of Omics data, many groups of scientists have been working to establish the relevant data standards. The main components of data sharing standards are experiment description standards, data exchange standards, terminology standards, and experiment execution standards. Here we provide a survey of existing and emerging standards that are intended to assist the free and open exchange of large-format data.

Keywords

Data sharing; Data exchange; Data standards; MGED; MIAME; Ontology; Data format; Microarray; Proteomics; Metabolomics

1. Introduction

The advent of genome sequencing efforts in the 1990s led to a dramatic change in the scale of biomedical experiments. With the comprehensive lists of genes and predicted gene products that resulted from genome sequences, researchers could design experiments that assayed every gene, every protein, or every predicted metabolite. When exploiting transformative Omics technologies such as microarrays, proteomics or high-throughput cell assays, a single experiment can generate very large amounts of raw data as well as summaries in the form of lists of sequences, genes, proteins, metabolites, or SNPs. Managing, analyzing, and sharing the large data set from Omics experiments present challenges because the standards and conventions developed for single-gene or single-protein studies do not accommodate the needs of Omics studies (1) (see Note 1).

The development and applications of Omics technologies is evolving rapidly, and so is awareness of the need for, and value of, data-sharing standards in the life sciences community. Standards that become widely adopted can help scientists and data analysts better utilize, share, and archive the ever-growing mountain of Omics data sets. Also, such standards are essential for the application of Omics approaches in healthcare environments. This chapter provides an introduction to the major Omics data sharing standards initiatives

[1]Tools for programmers: Many labs need to implement their own tools for managing and analyzing data locally. There are a number of parsers and tool kits for common data formats that can be reused in this context. These are listed in Table 11.

in the domains of genomics, transcriptomics, proteomics, and metabolomics, and includes summaries of goals, example applications, and references for further information. New standards and organizations for standards development may well arise in the future that will augment or supersede the ones described here. Interested readers are invited to further explore the standards described in this chapter (as well as others not mentioned) and keep up with the latest developments by visiting the website http://biostandards.info.

1.1. Goals and Motivations for Standards in the Life Sciences

Standards within a scientific domain have the potential to provide uniformity and consistency in the data generated by different researchers, organizations, and technologies. They thereby facilitate more effective reuse, integration, and mining of those data by other researchers and third-party software applications, as well as enable easier collaboration between different groups. Standards-compliant data sets have increased value for scientists who must interpret and build on earlier efforts. And, of course, software analysis tools which – of necessity – require some sort of regularized data input are very often designed to process data that conform to public data formatting standards, when such are available for the domain of interest. Standard laboratory procedures and reference materials enable the creation of guidelines, systems benchmarks, and laboratory protocols for quality assessment and cross-platform comparisons of experimental results that are needed in order to deploy a technology within research, industrial, or clinical environments. The value of standards in the life sciences for improving the utility of data from high-throughput post-genomic experiments has been widely noted for some years (2–6).

To understand how the conclusions of a study were obtained, not only do the underlying data need to be available, but also the details of how the data were generated need to be adequately described (i.e., samples, procedural methods, and data analysis). Depositing data in public repositories is necessary but not sufficient for this purpose. Several standard types of associated data are also needed. Reporting, or "minimum information," standards are needed to ensure that submitted data are sufficient for clear interpretation and querying by other scientists. Standard data formats greatly reduce the amount of effort required to share and make use of data produced by different investigators. Standards for the terminology used to describe the study and how the data were generated enable not only improved understanding of a given set of experimental results but also improved ability to compare studies produced by different scientists and organizations. Standard physical reference materials as well as standard methods for data collection and analysis can also facilitate such comparisons as well as aid the development of reusable data quality metrics.

Ideally, any standards effort would take into account the usability of the proposed standard. A standard that is not widely used is not really a standard and the successful adoption of a standard by end-user scientists requires a reasonable cost-benefit ratio. The effort of producing a new standard (development cost) and, more importantly, the effort needed to learn how to use the standard or to generate standards-conforming data (end-user cost) has to be outweighed by gains in the ability to publish experimental results, the ability to use other published results to advance one's own work, and higher visibility bestowed on standards-compliant publications (7). Thus, a major focus of standards initiatives is

minimizing end-user usability barriers, typically done by educational outreach via workshops and tutorials as well as fostering the development of software tools that help scientists utilize the standard in their investigations. There also must be a means for incorporating feedback from the target community both at the initiation of standard development and on a continuing basis so that the standard can adapt to user needs that can change over time. Dr. Brazma and colleagues (8) discuss some additional factors that contribute to the success of standards in systems biology and functional genomics.

1.2. History of Standards for Omics

The motivation for standards for Omics initially came from the parallel needs of the scientific journals, which wanted standards for data publication, and the needs of researchers, who recognized the value of comparing the large and complex data sets characteristic of Omics experiments. Such data sets, often with thousands of data points, required new data formats and publication guidelines. Scientists using DNA microarrays for genome-wide gene expression analysis were the first to respond to these needs. In 2001, the Microarray and Gene Expression Data (MGED) Society (http://www.mged.org) published the Minimum Information About a Microarray Experiment (MIAME) standard (9), a guideline for the minimum information required to describe a DNA microarray-based experiment. The MIAME guidelines specify the information required to describe such an experiment so that another researcher in the same discipline could either reproduce the experiment or analyze the data de novo.

Adoption of the MIAME guidelines was expedited when a number of journals and funding agencies required compliance with the standard as a precondition for publication. In parallel with MIAME, data modeling and XML-based exchange standards called Microarray Gene Expression Object Model (MAGE-OM) and Markup Language (MAGE-ML) (10), and a controlled vocabulary called the MGED Ontology (11), were created. These standards facilitated the creation and growth of a number of interoperable databases and public data repositories. Use of these standards also led to the establishment of open-source software projects for DNA microarray data analysis. Resources such as the ArrayExpress database (12–14) at the European Bioinformatics Institute (EBI), the Gene Expression Omnibus (GEO) (15–18), at the National Center for Biotechnology Information (NCBI), and others were advertised as "MIAME-compliant" and capable of importing data submitted in the MAGE-ML format (10).

Minimum information guidelines akin to MIAME then arose within other Omics communities. For example, the Minimum Information about a Proteomics Experiment (MIAPE) guidelines for proteomics studies (19) have been developed. More recent initiatives have been directed towards technology-independent standards for reporting, modeling, and exchange that support work spanning multiple Omics technologies or domains, and directed toward harmonization of related standards. These projects have, of necessity, required extensive collaboration across disciplines. The resulting standards have gained in sophistication, benefiting from insights gained in the use and implementation of earlier standards, in the use of formalisms imposed by the need to make the data

computationally tractable and logically coherent, and in the experience in engagement of multiple academic communities in the development of these prior standards.

Increasingly, the drive for standards in Omics is shifting from the academic communities to include the biomedical and healthcare communities as well. As application of Omics technologies and data expands into the clinical and diagnostic arena, organizations such as the US Food and Drug Administration (FDA) and technology manufacturers are becoming more involved in a range of standards efforts, for example the MicroArray Quality Control (MAQC) consortium brings together representatives of many such organizations (20). Quality control/assurance projects and reference standards that support comparability of data across different manufacturer platforms are of particular interest as Omics technologies mature and start to play an expanded role in healthcare settings.

2. Materials

Omics standards are typically scoped to a specific aspect of an Omics investigation. Generally speaking, a given standard will cover either the description of a completed experiment, or will target some aspect of performing the experiment or analyzing results. Standards are further stratified to handle more specific needs, such as reporting data for publication, providing data exchange formats, or defining standard terminologies. Such scoping reflects a pragmatic decoupling that permits different standards groups to develop complementary specifications concurrently and allows different initiatives to attract individuals with relevant expertise or interest in the target area (8).

As a result of this arrangement, a standard or standardization effort within Omics can be generally characterized by its domain and scope. The domain reflects the type of experimental data (transcriptomics, proteomics, metabolomics, etc.), while the scope defines the area of applicability of the standard or the methodology being standardized (experiment reporting, data exchange, etc.). Tables 1 and 2 list the different domains and scopes, respectively, which characterize existing Omics standardizations efforts (see Note 2).

The remainder of this section describes the different scopes of Omics standards, listing the major standards initiatives and organizations relevant to each scope. The next section then surveys the standards by domain, providing more in-depth description of the relevant standards, example applications, and references for further information.

2.1. Experiment Description Standards

Experiment description standards, also referred to generally as "data standards", concern the development of guidelines, conventions, and methodologies for representing and communicating the raw and processed data generated by experiments as well as the metadata for describing how an experiment was carried out, including a description of all reagents, specimens, samples, equipment, protocols, controls, data transformations, software

---

[2]Tips for using standards: Standards are commonly supported by tools and applications related to projects or to public repositories. One example is the ISA-TAB related infrastructure described in Subheading 3.8.3, others are provided in Table 12. These include simple conversion tools for formats used by standards compliant databases such as ArrayExpress and GEO, and tools that allow users to access these databases and load data into analysis applications.

algorithms, and any other factors needed to accurately communicate, interpret, reproduce, or analyze the experimental results.

Omics studies and the data they generate are complex. The diversity of scientific problems, experimental designs, and technology platforms creates a challenging landscape of data for any descriptive standardization effort. Even within a given domain and technology type, it is not practical for a single Specification to encompass all aspects of describing an experiment. Certain aspects are more effectively handled separately; for example, a description of the essential elements to be reported for an experiment is independent of the Specific data format in which that information should be encoded for import or export by software applications.

In recognition of this, experiment description standardization efforts within the Omics community are further scoped into more specialized areas that address distinct data handling requirements encountered during different aspects of or types of data encountered in an Omics study. Thus we have:

- Reporting.

- Data exchange & modeling.

- Terminology.

These different areas serve complementary roles and together, provide a complete package for describing an Omics experiment within a given domain or technology platform. For example, a data exchange/modeling standard will typically have elements to satisfy the needs of a reporting standard with a set of allowable values for those elements to be provided by an associated standard controlled vocabulary/terminology.

2.1.1. Reporting Standards: Minimum Information Guidelines—The scope of a reporting standard pertains to how a researcher should record the information required to unambiguously communicate experimental designs, treatments and analyses, to con-textualize the data generated, and underpin the conclusions drawn. Such standards are also known as data content or minimum information standards because they usually have an acronym beginning with "MI" standing for "minimum information" (e.g. MIAME). The motivation behind reporting standards is to enable an experiment to be interpreted by other scientists and (in principle) to be independently reproduced. Such standards provide guidance to investigators when preparing to report or publish their investigation or archive their data in a repository of experimental results. When an experiment is submitted to a journal for publication, compliance with a reporting standard can be valuable to reviewers, aiding them in their assessment of whether an experiment has been adequately described and is thus worthy of approval for publication.

A reporting Specification does not normally mandate a particular format in which to capture/ transport information, but simply delineates the data and metadata that their originating community considers appropriate to sufficiently describe how a particular investigation was carried out. Although a reporting standard does not have a Specific data formatting requirement, the often explicit expectation is that the data should be provided using a

technology-appropriate standard format where feasible, and that controlled vocabulary or ontology terms should be used in descriptions where feasible. Data repositories may impose such a requirement as a condition for data submission.

Omics experiments, in addition to their novelty, can be quite complex in their execution, analysis, and reporting. Minimum information guidelines help in this regard by providing a consistent framework to help scientists think about and report essential aspects of their experiments, with the ultimate aim of ensuring the usefulness of the results to scientists who want to understand or reproduce the study. Such guidelines also help by easing compliance with a related data exchange standard, which is often designed to support the requirements of a reporting standard (discussed below). Depending on the nature of a particular investigation, information in addition to what is specified by a reporting standard may be provided as desired by the authors of the study or as deemed necessary by reviewers of the study.

Table 3 lists the major reporting standards for different Omics domains. The MIBBI project (discussed later in this chapter) catalogues these and many other reporting standards and provides a useful introduction (21).

For some publishers, compliance with a reporting standard is increasingly becoming an important criterion for accepting or rejecting a submitted Omics manuscript (22). The journals Nature, Cell, and The Lancet have led the way in the enforcement of compliance for DNA microarray experiments by requiring submitted manuscripts to demonstrate compliance with the MIAME guidelines as a condition of publication. Currently, most journals that publish such experiments have adopted some elaboration of this policy. Furthermore, publishers such as the BioMed Central are moving to, or already endorse the MIBBI project, described below, as a portal to the diverse set of available guidelines for the biosciences.

2.1.2. Data Exchange and Modeling Standards—The scope of a data exchange standard is the definition of an encoding format for use in sharing data between researchers and organizations, and for exchanging data between software programs or information storage systems. A data exchange standard delineates what data types can be encoded and the particular way they should be encoded (e.g., tab-delimited columns, XML, binary, etc.), but does not specify what the document should contain in order to be considered complete. There is an expectation that the content will be constructed in accordance with a community-approved reporting standard and the data exchange standard itself is typically designed so that users can construct documents that are compliant with a particular reporting standard (e.g., MAGE-ML and MAGE-TAB contain placeholders that are designed to hold the data needed for the production of MIAME-compliant documents).

A data exchange standard often is designed to work in conjunction with a data modeling standard, which defines the attributes and behaviors of key entities and concepts (objects) that occur within an Omics data set. The model is intended to capture the exchange format-encoded data for the purpose of storage or downstream data mining by software applications. The data model itself is designed to be independent of any particular software

implementation (database schema, XML file, etc.) or programming language (Java, C++, Perl, etc.). The implementation decisions are thus left to the application programmer, to be made using the most appropriate technology(s) for the target user base. This separation of the model (or "platform-independent model") and the implementation (or "platform-Specific implementation") was first defined by the Object Management Group's Model Driven Architecture (http://www.omg.org/mda) and offers a design methodology that holds promise for building software systems that are more interoperable and adaptable to technological change. Such extensibility has been recognized as an essential feature of data models for Omics experiments (23). Data exchange and modeling standards are listed in Table 4.

2.1.3. Terminology Standards—The scope of a terminology standard is typically defined by the use cases it is intended to support and competency questions it is designed to answer. An example of a use case is annotating the data generated in an investigation with regard to materials, procedures, and results while associated competency questions would include those used in data mining (for example, "find all cancer studies done using Affymetrix microarrays"). Terminology standards generally provide controlled vocabularies and some degree of organization. Ontologies have become popular as mechanisms to encode terminology standards because they provide definitions for terms in the controlled vocabulary as well as properties of and relationships between terms. The Gene Ontology (24) is one such ontology created to address the use case of providing consistent annotation of gene products across different species and enabling questions such as "return all kinases". The primary goal of a terminology standard is to promote consistent use of terms within a community and thereby facilitate knowledge integration by enabling better querying and data mining within and across data repositories as well as across domain areas.

Use of standard terminologies by scientists working in different Omics domains can enable interrelation of experimental results from diverse data sets (see Note 3). For example, annotating results with standard terminologies could help correlate the expression profile of a particular gene, assayed in a transcriptomics experiment, to its protein modification state, assayed in a separate proteomics experiment. Using a suitably annotated metabolomics experiment, the gene/protein results could then be linked to the activity of the pathway(s) in which they operate, or to a disease state documented in a patient's sample record.

Consistent use of a standard terminology such as GO has enabled research advances. Data integration is possible across diverse data sets as long as they are annotated using terms from GO. Data analysis for association of particular types of gene products with results from investigations is also made possible because of the effort that has been made by the GO Consortium to consistently annotate gene products with GO. Numerous tools that do this are listed at the Gene Ontology site http://www.geneontology.org/GO.tools.microarray.shtml.

There is already quite a proliferation of terminologies in the life sciences. Key to their success is adoption by scientists, bio-informaticians, and software developers for use in the annotation of Omics data. However, the proliferation of ontologies which are not

---

[3] Annotation tools for biologists and bioinformaticians: Annotation of data to be compliant with standards is supported by several open-source annotation tools. Some of these are related to repositories supporting standards, but most are available for local installation as well. These are described in Table 13.

interoperable can be a barrier to integration (25) (see Note 4). The OBO Foundry targets this area and is delineating best practices underlying the construction of terminologies, maximizing their internal integrity, extensibility, and reuse. Easy access to standard terminologies is important and being addressed through sites such as the NCBO BioPortal (http://bioportal.bioontology.org) and the EBI Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup). These web sites allow browsing and downloading of ontologies. They also provide programmatic access through web services, which is important for integration with software tools and web sites that want to make use of these.

Terms in ontologies are organized into classes and typically placed in a hierarchy. Classes represent types of entities for which there can be different instances. Terms can be given accession numbers so that they can be tracked and can be assigned details, such as who is responsible for the term and what was the source of the definition. If the ontology is based on a knowledge representation language such as OWL (web ontology language, http://www.w3.org/TR/owl-ref) then restrictions on the usage of the term can be encoded. For example, one can require associations between terms (e.g. the inputs and outputs of a process). Building an ontology is usually done with a tool such as Protégé (http://protege.stanford.edu) or OBO-Edit (http://oboedit.org). These tools are also useful for navigating ontologies.

Table 5 lists some of the ontologies or controlled vocabularies relevant to Omics. For a complete listing and description of these and related ontologies, see the OBO Foundry website (http://www.obofoundry.org).

2.2. Experiment Execution Standards

2.2.1. Physical Standards—The scope of a physical standard pertains to the development of standard reagents for use as spike-in controls in assays. A physical standard serves as a stable reference point that can facilitate the quantification of experimental results and the comparison of results between different runs, investigators, organizations, or technology platforms. Physical standards are essential for quality metrics purposes and are especially important within applications of Omics technologies in regulated environments such as clinical or diagnostic settings.

In the early days of DNA microarray-based gene expression experiments, results from different investigators, laboratories, or array technology were notoriously hard to compare despite the use of reporting and data exchange standards (26). The advent of physical standards and the improved metrology promises to increase the accuracy of comparisons within cross-platform and cross-investigator experimental results. Such improvements are necessary for the adoption of Omics technologies in clinical and diagnostic applications within the regulated healthcare industry. Examples of physical standards are provided in Table 6.

---

[4]Tips for using Ontologies: Further introductory information on design and use of ontologies can be found at the Ontogenesis site (http://ontogenesis.knowledgeblog.org). Publicly available ontologies can be queried from the NCBO's website (http://www.bioportal.org) and tutorials for developing ontologies and using supporting tools such as the OWL-API are run by several organizations, including the NCBO, the OBO Foundry and the University of Manchester, UK.

2.2.2. Data Analysis and Quality Metrics—The scope of a data analysis or quality metrics standard is the delineation of best practices for algorithmic and statistical approaches to processing experimental results as well as methods to assess and assure data quality. Methodologies for data analysis cover the following areas:

- Data transformation (normalization) protocols.

- Background or noise correction.

- Clustering.

- Hypothesis testing.

- Statistical data modeling.

Analysis procedures have historically been developed in a tool-Specific manner by commercial vendors, and users of these tools would rely on the manufacturer for guidance. Yet efforts to define more general guidelines and protocols for data analysis best practices are emerging. Driving some of these efforts is the need for consistent approaches to measure data quality, which is critical for determining one's confidence in the results from any given experiment and for judging the comparability of results obtained under different conditions (days, laboratories, equipment operators, manufacturing batches, etc.). Data quality metrics rely on data analysis standards as well as the application of physical standards. Collecting or assessing data quality using quality metrics is facilitated by having data conforming to widely-adopted reporting standards as available in common data exchange formats. A number of data analysis and quality metrics efforts are listed in Table 7.

3. Methods

Here we review some of the more prominent standards and initiatives within the main Omics domains: genomics, transcriptomics, proteomics, and metabolomics. Of these, transcriptomics is the most mature in terms of standards development and community adoption, though proteomics is a close second.

3.1. Genomic Standards

Genomic sequence data is used in a variety of applications such as genome assembly, comparative genomics, DNA variation assessment (SNP genotype and copy number), epigenomics (DNA methylation analysis), and metagenomics (DNA sequencing of environment samples for organism identification). Continued progress in the development of high-throughput sequencing technology has lead to an explosion of new genome sequence data and new applications of this technology. A number of efforts are underway to standardize the way scientists describe and exchange this genomic data in order to facilitate better exchange and integration of data contributed by different laboratories using different sequencing technologies.

3.1.1. MIGS-MIMS—This term stands for Minimum Information About a Genome Sequence/Minimum Information about a Metagenomic Sequence: MIGS/MIMS (http://gensc.org).

MIGS (Minimum Information About a Genome Sequence) is a minimum information checklist that is aimed at standardizing the description of a genomic sequence, such as the complete assembly of a bacterial or eukaryotic genome. It is intended to extend the core information that has been traditionally captured by the major nucleotide sequence repositories (Genbank, EMBL, and DDBJ) in order to accommodate the additional requirements of scientists working with genome sequencing project data.

MIGS is maintained by the Genomic Standards Consortium (http://gensc.org) which also has developed an extension of MIGS for supporting metagenomic data sets called MIMS (Minimum Information about a Metagenomic Sequence/Sample). MIMS allows for additional metadata particular to a metagenomics experiment, such as the details about environmental sampling.

A data format called GCDML (Genomic Contextual Data Markup Language) is under development by the GSC for the purpose of providing a MIGS/MIMS-compliant data format for exchanging data from genomic/metagenomic experiments.

3.1.2. SAM Tools—The SAM format is an emerging data exchange format for efficiently representing large sequence alignments, driven by the explosion of data from high-throughput sequencing projects, such as the 1,000 Genomes Project (27). It is designed to be simple, compact, and to accommodate data from different alignment programs. The SAM Tools open source project provides utilities for manipulating alignments in the SAM format, including sorting, merging, indexing, and generating alignments (http://samtools.sourceforge.net).

3.1.3. PML and PaGE-OM—The Polymorphism Markup Language PML (http://www.openpml.org) was approved as an XML-based data format for exchange of genetic polymorphism data (e.g., SNPs) in June 2005. It was designed to facilitate data exchange among different data repositories and researchers who produce or consume this type of data. Phenotype and Genotype Experiment Object Model (PaGE-OM) is an updated, broader version of the PML standard and provides a richer object model and incorporates phenotypic information. It was approved as a standard by the OMG in March 2008. PaGE-OM defines a generic, platform-independent representation for entities such as alleles, genotypes, phenotype values, and relationships between these entities with the goal of enabling the capture of the minimum amount of information required to properly report most genetic experiments involving genotype and/or phenotype information (28). Further refinements of the PaGE-OM object model, harmonization with object models from other domains, and generation of exchange formats are underway at the time of writing. PaGE-OM is maintained by JBIC (http://www.pageom.org) in partnership with the Gen2Phen project (http://www.gen2phen.org).

3.2. Transcriptomics Standards

This section describes the organizations and standards related to technologies that measure transcription, gene expression, or its regulation on a genomic scale.

Transcriptomics standards pertain to the following technologies or types of investigation:

- Gene expression via DNA microarrays or ultra high-throughput sequencing.

- Tiling.

- Promoter binding (ChIP-chip, ChIP-seq).

- In situ hybridization studies of gene expression.

3.2.1. MIAME—The goal of MIAME (Minimum Information About a Microarray Experiment, http://www.mged.org/Workgroups/MIAME/miame.html) is to permit the unambiguous interpretation, reproduction, and verification of the results of a microarray experiment. MIAME was the original reporting standard which inspired similar "minimum information" requirements Specifications in other Omics domains (9).

MIAME defines the following six elements as essential for achieving these goals:

1. The raw data from each hybridization.

2. The final processed data for the set of hybridizations in the experiment.

3. The essential sample annotation, including experimental factors and their values.

4. The experiment design including sample data relationships.

5. Sufficient annotation of the array design.

6. Essential experimental and data processing protocols.

For example, the MIAME standard has proven useful for microarray data repositories that have used it both as a guideline to data submitters and as a basis for judging the completeness of data submissions. The ArrayExpress database provides a service to publishers of microarray studies wherein ArrayExpress curators will assess a dataset on the basis of how well it satisfies the MIAME requirements (29). A publisher can then choose whether to accept or reject a manuscript on the basis of the assessment.

ArrayExpress judges the following aspects of a report to be the most critical toward MIAME compliance:

- Sufficient information about the array design (e.g., reporter sequences for oligonucleotide arrays or database accession numbers for cDNA arrays).

- Raw data as obtained from the image analysis software (e.g. CEL files for Affymetrix technology, or GPR files for GenPix).

- Processed data for the set of hybridizations.

- Essential sample annotation, including experimental factors (variables) and their values (e.g., the compound and dose in a dose response experiment).

- Essential experimental and data processing protocols.

- Several publishers now have policies in place that require MIAME-compliance as a precondition for publication.

3.2.2. MINSEQE—The Minimum Information about a high-throughput Nucleotide SEQuencing Experiment (MINSEQE, http://www.mged.org/minseqe) provides a reporting guideline akin to MIAME that is applicable to high-throughput nucleotide sequencing experiments used to assay biological state. It does not pertain to traditional sequencing projects, where the aim is to assemble a chromosomal sequence or resequence a given genomic region, but rather to applications of sequencing in areas such as transcriptomics where high-throughput sequencing is being used to compare the populations of sequences between samples derived from different biological states, for example, sequencing cDNAs to assess differential gene expression.

Here, sequencing provides a means to assay the sequence composition of different biological samples, analogous to the way that DNA microarrays have traditionally been used. MINSEQE is now supported by both the Gene Expression Omnibus (GEO) and ArrayExpress. ArrayExpress and GEO have entered into a metadata exchange agreement, meaning that UHTS sequence experiments will appear in both databases regardless of where they were submitted. This complements the exchange of underlying raw data between the NBCI and EBI short read archives, SRA and ERA.

3.2.3. MAGE—The MAGE project (MicroArray Gene Expression, http://www.mged.org/Workgroups/MAGE/mage.html) aims to provide a standard for the representation of microarray gene expression data to facilitate the creation of software tools for exchanging microarray information between different users and data repositories. The MAGE family of standards does not have direct support for capturing the results of higher-level analysis (e.g., clustering of expression data from a microarray experiment).

MAGE includes the following sub-projects:

- MAGE-OM: MAGE Object Model

- MAGE-ML: MAGE Markup Language

- MAGEstk: MAGE Software Toolkit

- MAGE-TAB: MAGE Tabular Format

MAGE-OM is a platform independent model for representing gene expression microarray data. Using the MAGE-OM model, the MGED Society has implemented MAGE-ML (an XML-based format) as well as MAGE-TAB (tab-delimited values format). Both formats can be used for annotating and communicating data from microarray gene expression experiments in a MIAME-compliant fashion. MAGE-TAB evolved out of a need to create a simpler version of MAGE-ML. MAGE-TAB is easier to use and thus more accessible to a wider cross-section of the microarray-based gene expression community which has struggled with the often large, structured XML-based MAGE-ML documents. A limitation of MAGE-TAB is that only single values are permitted for certain data slots that may in practice be multivalued. Data that cannot be adequately represented by MAGE-TAB can be described using MAGE-ML, which is quite flexible.

MAGEstk is a collection of Open Source packages that implement the MAGE Object Model in various programming languages (10). The toolkit is meant for bioinformatics users that

develop their own applications and need to integrate functionality for managing an instance of a MAGE-OM. The toolkit facilitates easy reading and writing of MAGE-ML to and from the MAGE-OM, and all MAGE-objects have methods to maintain and update the MAGE-OM at all levels. However, the MAGE-stk is the glue between a software application and the standard way of representing DNA microarray data in MAGE-OM as a MAGE-ML file.

3.2.4. MAGE-TAB—MAGE-TAB (30) is a simple tab delimited format that is used to represent gene expression and other high throughput data such as high throughput sequencing (see Note 5). It is the main submission format for the ArrayExpress database at the European Bioinformatics Institute and is supported by the BioConductor package ArrayExpress. There are also converters available to MAGE-TAB from GEO soft format, from MAGE-ML to MAGE-TAB, and an open source template generation system (31). The MGED community maintains a complete list of applications using MAGE-TAB at http://www.mged.org/mage-tab (see Note 6).

3.2.5. MO—The MGED Ontology (MO, http://mged.sourceforge.net/ontologies/index.php) provides standard terms for describing the different components of a microarray experiment (11). MO is complementary to the other MGED standards, MIAME, and MAGE, which respectively specify what information should be provided and how that information should be structured. The Specification of the terminology used for labeling that information has been left to MO. MO is an ontology with defined classes, instances, and relations. A primary motivation for the creation of MO was to provide terms wherever needed in the MAGE Object Model. This has led to MO being organized along the same lines as the MAGE-OM packages. A feature of MO is that it provides pointers to other resources as appropriate to describe a sample, or biomaterial characteristics, and treatment compounds used in the experiment (e.g. NCBI Taxonomy, ChEBI) rather than importing, mapping, or duplicating those terms.

A major revision of MO (currently at version 1.3.1.1, released in Feb. 2007) was planned to address structural issues. However, such plans have been recently superseded by efforts aimed in incorporating MO into the Ontology for Biomedical Investigations (OBI).

The primary usage of MO has been for the annotation of microarray experiments. MO terms can be found incorporated into a number of microarray databases (e.g., ArrayExpress, RNA Abundance Database (RAD) (32), caArray (http://caarray.nci.nih.gov/). Stanford Microarray Database (SMD) (33–38), maxD (http://www.bioinf.manchester.ac.uk/microarray/maxd), MiMiR (39)) enable retrieval of studies consistently across these different sites. MO terms

---

[5]Format Conversion Tools: The MAGE-ML format described in Subheading 3.2.4 has been superseded by MAGE-TAB and the different gene expression databases use different formats to express the same standards compliant data. There are therefore a number of open source conversion tools that reformat data, or preprocess data for analysis application access. These are provided as downloadable applications, and are summarized in Table 12. Support for understanding and applying data formats is often available from repositories that use these formats for data submission and exchange. Validation tools and supporting code may also be available. Email their respective helpdesks for support.

[6]Tips for developing standards: Most standards bodies have affiliated academic or industry groups and fora who are developing applications and who welcome input from the community. For example MGED has mailing lists, workshops, and an open source project that provides tools for common data representation tasks.

have also been used as part of column headers for MAGE-TAB (30), a tab-delimited form of MAGE.

Example terms from MO v.1.3.1.1:

- BioMaterialPackage (MO_182): Description of the source of the nucleic acid used to generate labeled material for the microarray experiment (an abstract class taken from MAGE to organize MO).

- BioMaterialCharacteristics (MO_5): Properties of the biomaterial before treatment in any manner for the purposes of the experiment (a subclass of BioMaterialPackage).

- CellTypeDatabase (MO_141): Database of cell type information (a subclass of the Database).

- eVOC (MO_684): Ontology of human terms that describe the sample source of human cDNA and SAGE libraries (an instance of CellTypeDatabase).

3.2.6. MAQC—The MAQC project (MicroArray Quality Control project, http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc) aims to develop best practices for executing microarray experiments and analyzing results in a manner that maximizes consistency between different vendor platforms. The effort is spearheaded by the U.S. Food and Drug Administration (FDA) and has participants spanning the microarray industry. The work of the MAQC project is providing guidance for the development of quality measures and procedures that will facilitate the reliable use of microarray technology within clinical practice and regulatory decision-making, thereby helping realize the promises of personalized medicine (40).

The project consists of two phases:

1. MAQC-I demonstrated the technical performance of microarray platforms in the identification of differentially expressed genes (20).

2. MAQC-II is aimed at reaching consensus on best practices for developing and validating predictive models based on microarray data. This phase of the project includes genotyping data as well as gene expression data, which was the focus of MAQC-I. MAQC-II is currently in progress with results expected soon (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc).

3.2.7. ERCC—The External RNA Control Consortium (ERCC, http://www.cstl.nist.gov/biotech/Cell&TissueMeasurements/GeneExpression/ERCC.htm) aims to create well-characterized and tested RNA spike-in controls for gene expression assays. They have worked with the U.S. National Institute of Standards and Technology (NIST) to create certified reference materials useful for evaluating sample and system performance. Such materials facilitate standardized data comparisons among commercial and custom microarray gene expression platforms as well as by an alternative expression profiling method such as qRT-PCR.

The ERCC originated in 2003 and has grown to include more than 90 organizations spanning a cross-section of industry and academic groups from around the world. The controls developed by this group have been based on contributions from member organizations and have undergone rigorous evaluation to ensure efficacy across different expression platforms.

3.3. Proteomic Standards

This section describes the standards and organizations related to technologies that measure protein-related phenomena on a genomic scale.

3.3.1. HUPO PSI—The primary digital communications standards organization in this domain is the Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) (http://www.psidev.info/), which provides an official process for drafting, reviewing, and accepting proteomics-related standards (41). As with other standardization efforts, the PSI creates and promotes both minimum information standards, which define what metadata about a study should be provided, as well as data exchange standards, which define the standardized, computer-readable format for conveying the information.

Within the PSI are six working groups, which define standards in subdomains representing different components in typical workflows or different types of investigations:

- Sample processing

- Gel electrophoresis

- Mass spectrometry

- Proteomics informatics

- Protein modifications

- Molecular interactions

3.3.2. MIAPE—MIAPE (Minimum Information About a Proteomics Experiment, http://www.psidev.info/index.php?q=node/91) is a reporting standard for proteomics experiments analogous to use of MIAME for gene expression experiments. The main MIAPE publication (19) describes the overall goals and organization of the MIAPE Specifications. Each subdomain (e.g., sample processing, column chromatography, mass spectrometry, etc.) has been given a separate MIAPE module that describes the information needed for each component of the study being presented. The PSI has actively engaged the journal editors to refine the MIAPE modules to a level that the editors are willing to enforce.

3.3.3. Proteomics Mass Spectrometry Data Exchange Formats—Since 2003, several data formats for encoding data related to proteomics mass spectrometry experiments have emerged. Some early XML-based formats originating from the Institute for Systems Biology such as mzXML (42) and pepXML/protXML (43) were widely adopted and became de-facto standards. More recently, the PSI has built on these formats to develop official standards such as mzML (44) for mass spectrometer output, GelML for gel electrophoresis, and mzIdentML for the bioinformatics analysis results from such data and

others. See Deutsch et al. (45) for a review of some of these formats. These newer PSI formats are accompanied by controlled vocabularies, semantic validator software, example instance documents, and in some cases fully open-source software libraries to enable swift adoption of these standards.

3.3.4. MIMIx—The PSI Molecular Interactions (MI) Working Group (http://www.psidev.info/index.php?q=node/277) has developed and approved several standards to facilitate sharing of molecular interaction information. MIMIx (Minimum Information about a Molecular Interaction Experiment) (46) is the minimum information standard that defines what information must be present in a compliant list of molecular interactions. The PSI-MI XML (or MIF) standard is an XML-based data exchange format for encoding the results of molecular interaction experiments. A major component of the format is a controlled vocabulary (PSI-MI CV) that insures the terms to describe and annotate interactions are used consistently by all documents and software. In addition to the XML format, a simpler tab-delimited data exchange format MITAB2.5 has been developed. It supports a subset of the PSI-MI XML functionality and can be edited easily using widely available spreadsheet software (47).

3.4. Metabolomics Standards

This section describes the standards and organizations related to the study of metabolomics, which studies low molecular weight metabolite profiles on a comprehensive, genomic scale within a biological sample. Metabolomic standards initiatives are not as mature as those in the transcriptomic and proteomic domains, though there is a growing community interest in this area. (Note that no distinction is made in this text between metabolomics vs metabonomics. We use "metabolomics" to refer to both types of investigations, in so far as a distinction exists).

Metabolomic standards pertain to the following technologies or types of investigations:

- Metabolic profiling of all compounds in a Specific pathway

- Biochemical network modeling

- Biochemical network perturbation analysis (environmental, genetic)

- Network flux analysis

The metabolomics research community is engaged in development of a variety of standards, coordinated by the Metabolomics Standards Initiative (48, 49). Under development are reporting "minimum information" standards (48, 50), data exchange formats (51), data models (52–54), and standard ontologies (55). A number of Specific experiment description-related projects for metabolomics are described below.

3.4.1. CIMR—CIMR (Core Information for Metabolomics Reporting, http://msi-workgroups.sourceforge.net) is in development as a minimal information guideline for reporting metabolomics experiments. It is expected to cover all metabolomics application areas and analysis technologies. The MSI is also involved in collaborative efforts to develop ontologies and data exchange formats for metabolomics experiments.

3.4.2. MeMo and ArMet—MeMo (Metabolic Modelling, http://dbkgroup.org/memo) defines a data model and XML-based data exchange format for metabolomic studies in yeast (54).

ArMet (Architecture for Metabolomics, http://www.armet.org) defines a data model for plant metabolomics experiments and also provides guidance for data collection (52, 56).

3.5. Healthcare Standards

The health care community has a long history of using standards to drive data exchange and submission to regulatory agencies. Within this setting, it is vital to ensure that data from assays pass quality assessments and can be transferred without loss of meaning and in a format that can be easily used by common tools. The drive to translate Omics approaches from a research to a clinical setting has provided strong motivation for the development of physical standards and guidelines for their use in this setting. Omics technologies hold much promise to improve our understanding of the molecular basis of diseases and develop improved diagnostics and therapeutics tailored to individual patients (6, 57).

Looking forward, the health care community is now engaged in numerous efforts to define standards important for clinical, diagnostic, and toxicological applications of data from high-throughput genomics technologies. The types and amount of data from a clinical trial or toxicogenomics study are quite extensive, incorporating data from multiple Omics domains. Standards development for electronic submission of this data is still ongoing with best practices yet to emerge. While it is likely that high-throughput data will be summarized prior to transmission, it is anticipated that the raw files should be available for analysis if requested by regulators and other scientists.

Standards-related activities pertaining to the use of Omics technologies within a health care setting can be roughly divided into three main focus areas: experiment description standards, reference materials, and laboratory procedures.

3.5.1. Healthcare Experiment Description Standards—Orthogonal to the experiment description standards efforts in the basic research and technical communities, clinicians and biologists have identified the need to describe the characteristics of an organism or specimen under study in a way that is understandable to clinicians as well as scientists. Under development within these biomedical communities are reporting standards to codify what data should be captured and in what data exchange format to permit reuse of the data by others. As with the other minimum information standards, the goal is to create a common way to describe characteristics of the objects of a study, and identify the essential characteristics to include when publishing the study. Parallel work is underway in the arena of toxicogenomics (21, 58). Additionally, standard terminologies in the form of thesauri or controlled vocabularies and systematic annotation methods are also under development.

It is envisioned that clinically relevant standards (some of which are summarized in Table 8) will be used in conjunction with the experiment description standards being developed by the basic research communities that study the same biological objects and organisms. For example, ISA-TAB (described below) is intended to complement existing biomedical

formats such as the Study Data Tabulation Model (SDTM), a FDA-endorsed data model created by CDISC to organize, structure, and format both clinical and nonclinical (toxicological) data submissions to regulatory authorities (http://www.cdisc.org/models/sds/v3.1/index.html). It is inevitable that some information will be duplicated between the two frameworks, but this is not generally seen as a major problem. Links between related components of ISA-TAB and SDTM could be created using properties of the subject source, for example.

3.5.2. Reference Materials—Developing industry-respected standard reference materials, such as a reagent for use as a positive or negative control in an assay, is essential for any work in a clinical or diagnostic setting. Reference materials are physical standards (see above) that provide an objective way to evaluate the performance of laboratory equipment, protocols, and sample integrity. The lack of suitable reference materials and guidelines for their use has been a major factor in slowing the adoption of Omics technologies such as DNA microarrays within clinical and diagnostic settings (6).

The ERCC (described above) and the LGC (http://www.lgc.co.uk) are the key organizations working on development of standard reference materials, currently targeting transcriptomics experiments.

3.5.3. Laboratory Procedures—Standard protocols providing guidance in the application of reference materials, experiment design, and data analysis best practices are essential for performing high-throughput Omics procedures in clinical or diagnostic applications.

The Clinical Laboratory Standards Institute (CLSI, http://www.clsi.org) is an organization that provides an infrastructure for ratifying and publishing guidelines for clinical laboratories. Working with organizations such as the ERCC (described above), they have produced a number of documents (see Table 9) applicable to the use of multiplex, whole-genome technologies such as gene expression and genotyping within a clinical or diagnostic setting.

3.6. Challenges for Omics Standards in Basic Research

A major challenge facing Omics standards is proving their value to a significant fraction of the user base and facilitating widespread adoption. Given the relative youth of the field of Omics and of the standardization efforts for such work, the main selling point for use of a standard has been that it will benefit future scientists and application/database developers, with limited added value for the users who are being asked to comply with the standard at publication time. Regardless of how well designed a standard is, if complying with it is perceived as being difficult or complicated, widespread adoption will be unlikely to occur. Some degree of enforcement of compliance by publishers and data repositories most likely will be required to inculcate the standard and build a critical mass within the targeted scientific community that then sustains its adoption. Significant progress has been achieved here: for DNA microarray gene expression studies, for example, most journals now require MIAME compliance and there is a broad recognition of the value of this standard within the target community.

Here are the some of the "pressure points" any standard will experience from its community of intended users:

- Domain experts who want to ensure comprehensiveness of the standard

- End-user scientists who want the standard to be easy with which to comply

- Software developers who want tools for encoding and decoding standards-compliant data

- Standards architects who want to ensure formal correctness of the standard

Satisfying all of these interests is not an easy task. One complication is that the various interested groups may not be equally involved in the development of the standard. Balancing these different priorities and groups is the task of the group responsible for maintaining a standard. This is an ongoing process that must remain responsive to user feedback. The MAGE-TAB data exchange format in the DNA microarray community provides a case in point: it was created largely in response to users that found MAGE-ML difficult to work with.

3.7. Challenges for Omics Standards in Healthcare Settings

The handling of clinical data adds additional challenges on top of the intrinsic complexities of Omics data. Investigators must respect certain regulations imposed by regulatory authorities. For example, the Health Insurance Portability Accountability Act (HIPAA) mandates the de-identification of patient data to protect an individual's privacy. Standards and information systems used by the healthcare community therefore must be formulated to deal with such regulations (e.g., (59)). While the use of open standards poses risks to the release of protected health information, the removal of detailed patient metadata about samples can present barriers to research (60, 61). Enabling effective research while maintaining patient privacy remains an on-going issue (Joe White, Dana-Farber Cancer Institute, personal communication).

3.8. Upcoming Trends: Standards Harmonization

The field of Omics is not suffering from lack of interest in standards development, as the number of different standards discussed in this chapter attests. Such a complex landscape can have adverse effects on data sharing, integration, and systems interoperability – the very things that the standards are intended to help (62). To address this, there are a number of projects in the research and biomedical communities engaged in harmonization activities that focus on integrating standards with related or complementary scope and aim to enhance interoperability in the reporting and analysis of data generated by different technologies or within different Omics domains.

Some standards facilitate harmonization by having a sufficiently general-purpose design such that they can accommodate data from experiments in different domains. Such "multiomics" standards typically have a mechanism that allows them to be extended as needed in order to incorporate aspects Specific to a particular application area. The use of these domain- and technology-neutral frameworks is anticipated to improve the interoperability of data analysis tools that need to handle data from different types of Omics

experiments as well as to reduce wheel reinvention by different standards groups with similar needs. Harmonization and multiomics projects are collaborative efforts, involving participants from different domain-Specific standards developing organizations with shared interests. Indeed, the success of these efforts depends on continued broad-based community involvement. In Table 10, we describe major current efforts in such multiomics and harmonization.

3.8.1. FuGE—The FuGE (Functional Genomics Experiment, (http://fuge.sourceforge.net)) project aims to build generic components that capture common facets of different Omics domains (63). Its contributors come from different standards efforts, primarily MGED and HUPO-PSI, reflecting the desire to build components that provide properties and functionalities common across different Omics technologies and application areas.

The vision of this effort is that using FuGE-based components, a software developer will be better able to create and modify tools for handling Omics data, without having to reinvent the wheel for common tasks in potentially incompatible ways. Further, tools based on such shared components are expected to be more interoperable.

FuGE has several sub-projects that include the FuGE Object Model (FuGE-OM) and the FuGE Markup Language (FuGE-ML), a data exchange format. Technology-Specific aspects can be added by extending the generic FuGE components, building on the common functionalities. For example, a microarray-Specific framework equivalent to MAGE could be derived by extending FuGE, deriving microarray-Specific objects from the FuGE object model.

3.8.2. HITSP—The Healthcare Information Technology Standards Panel (HITSP) is a public-private sector partnership of standards developers, healthcare providers, government representatives, consumers, and vendors in the healthcare industry. It is administered by the American National Standards Institute (ANSI, http://www.ansi.org) to harmonize healthcare-related standards and improve interoperability of healthcare software systems. It produces recommendations and reports contributing to the development of a Nationwide Health Information Network for the United States (NHIN, http://www.hhs.gov/healthit/healthnetwork/background).

The HITSP is driven by use cases issued by the American Health Information Community (AHIC, http://www.hhs.gov/healthit/community/background). A number of use cases have been defined on a range of topics, such as personalized healthcare, newborn screening, and consumer adverse event reporting (http://www.hhs.gov/healthit/usecases).

3.8.3. ISA-TAB—The ISA-TAB format (Investigation Study Assay Tabular format, http://isatab.sourceforge.net) is a general purpose framework with which to communicate both data and metadata from experiments involving a combination of functional technologies (64). ISA-TAB therefore has a broader applicability and more extended structure compared to a domain-Specific data exchange format such as MAGE-TAB. An example where ISA-TAB might be applied would be an experiment looking at changes both in (1) the metabolite profile of urine, and (2) gene expression in the liver in subjects treated with a compound

inducing liver damage, using both mass spectrometry and DNA microarray technologies, respectively.

The ISA-Tab format is the backbone for the ISA Infrastructure – a set of tools that support the capture of multiomics experiment descriptions. It also serves as a submission format to compatible databases such as the BioInvestigation Index project at the EBI (http://www.ebi.ac.uk/bioinvindex). It allows users to create a common structured representation of the metadata required to interpret an experiment for the purpose of combined submission to experimental data repositories such as ArrayExpress, PRIDE, and an upcoming metabolomics repository (64). Additional motivation comes from a group of collaborative systems, part of the MGED's RSBI group (65), each of which is committed to pipelining Omics-based experimental data into EBI public repositories or willing to exchange data among themselves, or to enable their users to import data from public repositories into their local systems.

ISA-TAB has a number of additional features that make it a more general framework that can comfortably accommodate multidomain experimental designs. ISA-TAB builds on the MAGE-TAB paradigm, and shares its motivation for the use of tab-delimited text files; i.e., that they can easily be created, viewed, and edited by researchers, using spreadsheet software such as Microsoft Excel. ISA-TAB also employs MAGE-TAB syntax as far as possible, to ensure backward compatibility with existing MAGE-TAB files. It was also important to align the concepts in ISA-TAB with some of the objects in the FuGE model. The ISA-TAB format could be seen as competing with XML-based formats such as the FuGE-ML. However, ISA-TAB addresses the immediate need for a framework to communicate for multiomics experiments, whereas all existing FuGE-based modules are still under development. When these become available, ISA-TAB could continue serving those with minimal bioinformatics support, as well as finding utility as a user-friendly presentation layer for XML-based formats (via an XSL transformation); i.e. in the manner of the HTML rendering of MAGE-ML documents.

Initial work has been carried out to evaluate the feasibility of rendering FuGE-ML files (and FuGE-based extensions, such as GelML and Flow-ML) in the ISA-TAB format. Examples are available at the ISA-TAB website under the document section, along with a report detailing the issues faced during these transformations. When finalized, the XSL templates will also be released, along with Xpath expressions and a table mapping FuGE objects and ISA-TAB labels. Additional ISA-TAB-formatted examples are available, including a MIGS-MIMS-compliant data-set (see http://isatab.sourceforge.net/examples.html).

The decision on how to regulate the use of the ISA-TAB (marking certain fields as mandatory or enforcing the use of controlled terminology) is a matter for those who will implement the format in their system. Although certain fields would benefit from the use of controlled terminology, ISA-TAB files with all fields left empty are syntactically valid, as are those where all fields are filled with free text values rather than controlled vocabulary or ontology terms.

3.8.4. MIBBI—Experiments in different Omics domains typically share some reporting requirements (for example, specifying the source of a biological specimen). The MIBBI project (Minimal Information for Biological and Biomedical Investigations, http://mibbi.org; developers: http://mibbi.sourceforge.net) aims to work collab-oratively with different groups to harmonize and modularize their minimum information checklists (e.g., MIAME, MIGS-MIMS, etc.) refactoring the common requirements, to make it possible to use these checklists in combination (21). Additionally, the MIBBI project provides a comprehensive web portal providing registration of and access to different minimum information checklists for different types of Omics (and other) experiments.

3.8.5. OBI—An excellent description of the OBI project comes from its home web page: The Ontology for Biomedical Investigations (OBI, http://purl.obofoundry.org/obo/obi) project is developing an integrated ontology for the description of biological and medical experiments and investigations. This includes a set of "universal" terms that are applicable across various biological and technological domains, and domain-Specific terms relevant only to a given domain. This ontology will support the consistent annotation of biomedical investigations, regardless of the particular field of study. The ontology will model the design of an investigation, the protocols and instrumentation used, the material used, the data generated and the type of analysis performed on it. This project was formerly called the Functional Genomics Investigation Ontology (FuGO) project (66).

OBI is a collaborative effort of many communities representing particular research domains and technological platforms (http://obi-ontology.org/page/Consortium). OBI is meant to serve very practical needs rather than be an academic exercise. Thus it is very much driven by use cases and validation questions. The OBI user community provides valuable feedback about the utility of OBI and acts as a source of terms and use cases. As a member of the OBO Foundry (described below), OBI has made a commitment to be interoperable with other biomedical ontologies. Each term in OBI has a set of annotation properties, some of which are mandatory (minimal metadata defined at http://obi-ontology.org/page/OBI_Minimal_metadata). These include the term's preferred name, definition source, editor, and curation status.

3.8.6. OBO Consortium and the NCBO—The OBO Consortium (Open Biomedical Ontologies Consortium, http://www.obofoundry.org), a voluntary, collaborative effort among different OBO developers, has developed the OBO Foundry as a way to avoid the proliferation of incompatible ontologies in the biomedical domain (25). The OBO Foundry provides validation and assessment of ontologies to ensure interoperability. It also defines principles and best practices for ontology construction such as the Basic Formal Ontology, which serves as a root-level ontology from which other domain-Specific ontologies can be built, and the relations ontology, which defines a common set of relationship types (67). Incorporation of such elements within OBO is intended to facilitate interoperability between ontologies (i.e., for one OBO Foundry ontology to be able to import components of other ontologies without conflict) and the construction of "accurate representations of biological reality."

The NCBO (National Center for Biomedical Ontology, http://bioontology.org) supports the OBO Consortium by providing tools and resources to help manage the ontologies and to help the scientific community access, query, visualize, and use them to annotate experimental data (68). The NCBO's BioPortal website provides searches across multiple ontologies and contains a large library of these ontologies spanning many species and many scales, from molecules to whole organism. The ontology content comes from the model organism communities, biology, chemistry, anatomy, radiology, and medicine.

Together, the OBO Consortium and the NCBO are helping to construct a consistent arsenal of ontologies to promote their application in annotating Omics and other biological experiments. This is the sort of community-based ontology building that holds great potential to help the life science community convert the complex and daunting Omics data sets into new discoveries that expand our knowledge and improve human health.

3.9. Concluding on the Need for Standards

A key motivation behind Omics standards is to foster data sharing, reuse, and integration with the ultimate goal of producing new biological insights (within basic research environments) and better medical treatments (within healthcare environments). Widely adopted minimum information guidelines for publication and formats for data exchange are leading to increased and better reporting of results and submission of experimental data into public repositories, and more effective data mining of large Omics data sets. Standards harmonization efforts are in progress to improve data integration and interoperability of software within both basic research settings as well as within healthcare environments. Standard reference materials and protocols for their use are also under active development and hold much promise for improving data quality, systems benchmarking, and facilitating the use of Omics technologies within clinical and diagnostic settings.

High-throughput Omics experiments, with their large and complex data sets, have posed many challenges to the creation and adoption of standards. However, in recent years, the standards initiatives in this field have risen to the challenge and continue to engage their respective communities to improve the fit of the standards to user and market needs.

Omics communities have recognized that standards-compliant software tools can go a long way towards enhancing the adoption and usefulness of a standard by enabling ease-of-use. For data exchange standards, such tools can "hide the technical complexities of the standard and facilitate manipulation of the standard format in an easy way" (8). Some tools can themselves become part of standard practice when they are widely used throughout a community. Efforts are underway within organizations such as MGED and HUPO PSI to enhance the usefulness of tools for end user scientists working with standard data formats in order to ease the process of data submission, annotation, and analysis.

The widespread adoption of some of the more mature Omics standards by large numbers of life science researchers, data analysts, software developers, and journals has had a number of benefits. Adoption has promoted data sharing and reanalysis, facilitated publication, and spawned a number of data repositories to store data from Omics experiments. A higher citation rate and other benefits have been detected for researchers who share their data (7,

69). Estimates of total volume of high-throughput data available in the public domain are complex to calculate, but a list of databases maintained by the Nucleic Acids Research journal (http://www3.oup.co.uk/nar/database/a) contained more than 1,000 databases in areas ranging from nucleic acid sequence data to experimental archives and specialist data integration resources (70). More public databases appear every year and as technologies change, so that deep sequencing of genomes and transcriptomes becomes more cost effective, the volume will undoubtedly rise even further.

Consistent annotation of this growing volume of Omics data using interoperable ontologies and controlled vocabularies will play an important role in enabling collaborations and reuse of the data by other third parties. More advanced forms of knowledge integration that rely on standard terminologies are beginning to be explored using semantic web approaches (71–73).

Adherence to standards by public data repositories is expected to facilitate data querying and reuse. Even in the absence of strict standards (such as compliance requirements upon data submission), useful data mining can be performed from large bodies of raw data originating from the same technology platform (74), especially if standards efforts make annotation guidelines available and repositories encourage their use. Approaches such as this may help researchers better utilize the limited levels of consistently annotated data in the public domain.

It was recently noted that only a fraction of data generated is deposited in public data repositories (75). Improvements in this area can be anticipated through the proliferation of better tools for bench scientists that make it easier for them to submit their data in a consistent, standards-compliant manner. The full value of Omics research will only be realized once scientists in the laboratory and the clinic are able to share and integrate over large amounts of Omics data as easily as they can now do so with primary biological sequence data.

References

1. Boguski MS. Biosequence exegesis. Science. 1999; 286(5439):453–455. [PubMed: 10521334]

2. Brazma A. On the importance of standardisation in life sciences. Bioinformatics. 2001; 17(2):113–114. [PubMed: 11238066]

3. Stoeckert CJ Jr. Causton HC, Ball CA. Microarray databases: standards and ontologies. Nat Genet. 2002; 32:469–473. [PubMed: 12454640]

4. Brooksbank C, Quackenbush J. Data standards: a call to action. OMICS. 2006; 10(2):94–99. [PubMed: 16901212]

5. Rogers S, Cambrosio A. Making a new technology work: the standardization and regulation of microarrays. Yale J Biol Med. 2007; 80(4):165–178. [PubMed: 18449388]

6. Warrington, JA. Book of Genes and Medicine. Osaka: Medical Do Co; 2008. Standard controls and protocols for microarray based assays in clinical applications.

7. Piwowar HA, et al. Towards a data sharing culture: recommendations for leadership from academic health center. PLoS Med. 2008; 5(9):e183. [PubMed: 18767901]

8. Brazma A, Krestyaninova M, Sarkans U. Standards for systems biology. Nat Rev Genet. 2006; 7(8):593–605. [PubMed: 16847461]

9. Brazma A, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001; 29(4):365–371. [PubMed: 11726920]

10. Spellman PT, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. 2002; 3(9) RESEARCH0046.

11. Whetzel PL, et al. The MGED ontology: a resource for semantics-based description of microarray experiments. Bioinformatics. 2006; 22(7):866–873. [PubMed: 16428806]

12. Parkinson H, et al. ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009; 37(Database issue):D868–D872. [PubMed: 19015125]

13. Parkinson H, et al. ArrayExpress – a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007; 35(Database issue):D747–D750. [PubMed: 17132828]

14. Parkinson H, et al. ArrayExpress – a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2005; 33(Database issue):D553–D555. [PubMed: 15608260]

15. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 2006; 411:352–369. [PubMed: 16939800]

16. Barrett T, et al. NCBI GEO: mining millions of expression profiles – database and tools. Nucleic Acids Res. 2005; 33(Database issue):D562–D566. [PubMed: 15608262]

17. Barrett T, et al. NCBI GEO: mining tens of millions of expression profiles – database and tools update. Nucleic Acids Res. 2007; 35(Database issue):D760–D765. [PubMed: 17099226]

18. Barrett T, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009; 37(Database issue):D885–D890. [PubMed: 18940857]

19. Taylor CF, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007; 25(8):887–893. [PubMed: 17687369]

20. Shi L, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006; 24(9):1151–1161. [PubMed: 16964229]

21. Taylor CF, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol. 2008; 26(8):889–896. [PubMed: 18688244]

22. DeFrancesco L. Journal trio embraces MIAME. Genome Biol. 2002; 8(6):R112.

23. Jones AR, Paton NW. An analysis of extensible modelling for functional genomics data. BMC Bioinformatics. 2005; 6:235. [PubMed: 16188029]

24. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25–29. [PubMed: 10802651]

25. Smith B, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007; 25(11):1251–1255. [PubMed: 17989687]

26. Salit M. Standards in gene expression microarray experiments. Methods Enzymol. 2006; 411:63–78. [PubMed: 16939786]

27. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16):2078–2079. [PubMed: 19505943]

28. Brookes AJ, et al. The phenotype and genotype experiment object model (PaGE-OM): a robust data structure for information related to DNA variation. Hum Mutat. 2009; 30(6):968–977. [PubMed: 19479963]

29. Brazma A, Parkinson H. ArrayExpress service for reviewers/editors of DNA microarray papers. Nat Biotechnol. 2006; 24(11):1321–1322. [PubMed: 17093465]

30. Rayner TF, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics. 2006; 7:489. [PubMed: 17087822]

31. Rayner TF, et al. MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB. Bioinformatics. 2009; 25(2):279–280. [PubMed: 19038988]

32. Manduchi E, et al. RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. Bioinformatics. 2004; 20(4):452–459. [PubMed: 14990441]

33. Ball CA, et al. The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res. 2005; 33(Database issue):D580–D582. [PubMed: 15608265]

34. Demeter J, et al. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res. 2007; 35(Database issue):D766–D770. [PubMed: 17182626]

35. Gollub J, et al. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res. 2003; 31(1):94–96. [PubMed: 12519956]

36. Gollub J, Ball CA, Sherlock G. The Stanford Microarray Database: a user's guide. Methods Mol Biol. 2006; 338:191–208. [PubMed: 16888360]

37. Hubble J, et al. Implementation of GenePattern within the Stanford Microarray Database. Nucleic Acids Res. 2009; 37(Database issue):D898–D901. [PubMed: 18953035]

38. Sherlock G, et al. The Stanford Microarray Database. Nucleic Acids Res. 2001; 29(1):152–155. [PubMed: 11125075]

39. Navarange M, et al. MiMiR: a comprehensive solution for storage, annotation and exchange of microarray data. BMC Bioinformatics. 2005; 6:268. [PubMed: 16280078]

40. Allison M. Is personalized medicine finally arriving? Nat Biotechnol. 2008; 26(5):509–517. [PubMed: 18464779]

41. Orchard S, Hermjakob H. The HUPO proteomics standards initiative - easing communication and minimizing data loss in a changing world. Brief Bioinform. 2008; 9(2):166–173. [PubMed: 18065433]

42. Pedrioli PG, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol. 2004; 22(11):1459–1466. [PubMed: 15529173]

43. Keller A, et al. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. 2005; 1:0017. [PubMed: 16729052]

44. Deutsch E. mzML: a single, unifying data format for mass spectrometer output. Proteomics. 2008; 8(14):2776–2777. [PubMed: 18655045]

45. Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiol Genomics. 2008; 33(1):18–25. [PubMed: 18212004]

46. Orchard S, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol. 2007; 25(8):894–898. [PubMed: 17687370]

47. Kerrien S, et al. Broadening the horizon - level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. 2007; 5:44. [PubMed: 17925023]

48. Fiehn O, et al. Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. OMICS. 2006; 10(2):158–163. [PubMed: 16901221]

49. Sansone SA, et al. The metabolomics standards initiative. Nat Biotechnol. 2007; 25(8):846–848. [PubMed: 17687353]

50. Goodacre R, et al. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics. 2007; 3(3):231–241.

51. Hardy N, Taylor C. A roadmap for the establishment of standard data exchange structures for metabolomics. Metabolomics. 2007; 3(3):243–248.

52. Jenkins H, Johnson H, Kular B, Wang T, Hardy N. Toward supportive data collection tools for plant metabolomics. Plant Physiol. 2005; 138(1):67–77. [PubMed: 15888680]

53. Jenkins H, et al. A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol. 2004; 22(12):1601–1606. [PubMed: 15583675]

54. Spasic I, et al. MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. BMC Bioinformatics. 2006; 7:281. [PubMed: 16753052]

55. Sansone S-A, Schober D, Atherton H, Fiehn O, Jenkins H, Rocca-Serra P, et al. Metabolomics standards initiative: ontology working group work in progress. Metabolomics. 2007; 3(3):249–256.

56. Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, et al. A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol. 2004; 22(12):1601–1606. [PubMed: 15583675]

57. Kumar D. From evidence-based medicine to genomic medicine. Genomic Med. 2007; 1(3–4):95–104. [PubMed: 18923934]

58. Fostel JM. Towards standards for data exchange and integration and their impact on a public database such as CEBS (Chemical Effects in Biological Systems). Toxicol Appl Pharmacol. 2008; 233(1):54–62. [PubMed: 18680759]

59. Bland PH, Laderach GE, Meyer CR. A web-based interface for communication of data between the clinical and research environments without revealing identifying information. Acad Radiol. 2007; 14(6):757–764. [PubMed: 17502264]

60. Meslin EM. Shifting paradigms in health services research ethics. Consent, privacy, and the challenges for IRBs. J Gen Intern Med. 2006; 21(3):279–280. [PubMed: 16637827]

61. Ferris TA, Garrison GM, Lowe HJ. A proposed key escrow system for secure patient information disclosure in biomedical research databases. Proc AMIA Symp. 2002:245–249. [PubMed: 12463824]

62. Quackenbush J, et al. Top-down standards will not serve systems biology. Nature. 2006; 440(7080):24. [PubMed: 16511469]

63. Jones AR, et al. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol. 2007; 25(10):1127–1133. [PubMed: 17921998]

64. Sansone SA, et al. The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?". OMICS. 2008; 12(2):143–149. [PubMed: 18447634]

65. Sansone SA, et al. A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. OMICS. 2006; 10(2):164–171. [PubMed: 16901222]

66. Whetzel PL, et al. Development of FuGO: an ontology for functional genomics investigations. OMICS. 2006; 10(2):199–204. [PubMed: 16901226]

67. Smith B, et al. Relations in biomedical ontologies. Genome Biol. 2005; 6(5):R46. [PubMed: 15892874]

68. Rubin DL, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS. 2006; 10(2):185–198. [PubMed: 16901225]

69. Piwowar HA, Chapman WW. Identifying data sharing in biomedical literature. AMIA Annu Symp Proc. 2008:596–600. [PubMed: 18998887]

70. Galperin MY, Cochrane GR. Nucleic Acids Research annual Database issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic Acids Res. 2009; 37(Database issue):D1–D4. [PubMed: 19033364]

71. Ruttenberg A, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007; 8(Suppl 3):S2. [PubMed: 17493285]

72. Sagotsky JA, et al. Life Sciences and the web: a new era for collaboration. Mol Syst Biol. 2008; 4:201. [PubMed: 18594519]

73. Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. Nat Rev Genet. 2008; 9(9):678–688. [PubMed: 18714290]

74. Day A, et al. Celsius: a community resource for Affymetrix microarray data. Genome Biol. 2007; 8(6):R112. [PubMed: 17570842]

75. Ochsner SA, et al. Much room for improvement in deposition rates of expression microarray datasets. Nat Methods. 2008; 5(12):991. [PubMed: 19034265]

Table 1

Domains of Omics standards. The domain indicates the type of experimental data that the standard is designed to handle

Domain	Description
Genomics	Genome sequence assembly, genetic variations, genomes and metag-enomes, and DNA modifcations
Transcriptomics	Gene expression (transcription), alternative splicing, and promoter activity
Proteomics	Protein identifcation, protein–protein interactions, protein abundance, and posttranslational modifcations
Metabolomics	Metabolite profiling, pathway flux, and pathway perturbation analysis
Healthcare and Toxicogenomics[a]	Clinical, diagnostic, or toxicological applications
Harmonization and Multiomics[a]	Cross-domain compatibility and interoperability

[a]Healthcare, toxicological, and harmonization standards may be applicable to one or more other domain areas. These domains impose additional requirements on top of the needs of the pure Omics domains

Table 2

Scope of Omics standards. Scope defines the area of applicability or methodology to which the standard pertains. Scope-General: Standards can be generally partitioned based on whether they are to be used for describing or executing an experiment. Scope-Specific: The scope can be further narrowed to cover more Specific aspects of the general scope

Scope-General	Scope-Specific	Description
Experiment description	Reporting (Minimum information)	Documentation for publication or data deposition
	Data exchange & modeling	Communication between organizations and tools
	Terminology	Ontologies and CV's to describe experiments or data
Experiment execution	Physical standards	Reference materials, spike-in controls
	Data analysis & quality metrics	Analyze, compare, QA/QC experimental results

CV controlled vocabulary, *QA/QC* quality assurance/quality control

Table 3

Existing reporting standards for Omics

Acronym	Full name	Domain	Organization
CIMR	Core Information for Metabolomics Reporting	Metabolomics	MSI
MIAME	Minimum Information about a Microarray Experiment	Transcriptomics	MGED
MIAPE	Minimum Information about a Proteomics Experiment	Proteomics	HUPO-PSI
MIGS-MIMS	Minimum Information about a Genome/Metagenome Sequence	Genomics	GSC
MIMIx	Minimum Information about a Molecular Interaction eXperiment	Proteomics	HUPO-PSI
MINIMESS	Minimal Metagenome Sequence Analysis Standard	Metagenomics	GSC
MINSEQE	Minimum Information about a high-throughput Nucleotide Sequencing Experiment	Genomics, Transcriptomics (UHTS)	MGED
MISFISHIE	Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments	Transcriptomics	MGED

Acronyms and definitions of the major reporting standards efforts are shown, indicating their target domain and the maintaining organization, which are as follows: *MGED* MGED Society, http://mged.org; *GSC* Genomic Standards Consortium, http://gensc.org; *HUPO-PSI* Human Proteome Organization Proteomics Standards Initiative, http://www.psidev.info; *MSI* Metabolomics Standards Initiative, http://msi-workgroups.sourceforge.net

Table 4

A sampling of data exchange and modeling standards for Omics

Acronym				
Data format	Object model	Full name	Domain	Organization
FuGE-ML	FuGE-OM	Functional Genomics Experiment Markup Language/Object Model	Multiomics	FuGE
ISA-TAB		Investigation Study Assay – Tabular	Multiomics	RSBI
MAGE-ML	MAGE-OM	MicroArray and Gene Expression Markup Language	Transcriptomics	MGED
MAGE-TAB		MicroArray and Gene Expression Tabular Format		
MIF (PSI-MI XML)		Molecular Interactions Format	Proteomics	HUPO-PSI
mzML		Mass Spectrometry Markup Language		
mzIdentML		Mass Spectrometry Identifcations Markup Language		
PML	PAGE-OM	Polymorphism Markup Language/ Phenotype and Genotype Object Model	Genomics	GEN2PHEN
	SDTM	Study Data Tabulation Model	Healthcare	CDISC

Acronyms and names of some of the major data exchange standards efforts are shown, indicating their target domain and the maintaining organization, which are as described in the legend to Table 3 with the following additions: *RSBI* Reporting Structure for Biological Investigations, http://www.mged.org/Workgroups/rsb; *FuGE* Functional Genomics Experiment, http://fuge.sourceforge.net; *GEN2PHEN* Genotype to phenotype databases, http://www.gen2phen.org; *CDISC* Clinical Data Interchange Standards Consortium, http://www.cdisc.org. Additional proteomics exchange standards are described on the HUPO-PSI website, http://www.psidev.info

Table 5

Terminology standards

Acronym	Full name	Domain	Organization
EVS	Enterprise Vocabulary Services	Healthcare	NCI
GO	Gene Ontology	Multiomics	GOC
MS	Proteomics Standards Initiative Mass Spectrometry controlled vocabulary	Proteomics	HUPO-PSI
MO	MGED Ontology	Transcriptomics	MGED
OBI	Ontologies for Biomedical Investigators	Multiomics	OBI
OBO	Open Biomedical Ontologies	Multiomics	NCBO
PSI-MI	Proteomics Standards Initiative Molecular Interactions ontology	Proteomics	HUPO-PSI
sepCV	Sample processing and separations controlled vocabulary	Proteomics	HUPO-PSI
SO	Sequence Ontology	Multiomics	GOC

Acronyms and names of some of the major terminology standards in use with Omics data are shown, indicating their target domain and the maintaining organization, which are as described in the legends to Tables 3 and 4 with the following additions: *GOC* Gene Ontology Consortium, http://geneontology.org/GO.consortiumlist.shtml; *NCI* National Cancer Institute, http://www.cancer.gov; *NCBO* National Center for Biomedical Ontology, http://bioontology.org; *OBI* Ontology Biomedical Investigations, http://purl.obofoundry.org/obo/obi

Table 6

Organizations involved in the creation of physical standards relevant to Omics experiments

Acronym	Full name	Domain	Website
ERCC	External RNA Control Consortium	Transcriptomics	http://www.cstl.nist.gov/biotechCell&TissueMeasurements/GeneExpression/ERCC.htm
LGC	Laboratory of the Government Chemist	Transcriptomics, Proteomics	http://www.lgc.co.uk
NIST	National Institute for Standards Technology	Transcriptomics	http://www.cstl.nist.gov/biotech/Cell&TissueMeasurements/Main_Page.htm
NMS	National Measurement System (NMS) Chemical and Biological Metrology	Multiomics	http://www.nmschembio.org.uk
ATCC	American Type Culture Collection Standards Development Organization	Healthcare	http://www.atcc.org/Standards/ATCCStandardsDevelopmentOrganizationSDO/tabid/233/Default.aspx

Table 7

Data analysis and quality metrics projects

Acronym	Full name	Domain	Organization
arrayQuality-Metrics	Quality assessment software package	Transcriptomics	BioConductor
CAMDA	Critical Assessment of Microarray Data Analysis	Transcriptomics	n/a
CAMSI	Critical Assessment of Mass Spectrometry Informatics	Proteomics	n/a
iPRG	Proteome Informatics Research Group	Proteomics	ABRF
MAQC	Microarray Quality Control Project	Transcriptomics	FDA
NTO	Normalization and Transformation Ontology	Transcriptomics	EMERALD

BioConductor's arrayQualityMetrics: http://bioconductor.org/packages/2.3/bioc/html/arrayQualityMetrics.html. CAMDA is managed by a local organizing committee at different annual venues: http://camda.bioinfo.cipf.es. EMERALD's NTO: http://www.microarray-quality.org/ontology_work.html. MAQC is described in Subheading 3.2.6

Table 8

Summary of healthcare experiment description standards initiatives

Acronym	Full name	Description	Scope	Website
BIRN	Biomedical Informatics Research Network	Collaborative informatics resources medical/clinical data	Data analysis; Terminology	http://www.nbirn.net
CDISC	Clinical Data Interchange Standards Consortium	Regulatory submissions of clinical data	Data exchange & modeling	http://www.cdisc.org
CONSORT	Consolidated Standards of Reporting Trials	Minimum requirements for reporting randomized clinical trials	Reporting	http://www.consort-statement.org
EVS	Enterprise Vocabulary Services	Controlled vocabulary by the National Cancer Institute in support of cancer	Terminology	http://www.cancer.gov/cancertopics/terminologyresources
HL7	Health Level 7	Programmatic data exchange for healthcare applications	Data exchange	http://www.hl7.org
SEND	Standards for Exchange of Preclinical Data	Regulatory submissions of preclinical data; based on CDISC	Data exchange & modeling	http://www.cdisc.org/standards
ToxML	Toxicology XML	Toxicology data exchange; based on controlled vocabulary	Data exchange; terminology	http://www.leadscope.com/toxml.php

Table 9

CLSI documents most relevant to functional genomics technologies

Document	Description	Status
MM12-A	Diagnostic Nucleic Acid Microarrays	Approved guideline
MM14-A	Proficiency Testing (External Quality Assessment) for Molecular Methods	Approved guideline
MM16-A	Use of External RNA Controls in Gene Expression Assays	Approved guideline
MM17-A	Verification and Validation of Multiplex Nucleic Acid Assays	Approved guideline

Table 10

Existing Omics standards harmonization projects and initiatives

Acronym	Full name	Scope	Organization
FuGE-ML FuGE-OM	Functional Genomics Experiment Markup Language/Object Model	Data exchange & modeling	FuGE
ISA-TAB	Investigation Study Assay Tabular Format	Data exchange	RSBI, GSC, MSI, HUPO-PSI
HITSP	Healthcare Information Technology Standards Panel	(various)	ANSI
MIBBI	Minimum Information for Biological and Biomedical Investigations	Reporting	MIBBI
OBI	Ontologies for Biomedical Investigations	Terminology	OBI
OBO	Open Biomedical Ontologies	Terminology	NCBO
P^{3}G	Public Population Project in Genomics	(various)	International Consortium

P^{3}G covers harmonization between genomic biobanks and longitudinal population genomic studies including technical, social, and ethical issues: http://www.p3gconsortium.org. The other projects noted in this table are described further in the chapter

Table 11

Programmatic tools for dealing with standards, ontologies and common data formats

Tool name	Language	Purpose	Website
Limpopo	Java	MAGE-TAB parser	http://sourceforge.net/projects/limpopo/
MAGEstk	Perl and Java	MAGE-ML toolkit	http://www.mged.org/Workgroups/MAGE/magestk.html
MAGE-Tab module	Perl	MAGE-TAB API	http://magetabutils.sourceforge.net/
OntoCat	Java	Ontology access tool for OWL, OBO format files and ontology web services	http://ontocat.sourceforge.net/
OWL-API	Java	Reading and querying OWL and OBO format files	http://owlapi.sourceforge.net/

Table 12

Freely available standards related format conversion tools

Tool name	Language	Formats supported	Website
MAGETabulator	Perl	SOFT to MAGE-TAB	http://tab2mage.sourceforge.net
MAGETabulator	Perl	MAGE-TAB to MAGE-ML	http://tab2mage.sourceforge.net
ArrayExpress Package	R (Bioconductor)	MAGE-TAB to R objects	http://www.bioconductor.org/packages/bioc/html/ArrayExpress.html
GEOquery	R (Bioconductor)	GEO SOFT to R objects	http://www.bioconductor.org/packages/1.8/bioc/html/GEOquery.html
ISA-Creator	Java	ISA-TAB to MAGE-TAB	http://isatab.sourceforge.net/tools.html
ISA-Creator	Java	ISA-TAB to Pride XML	http://isatab.sourceforge.net/tools.html
ISA-Creator	Java	ISA-TAB to Short Read Archive XML	http://isatab.sourceforge.net/tools.html

Table 13

Standards compliant data annotation tools

Tool name	Language	Purpose	Website
Annotare	Adobe Air/Java	Desktop MAGE-TAB annotation application	http://code.google.com/p/annotare/
MAGETabulator	Perl	MAGE-TAB template generation and related database	http://tab2mage.sourceforge.net
caArray	Java	MAGE-TAB Data management solution	https://array.nci.nih.gov/caarray/home.action
ISA-Creator	Java	ISA-TAB annotation application	http://isatab.sourceforge.net/tools.html