Abstract

Biological processes are fundamentally driven by complex interactions between biomolecules. Integrated high-throughput omics studies enable multifaceted views of cells, organisms, or their communities. With the advent of new post-genomics technologies, omics studies are becoming increasingly prevalent; yet the full impact of these studies can only be realized through data harmonization, sharing, meta-analysis, and integrated research. These essential steps require consistent generation, capture, and distribution of metadata. To ensure transparency, facilitate data harmonization, and maximize reproducibility and usability of life sciences studies, we propose a simple common omics metadata checklist. The proposed checklist is built on the rich ontologies and standards already in use by the life sciences community. The checklist will serve as a common denominator to guide experimental design, capture important parameters, and be used as a standard format for stand-alone data publications. The omics metadata checklist and data publications will create efficient linkages between omics data and knowledge-based life sciences innovation and, importantly, allow for appropriate attribution to data generators and infrastructure science builders in the post-genomics era. We ask that the life sciences community test the proposed omics metadata checklist and data publications and provide feedback for their use and improvement.

A Common Omics Metadata Checklist Proposal

Modern life science technologies enable rapid and efficient acquisition of omics data. These data comprehensively measure multilayered molecular networks and provide a snapshot of biological processes in a cell, organism, or their communities. Collected on the same sample at the same time, omics data provide information on the functioning of biomolecules and their interactions. Omics studies are essential for the
systemic investigation of biological systems—an endeavor that is crucial to improve our ability to manage and cure diseases, identify drug targets, understand regulatory cascades, and predict ecosystem responses to environmental changes.

Through the pioneering efforts of Drs. Smarr and Snyder,1–4 two powerful multi-omics human datasets were recently made available. Smarr’s dataset includes a wide variety of molecular measures and clinical parameters meticulously collected and cataloged for years, while Snyder’s integrative personal multi-omics study presents his personal genomics, transcriptomics, proteomics, metabolomics, and autoantibody profiles collected over a 14-month period. Both studies yielded unique physiological insights not previously possible, including early indications of vulnerabilities to specific diseases.

In the near future these kinds of personal omics studies will become routine and will inevitably result in vast and diverse volumes of omics data. Therefore, the scientific community must commit to a common format for publishing the design and analysis of these studies that will ensure the compatibility, reproducibility, and reuse of the resulting data.

The use, integration, and reuse of data require accurate and comprehensive capture of the associated metadata, including details describing experimental design, sample acquisition and preparation, instrument protocols, and processing steps. The data and metadata must be captured together in a rigorous and consistent manner to allow the integration of data across omics experiments. The use of ontologies, naming conventions, and standards can increase the compatibility and usability of these diverse data. However, life sciences data have certain core similarities. However, combined with these similarities come the different nuances among various technology platforms, such as transcriptomics, proteomics, and metabolomics, as well as application contexts such as neuroscience and hematology. The differences are compounded by the multiplicity of standards within a field—transcriptomics alone has at least 15 standards potentially applicable to the data.5,6 Such complexities not only make reproducible, integrative, accurate, and comprehensive capture of data and metadata an intricate challenge that must be overcome but also place an excessive burden on researchers trying to convey metadata.5,7

Pioneering attempts in this area were made in 2007 when the Minimum Information about a Biomedical or Biological Investigation project brought many of these efforts for the life sciences together into an umbrella organization: MIBBI.8,9 In MIBBI, each set of guidelines is developed by a working group concentrated in a specific field (for example, functional magnetic resonance imaging [fMRI] or quantitative trait locus [QTL] and association studies). Other types of data sharing tools that have also been harmonized in MIBBI include single omics checklists such as Minimum Information About a Proteomics Experiment (MIAPE) by the Human Proteome Organization10 and the Minimum Information About a Microarray Experiment (MIAME) by the Microarray Gene Expression Data Society.11 Through this approach, MIBBI aspires to capture all essential metadata and data that are necessary to replicate any given experiment within a field. Also, the framework known as Minimal Information about any Sequence (MiXSiS) expands the breadth of information available.
by integrating the individual genomics checklists developed by the Genomics Standards Consortium with environmental information. The checklist requires the researcher and/or corresponding author to enter specific information on experimental design, statistical analysis, and reagents. This checklist is endorsed by the Data-Enabled Life Sciences Alliance (DELSA Global).

The unveiling of the Nature publication initiative brought into focus the need for a complementary omics checklist that allows the capture and publication of critical metadata associated with omics data sets. To this end, life sciences researchers from DELSA Global propose a single common omics metadata checklist as described below. By integrating DELSA researchers’ collective experiences with omics guidelines and publication requirements, one simplified, yet informative and flexible checklist was created to capture the essential aspects of omics studies.

Publication of a completed checklist will serve to inform the life sciences community of the details needed to properly utilize the given data set. This type of “resource publication” has long been done by Nucleic Acid Research in its annual database issue. NanoPubs and MicroPubs are two newer publication avenues that could serve to quickly and accurately share information. There are also other forms of data publications including, for example, Investigation, Study, Assay (ISA) metadata tracking tools and the journal Scientific Data.

It is worth noting that multi-omics data from a longitudinal study of a single individual (e.g., the Smarr and Snyder datasets) in their entirety constitute essentially a whole new data type. Supplied with detailed metadata, these data could become a part of a greater, well-documented collage of data within a specific domain. Because of the large amount of data and the complexity of data acquisition, it is exceedingly difficult to capture, disseminate, and interpret the metadata. Generally, minimal reporting requirements are aimed at enabling replication of an experiment, a concept that is not easily applied to the longitudinal personal omics studies.

Reuse of data can be enabled with more succinct and concise reporting.

The checklist we propose, therefore, has a simple structure covering four concise sections: experiment information, experimental design, experimental methods, and data processing. The experiment information section includes details of the lab, funding sources, data identification, and a brief abstract to address why the experiment was done. The experimental design section is meant to capture the high-level data about the experiment and its statistical design, including sample selection, replication, and randomization. The experimental methods section contains details about instrumentation and sample preparation. The data processing section captures information regarding methods and tools used in experimental data processing and data analysis (see Table 1). The metadata captured by this checklist will serve as interlocking bridges for data harmonization. Therefore the checklist focuses on details of the experimental design and subsequent data analyses. In multi-omics studies, the researcher would fill a checklist for each omics data-type measured. As test cases, two datasets of the integrative personal multi-omics study were used. The proposed checklist integrates existing ontologies and standards in order to standardize terminology and simplify data input. In its short, structured form, the checklist captures important experimental parameters and strikes a balance between comprehensiveness and ease of use. As such, the checklist can serve as a guide to the design of omics studies.

Implementation of this checklist will enable efficient portability and meta-analysis of the data, as well as transparent communication and greater reproducibility of omics studies. Yet the checklist is just the first step toward full utilization of the data. Traditional publication avenues and new data publications, for example, OMICS Journal of Integrative Biology, Journal of Proteome Research, Big Data, eLife, and Scientific Data, could test and adopt the format to ensure that the crucial information needed to allow data to be harmonized for broader usage is published. The assessment of the metadata quality and the data they accompany could be done through community resources like PubMed Commons.

Data submissions to single omics databases such as, for example, ArrayExpress and GEO for transcriptomics, or PRIDE and ProteomeXchange for proteomics, would benefit from both additional omics metadata within the given database and robust harmonization with other data-types in other databases.
multi-omics databases, data repositories, or data clouds. Examples include data clouds, such as The Open Science Data Cloud; data repositories, such as Dryad for raw data; and MOPED for processed data. When compatibility and sharing of data and metadata cease to be an issue, a deeper understanding of cells, organisms, and their communities will ensue.

Conclusions

The proposed metadata checklist offers a much-needed and balanced approach to bring about data harmonization across omics studies. This is accomplished while also maintaining the flexibility needed to adapt to complex and ever-evolving study designs and omics application contexts in the post-genomics era of the life sciences.

Acknowledgments

All authors of this publication are members of the Data-Enabled Life Sciences Alliance (DELSA Global). This manuscript has been reviewed and endorsed by the Alliance. Research reported in this study was supported by the National Science Foundation (NSF) under the Division of Biological Infrastructure award 0969929, National Institute of...
Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (NIH) under awards U01DK089571 and U01DK072473, Seattle Children’s Research Institute (SCRI), The Robert B. McMillen Foundation, EMC, Intel, and The Gordon and Betty Moore Foundation award to E.K. This support is very much appreciated. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH, SCRI, the McMillen Foundation, EMC, Intel or The Moore Foundation. The omics metadata checklist presented here is being published in parallel in OMICS: A Journal of Integrative Biology, 2014, Volume 18, Issue 1, due to its broad and community-wide importance.

Author Disclosure Statement

No competing financial interests exist.

References

25. ISA tools. 2013. isa-tools.org

Address correspondence to:
Eugene Kolker
Bioinformatics and High-Throughput Analysis Laboratory
Seattle Children’s Research Institute
Predictive Analytics, Seattle Children’s
1900 Ninth Avenue
Seattle, WA 98101
E-mail: eugene.kolker@seattlechildrens.org
This article has been cited by:

1. Dumbill Edd, Kolker Eugene. 2013. Introducing a Metadata Checklist for Omics Data. *Big Data* 1:4, 195-195. [Citation] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links]