Commentary

Toward More Transparent and Reproducible Omics Studies Through a Common Metadata Checklist and Data Publications

Eugene Kolker,[1–3] Vural Özdemir,[3–5] Lennart Martens,[3,6,7] William Hancock,[3,8] Gordon Anderson,[3,9] Nathaniel Anderson,[1,3] Sukru Aynacioglu,[3,10] Ancha Baranova,[3,11] Shawn R. Campagna,[3,12] Rui Chen,[3,13] John Choiniere,[1,3] Stephen P. Dearth,[3,12] Wu-Chun Feng,[3,14–16] Lynnette Ferguson,[3,17] Geoffrey Fox,[3,18] Dmitrij Frishman,[3,19] Robert Grossman,[3,20,21] Allison Heath,[3,20,22] Roger Higdon,[1–3] Mara H. Hutz,[3,23] Imre Janko,[3,24] Lihua Jiang,[3,13] Sanjay Joshi,[3,25] Alexander Kel,[3,26] Joseph W. Kemnitz,[3,27,28] Isaac S. Kohane,[3,29,30] Natali Kolker,[2,3,24] Doron Lancet,[3,31] Elaine Lee,[3,24] Weizhong Li,[3,32] Andrey Lisitsa,[3,33,34] Adrian Llerena,[3,35] Courtney MacNealy-Koch,[1,3] Jean-Claude Marshall,[3,36] Paola Masuzzo,[3,6,7] Amanda May,[3,12] George Mias,[3,13] Matthew Monroe,[3,37] Elizabeth Montague,[1–3] Sean Mooney,[3,38] Alexey Nesvizhskii,[3,39,40] Santosh Noronha,[3,41] Gilbert Omenn,[3,42–45] Harsha Rajasimha,[3,46] Preveen Ramamoorthy,[3,47] Jerry Sheehan,[3,48] Larry Smarr,[3,48] Charles V. Smith,[3,49] Todd Smith,[3,50] Michael Snyder,[3,13,51] Srikanth Rapole,[3,52] Sanjeeva Srivastava,[3,53] Larissa Stanberry,[1–3] Elizabeth Stewart,[1,3] Stefano Toppo,[3,54] Peter Uetz,[3,55] Kenneth Verheggen,[3,6,7] Brynn H. Voy,[3,56] Louise Warnich,[3,57] Steven W. Wilhelm,[3,58] and Gregory Yandl[1,3]

[1]Bioinformatics and High-Throughput Analysis Laboratory, Seattle Children's Research Institute, Seattle, Washington; [2]Predictive Analytics, Seattle Children's, Seattle, Washington; [3]Data-Enabled Life Sciences Alliance (DELSA Global), Seattle, Washington; [4]Office of the President, Gaziantep University, International Affairs and Global Development Strategy; [5]Faculty of Communications, Universite Bulvarı, Kilis Yolu, Turkey; [6]Department of Medical Protein Research, Vlaams Instituut voor Biotechnologie, Ghent, Belgium; [7]Department of Biochemistry, Ghent University; Ghent, Belgium; [8]Department of Chemistry, Barnett Institute, Northeastern University, Boston, Massachusetts; [9]Fundamental and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington; [10]Department of Pharmacology, Gaziantep University, Gaziantep, Turkey; [11]School of Systems Biology, George Mason University, Manassas, Virginia; [12]Department of Chemistry, University of Tennessee Knoxville, Knoxville, Tennessee; [13]Department of Genetics, Stanford University, Stanford, California; Departments of [14]Computer Science and [15]Electrical and Computer Engineering, and [16]SyNeRGy Laboratory, Virginia Tech, Blacksburg, Virginia; [17]Department of Nutrition, Auckland Cancer Society Research Centre, University of Auckland, Auckland, New Zealand; [18]School of Informatics and Computing, Indiana University, Bloomington, Indiana; [19]Technische Universitat Munchen, Wissenshaftzentrum Weihenstephan, Freising, Germany; [20]Institute for Genomics and Systems Biology, [21]Department of Medicine, and [22]Knapp Center for Biomedical Discovery, University of Chicago, Chicago, Illinois; [23]Departamento de Genetica, Instituto de Biociencias, Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [24]High-Throughput Analysis Core, Seattle Children's Research Institute, Seattle, Washington; [25]Life Sciences, EMC, Hopkinton, Massachusetts; [26]GeneXplain GmbH, Wolfenbüttel, Germany; [27]Department of Cell and Regenerative Biology, and [28]Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, Wisconsin; [29]Pediatrics and Health Sciences Technology, Children's Hospital and Harvard Medical School, Boston, Massachusetts; [30]HMS Center for Biomedical Informatics, Countway Library of Medicine, Boston, Massachusetts; [31]Department of Molecular Genetics, Crown Human Genome Center, Weizmann Institute of Science, Rehovot, Israel; [32]Center for Research in Biological Systems, University of California, San Diego, La Jolla, California; [33]Russian Human Proteome Organization (RHUPO), Moscow, Russia; [34]Institute of Biomedical Chemistry, Moscow, Russia; [35]Clinical Research Center, Extremadura University Hospital and Medical School, Badajoz, Spain; [36]Center for Translational Research, Catholic Health Initiatives, Towson, Maryland; [37]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington; [38]The Buck Institute for Research on Aging, Novato, California; [39]Department of Pathology, University of Michigan, Ann Arbor, Michigan; [40]Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan; [41]Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai, India; [42]Center for Computational Medicine and Bioinformatics, and Departments of [43]Molecular Medicine & Genetics and Human Genetics and [44]Computational Medicine and Bioinformatics, and [45]School of Public Health, University of Michigan, Ann Arbor Michigan; [46]Jeeva Informatics Solutions LLC, Derwood, Maryland; [47]Molecular Diagnostics Department, National Jewish Health, Denver, Colorado; [48]California Institute for Telecommunications and Information Technology, University of California-San Diego, La Jolla, California; [49]Center for Developmental Therapeutics, Seattle Children's Research Institute, Seattle, Washington; [50]Digital World Biology, Seattle, Washington; [51]Stanford Center for Genomics and Personalized Medicine, Stanford University, Stanford, California; [52]Proteomics Laboratory, National Centre for Cell Science, University of Pune, Pune, India; [53]Proteomics Laboratory, Indian Institute of Technology Bombay, Mumbai, India; [54]Department of Molecular Medicine, University of Padova, Padova, Italy; [55]Center for the Study of Biological Complexity (CSBC), Virginia Commonwealth University, Richmond, Virginia; [56]Department of Animal Science, University of Tennessee Institute of Agriculture, Knoxville, Tennessee; [57]Department of Genetics, Faculty of AgriSciences, University of Stellenbosch, Stellenbosch, South Africa; [58]Department of Microbiology, University of Tennessee-Knoxville, Knoxville, Tennessee.

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, (Smarr, 2012; Bowden, 2012; Chen, 2012; Mias, 2013) 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. Fortunately, 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 (Tennenbaum, 2013; Field, 2009). 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 (Tennenbaum, 2013; Editorial, 2011).

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 (Taylor, 2008; Kettner, 2010). 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 Organization (Taylor, 2007) and the Minimum Information About a Microarray Experiment (MIAME) by the Microarray Gene Expression Data Society (Alvis, 2001). 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 (MIxS) expands the breadth of information available by integrating the individual genomics checklists developed by the Genomics Standards Consortium with environmental information (Yilmaz, 2011). In addition, the NIH's National Center for Biotechnology Information developed a format for cataloging information about samples enabling further metadata availability (Barrett, 2012). While these frameworks are critical to the reuse of data, they do not fully take into account the interlocking aspects needed for harmonization of diverse omics data types.

Recently, the Nature Publishing Group implemented a publication checklist that provides another example of an approach to improve the transparency and reproducibility of life sciences publications (Editorial, 2013; Reporting Checklist for Life Sciences Articles, 2013). 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) (Kolker, 2013).

TABLE 1. MULTI-OMICS METADATA CHECKLIST

Checklist version	1.0
Experiment information	*Description*
Lab name	Lab conducting the experiment
Date	Checklist submission date
Author information	Name, organization, contacts
Title of experiment	One-sentence description of the particular experiment
Project	Project name, ID, organization
Funding	Funding sources for the project
Digital ID	Multiple digital IDs may be listed, such as those to GEO, MOPED, PRIDE, DOIs, etc.
Abstract	A short description of the experiment briefly stating the goals of the research and principal outcomes if any (100 words or less)
Experimental design	*Description*
Organism	For example, Human, mouse
OMICS type(s) utilized	For example, Proteomics, metabolomics
Reference	Published articles that utilize these data, their PubMed identifier (PMID), or other relevant IDs or links
Experimental design	Design specifications; type of replication (biological, technical, time points); grouping of subjects; samples or replicates; randomization; comparisons; other salient design attributes
Sample description	Description of samples
Tissue/cell type ID	For example, BRENDA
Localization ID	For example, GO
Condition ID	DOID, text
Experimental methods	*Description*
Sample prep description	Description of steps taken, kits used
Platform type	For example, Microarray, LC-MS/MS, GC-MS, sequencing platform
Instrument name	For example, LTQ-Orbitrap, psi-MS ontology, HiSeq, IonTorrent, chip name (microarray)
Instrument details	For example, Ion source, mass analyzer
Instrument protocol	For example, Fragmentation method (CID, HCD, ETD), MS/MS scans per MS scan, sequencing cycles, paired ends, single reads, hybridization methods (microarray)
Data processing	*Description*
Processing/normalization methods/software	Description of processing and normalization methods & software
Sequence/annotation database	Source, version or date
ID method/software	Name of search engine used + post processing to ID molecules
ID/expression measures	For example, Thresholds and cutoffs for ID, spectral counts, peak area, reads, log2 expression
Data analysis method/software	Methods and software used for expression analysis, error estimation
I/O data file formats	List of file formats for raw and processed data (e.g., txt, xml, etc) specifications and software tools to ensure readability
Additional Information	Any additional information related to the experiment

ID, identification.

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 (Data-Enabled Life Sciences Alliance, 2013a; Kolker, 2012b; Kolker, 2012d; Kolker, 2012c; Stewart, 2013) 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 (Data-Enabled Life Sciences Alliance, 2013b).

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 (Nanopub, 2013; Micropub, 2013). There are also other forms of data publications including, for example, Investigation, Study, Assay (ISA) metadata tracking tools and the journal *Scientific Data* (ISA tools, 2013; Scientific Data, 2013).

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 (Snyder, 2013). 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 (OMICS, 2013; Journal of Proteome Research, 2013; eLife, 2013). The assessment of the metadata quality and the data they accompany could be done through community resources like PubMed Commons (PubMed Commons, 2013; Swartz, 2013).

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 (Rustici, 2013; Barrett, 2013; Vizcaino, 2013; ProteomeXchange, 2013). The checklist could also aid submissions to 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 (Open Science Data Cloud, 2013; Dryad, 2013; Higdon, 2013; Kolker, 2012a). 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

Author Disclosure Statement

No competing financial interests exist.

References

Alvis B, Hingamp P, Quackenbush J, et al. (2001). Minimum information about a microarray experiment (MIAME) – Towards standards for microarray data. Nat Genet 29, 365–371.

Barrett T, Clark K, Gevorgyan R, et al. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res 40, D57–D63.

Barrett T, Wilhite SE, Ledoux P, et al. (2013). NCBI GEO: Archive for functionally genomics data sets – Update. Nucleic Acids Res 41, D991–995.

Bowden M. (2012). The Measured Man. The Atlantic June; 13 2012. www.theatlantic.com/magazine/print/2012/0.7/the-measured-man/309018/. Last accessed October 17, 2013.

Chen R, Mias G, Li-Pook-Than J, et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148, 1293–1307.

Data-Enabled Life Sciences Alliance. (2013a). www.delsaglobal.org.

Data-Enabled Life Sciences Alliance. (2013b). Multi-omics metadata checklist. http://www.delsaglobal.org/news/publications/item/84-checklist; Last accessed October 23, 2013.

Editorial. (2103). Announcement: Reducing our irreproducibility. Nature 496, 398.

Editorial. (2011). On the table. Nature Genetics 43, 1.

Dryad. (2013). datadryad.org; Last accessed October 23, 2013.

eLife. (2013). elife.elifesciences.org; Last accessed October 23, 2013.

Field D, Sansone SA, Collis A, et al. (2009). Omics Data Sharing. Science 326, 234–236.

Higdon R, Stewart E, Stanberry L, et al. (2013). MOPED enables discoveries through consistently processed proteomics data. J Proteome Research 13, 107–113.

ISA tools. (2013). isa-tools.org; Last accessed October 23, 2013.

Journal of Proteome Research. (2013). pubs.acs.org/journal/jprobs; Last accessed October 23, 2013.

Kettner C, Field D, Sansone SA, et al. (2010). Meeting Report from the Second ''Minimum Information for Biological and Biomedical Investigations'' (MIBBI) workshop. Stand Genomic Sci 3, 259–266.

Kolker E, Altinas I, Bourne P, et al. (2013). In praise of open research measures. Nature 498, 170.

Kolker E, Higdon R, Haynes W, et al. (2012a). MOPED: Model Organism Protein Expression Database. Nucleic Acids Res 40, D1093-D1099.

Kolker E and Stewart E. (2012b). Opinion: Data to Knowledge to Action: Introducing DELSA Global, a community initiative to connect experts, share data, and democratize science. The Scientist, April 18, 2012. http://www.the-scientist.com/?articles.view/articleNo/31985/title/Opinion–Data-to-Knowledge-to-Action/; Last accessed October 23, 2013.

Kolker E, Stewart E, and Ozdemir V. (2012c). DELSA Global for ''big data'' and the Bioeconomy: Catalyzing collective innovation. Indust Biotechnol 8, 176–178.

Kolker E, Stewart E, and Ozdemir V. (2012d). Opportunities and challenges for the life sciences community. OMICS 16, 138–147.

Mias G, and Snyder M. (2013). Personal genomes, quantitative dynamic omics and personalized medicine. Quant Biol I, 71–90.

Micropub. (2013). micropub.org; Last accessed October 23, 2013.

Nanopub. (2013). nanopub.org; Last accessed October 23, 2013.

OMICS Journal of Integrative Biology. (2013). www.liebertpub.com/OMI; Last accessed October 23, 2013.

Open Science Data Cloud. (2013). www.opensciencedatacloud.org; Last accessed October 23, 2013.

ProteomeXchange. (2013). proteomexchange.org; Last accessed October 23, 2013.

PubMed commons. (2013). www.ncbi.nlm.nih.gov/pubmed commons; Last accessed October 23, 2013.

Reporting Checklist for Life Sciences Articles. (2013). Nature Publishing Group, May 2013. http://www.nature.com/authors/policies/checklist.pdf; Last accessed on August 1, 2013.

Rustici G, Kalesnikov N, Brandizi M, et al. (2013). ArrayExpress update – Trends in database growth and links to data analysis tools. Nucleic Acids Res 41, 987–990.

Scientific Data. (2013). www.nature.com/scientificdata; Last accessed October 13, 2013.

Smarr L. (2012). Quantifying your body: A how-to guide from a systems biology perspective. Biotechnol J 7, 980–991.

Snyder M, Mias G, Stanberry L, et al. (2014). Metadata checklist for the integrated personal omics study: proteomics and metabolomics experiments. Published in parallel: OMICS, 18 pp-pp ; Big Data 1 pp-pp.

Stewart E, Smith T, De Souza A, et al (2013). DELSA Workshop IV: Launching the Quantified Human Initiative. Big Data 1, 187–190.

Swartz A. (2013). Post-publication peer review mainstreamed. The Scientist. www.the-scientist.com/?articles.view/articleNo/37969/title/Post-Publication-Peer-Review-Mainstreamed/; Last accessed October 23, 2013

Taylor C, Field D, Sansone SA, et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 8, 889–96.

Taylor CF, Paton NW, Lilley KS, et al. (2007). The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25, 887–893.

Tennenbaum JD, Sansone SA, and Haendel M. (2013). A sea of standards for omics data: sink or swim? J Am Med Inform Assoc http://jamia.bmj.com/content/early/2013/09/27/amiajnl-2013-002066.full.html; Last accessed on October 17, 2013.

Vizcaino JA, Cote RG, Csordas A, et al. (2013). The Poteomics IDEntifications (PRIDE) database and associated tools: Status in 2013. Nucleic Acids Res 41, D1063–1069.

Yilmaz P, Kottmann R, Field D, et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnol 29, 415–420.

Address correspondence to:
Eugene Kolker
Bioinformatics and High-Throughput Analysis Laboratory
Seattle Children's Research Institute
Predictive Analytics
Seattle Children's Hospital
1900 Ninth Avenue
Seattle, WA 98101

E-mail: eugene.kolker@seattlechildrens.org