 DOI 10.1002/prca.201100097 *Proteomics Clin. Appl.* 2011, *5*, 580–589

STANDARDISATION & GUIDELINES

Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles)

Christopher R. Kinsinger[1], James Apffel[2], Mark Baker[3], Xiaopeng Bian[4], Christoph H. Borchers[5], Ralph Bradshaw[6], Mi-Youn Brusniak[7], Daniel W. Chan[8], Eric W. Deutsch[9], Bruno Domon[10], Jeff Gorman[11], Rudolf Grimm[12], William Hancock[13], Henning Hermjakob[14], David Horn[15], Christie Hunter[16], Patrik Kolar[17], Hans-Joachim Kraus[18], Hanno Langen[19], Rune Linding[20], Robert L. Moritz[21], Gilbert S. Omenn[22], Ron Orlando[23], Akhilesh Pandey[24], Peipei Ping[25], Amir Rahbar[1], Robert Rivers[16], Sean L. Seymour[26], Richard J. Simpson[27], Douglas Slotta[28], Richard D. Smith[29], Stephen E. Stein[30], David L. Tabb[31], Danilo Tagle[32], John R. Yates III[33] and Henry Rodriguez[1]

[1] Office of Cancer Clinical Proteomics Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA*

Received: October 27, 2011
Accepted: October 27, 2011

Policies supporting the rapid and open sharing of proteomic data are being implemented by the leading journals in the field. The proteomics community is taking steps to ensure that data are made publicly accessible and are of high quality, a challenging task that requires the development and deployment of methods for measuring and documenting data quality metrics. On September 18, 2010, the U.S. National Cancer Institute (NCI) convened the "International Workshop on Proteomic Data Quality Metrics" in Sydney, Australia, to identify and address issues facing the development and use of such methods for open access proteomics data. The stakeholders at the workshop enumerated the key principles underlying a framework for data quality assessment in mass spectrometry data that will meet the needs of the research community, journals, funding agencies, and data repositories. Attendees discussed and agreed up on two primary needs for the wide use of quality metrics: (i) an evolving list of comprehensive quality metrics and (ii) standards accompanied by software analytics. Attendees stressed the importance of increased education and training programs to promote reliable protocols in proteomics. This workshop report explores the historic precedents, key discussions, and necessary next steps to enhance the quality of open access data. By agreement, this article is published simultaneously in *Proteomics, Proteomics Clinical Applications, Journal of Proteome Research, and Molecular and Cellular Proteomics,* as a public service to the research community. The peer review process was a coordinated effort conducted by a panel of referees selected by the journals.

Keywords:
Amsterdam principles / Bioinformatics / Data quality / Metrics / Open access / Selected reaction monitoring / Standards

Correspondence: Dr. Christopher R. Kinsinger, Office of Cancer Clinical Proteomics Research; National Cancer Institute; National Institutes of Health; 31 Center Drive, MSC 2580; Bethesda, Maryland 20892, USA
E-mail: kinsingc@mail.nih.gov
Fax: +1-301-496-7807

*The remaining author affiliations are provided in the Addendum.

1 Introduction

On September 18, 2010, members of the international proteomics community met for a one-day workshop in Sydney, Australia, that was convened by the National Cancer Institute (NCI) of the U.S. National Institutes of Health (NIH). This workshop was held to address the lack of widely implementable policies governing the adoption and use of quality metrics for open access proteomic data, particularly concerning tandem mass spectrometry data used to identify and quantify proteins based on fragmentation of peptide ions. Parallel efforts can be implemented for protein capture-based data.

Clear data policy has aided the advance of other data-driven fields of research. Data release policies for DNA sequencing have enabled widespread access to data from the Human Genome Project and other large-scale sequencing efforts [1, 2]. The development of Minimum Information About a Microarray Experiment (MIAME) defined reporting standards for microarray experiments [3]. Researchers, funding agencies, and journals would all benefit from data policy clarification in proteomics as well.

This workshop was intended to further a process that, for proteomics, began in 2004. Editors from the journal, *Molecular and Cellular Proteomics* (MCP), recognized the need for manuscript submission guidelines to assist reviewers and readers of the paper in understanding the methods by which the authors acquired, processed, and analyzed their data [4]. This was followed by an expanded set of guidelines developed by an invited group of participants that met in Paris, France in 2005 [5, 6]. These guidelines delineate elements required in manuscripts detailing LC-MS/MS proteomic inventories (such as assignment of peptides, assessment of false-positive rate, and inference of proteins) and other types of experiments (such as quantitative measurements, post-translational modification assignments, and peptide-mass fingerprinting). Soon after the journal *MCP* adopted these "Paris Guidelines" as its standard, other journals began to adopt variants of these guidelines [7]. A 2009 Philadelphia workshop revisited these guidelines to require, for the first time, deposition of raw data in a public repository. (As of this writing, this requirement has been temporarily suspended due to technical difficulties associated with the presently available repositories. (http://www.mcponline.org/site/home/news/index.xhtml#rawdata)) In 2008, MCP developed the first set of guidelines for clinical proteomics papers to establish a baseline of credibility for articles in this field [8]. Concomitantly, the Human Proteome Organization's Proteomics Standards Initiative (HUPO-PSI) developed guidance for reporting data and metadata (MIAPE) and sought to standardize data formats for mass spectrometry data (mzML) [9, 10]. These efforts provide a solid foundation for advancing proteomics data policy.

To complement the activities of MCP and HUPO-PSI, the NCI sponsored a 2008 International Summit at which the proteomics community produced the six Amsterdam Principles to promote open access to proteomic data [11], addressing issues of: timing, comprehensiveness, format, deposition to repositories, quality metrics, and responsibility for proteomic data release. MCP has defined policies for its published articles for the principles involving timing, deposition to repositories, and responsibility for proteomic data release. MIAPE and mzML provide solid frameworks for advancing the principles of comprehensiveness and format. The principle for quality metrics, however, has not yet been matched with community-derived, comprehensive guidance. The Amsterdam Principles state that ''central repositories should develop threshold metrics for assessing data quality. These metrics should be developed in a coordinated manner, both with the research community and among each other, so as to ensure interoperability. As data become shared through such repositories, their value will become obvious to the community, and momentum will grow to sustain data release.''

The NCI invited key thought-leaders and stakeholders in proteomics to the Sydney Workshop, including data producers and users, managers of database repositories, editors of scientific journals, and representatives of funding agencies. These stakeholders were asked, "*How good is good enough, when it comes to MS-based proteomic data?*" The answers to this question would form the quality criteria for shared data, benefiting both producers and users of proteomic data.

The workshop focused on three use cases: (i) users of public proteomic data, (ii) reviewers of journal articles, and (iii) multi-site, data production projects in which unpublished data are shared among laboratories. In the first case, public proteomic data may be used for a variety of applications, covering a range of prerequisites for data quality and metadata. What level of annotation and which quality metrics would help a potential user quickly determine whether a data set is appropriate for the intended application? In the second case, reviewers need to determine the quality of data that accompany a submitted manuscript as a means to ascertain the strength by which the data support experimental findings in the manuscript. Apart from repeating the full analysis of the data, what set of metrics would facilitate the task of reviewers? Third, team science in the context of systems biology presents opportunities and challenges for integrating data sets generated by different laboratories with distinct quality assurance protocols, often on different analysis platforms. Participants in the workshop stated that standards and metrics must not be oppressive, stifling innovation and methods development and potentially hindering the evolution of proteomic technologies. However, for data produced to answer a biological question rather than to demonstrate technology, what metrics would bolster confidence that data are of sufficient quality to be integrated with experimental results spanning multiple laboratories and platforms?

Data from proteomics experiments can assume a variety of forms, depending on the experimental design. Evaluating data set quality is heavily dependent upon accompanying

annotation. Metadata relate the *experimental design* to produced data, making clear which files or records are associated with a particular sample. *Protocols* may be documented via Standard Operating Procedures, but most data sets are presently dependent on methods explained briefly or by reference in associated publications. *Instrument data* are generally considered the core of public data sets, but the format in which they are published may have a considerable impact on their usability. *Derived information* from these data may include raw peptide identifications or transition peak areas. Often this level is bypassed, with a more heavily processed protein spectral count table or protein ratio table presented with minimal explanation of software tools used to move from instrument data to derived information. Finally, *experimental findings* reflect the high-level analysis motivated by the experimental design. A list of differentially expressed proteins, enriched networks, or biomarker candidates connects to derived information through higher level analyses conducted in statistical or systems biology tools.

This meeting report discusses the basic principles underlying data quality metrics in proteomics and the challenges to develop and implement these principles. Its purpose is to serve as a focus of discussion and action that will guide the continuing commitment of the field to conduct high-quality research and to make the product of that research useful to the global research community.

2 Issues and challenges

2.1 Data access and annotation

One of the major conclusions from the Amsterdam Summit was that data sharing among members of the proteomics community should be required. While there is widespread consensus on the importance of making proteomics data publicly available, the full infrastructure to do so does not yet exist. ProteomeCommons/Tranche [12] and ProteomeXchange [13] have been developed to meet key needs, but do not yet constitute the comprehensive system required to fully support the field. Participants at the Sydney Workshop called for a sustainably funded repository or set of linked repositories that would work closely with journals to ensure long-term access to proteomics data.

A major problem confronting the reuse of published data files is the lack of information concerning the experiment that led to data generation. Even when experimental annotation is available, it is often cryptic or missing key information, inviting errors in mapping data to results. Contacting authors is often necessary to seek clarification.

Given public access and annotation for a data set, users should next be able to assess its quality, preferably prior to download. Data quality has become more important due to rapid evolution of technologies as well as the growing size and quantity of repository data sets. Given the option of several potential data sets, users should be able to compare among them on a common basis and evaluate their comparability or compatibility.

With major repositories joining forces to create ProteomeXchange [13], quality assurance and quality control of public data sets will be necessary to separate what is useful from the rest. Moreover, research societies and funding agencies are placing importance on proteomic technologies and findings with the launch of the HUPO Human Proteome Project [14], the NCI-sponsored Clinical Proteomic Technologies for Cancer (CPTC) Initiative [15], and other related initiatives, raising the importance of both data access and quality. Data quality standards will enable tools for tracking samples, measuring analytical variation, flagging incomplete data sets, and protecting against experimental misinterpretation.

2.2 Peptide to spectra matching

While the assembly of proteome inventories was the principal driver for technologies prior to the year 2000 and continues to be important as much more powerful instruments are deployed, the most recent decade has seen the emergence of quantitative proteomics as a complementary discipline. While both sets of technologies depend upon the dissociation of peptides into fragment ions, discovery technologies generally scan the fragments produced from sampled peptide ions while targeted strategies quantify fragment ions from specified sets of precursor ions. The ion-chromatograms of these fragment intensities comprise the raw output of selected reaction monitoring (SRM, also commonly referred to as multiple reaction monitoring or MRM) experiments just as tandem mass spectra are the raw output of sampling dominant peptide precursor ions.

Discovery platforms generate tandem mass spectra from selected peptides in complex mixtures. These tandem mass spectra are then matched to lists of fragment ions predicted from sequences in reference protein databases or by direct match to qualified spectral databases. Once identifications have been filtered to an acceptably stringent error rate [16, 17], the confident set of peptides can be assembled to derive protein-level information. Tables of spectral counts for proteins observed across several samples can be used to recognize sets of proteins that vary in concentration level between cohorts of samples.

Quality metrics can target any of several levels of data in such experiments:

(i) A single peptide-spectrum match (PSM) overlays a peptide sequence on a tandem mass spectrum, showing the correspondence of sequence-derived fragments with peaks observed in the MS/MS. The most straightforward quality assessment for an identification may be based in the scores produced from a search engine [18].

Posterior error probabilities, on the other hand, characterize the chance of error for a given PSM [19]. Alternatively, one may evaluate the quality of a spectrum without recourse to an associated sequence [20].

(ii) Evaluating a single LC-MS/MS analysis broadens the scope to thousands of tandem mass spectra. Evaluating error rates for the aggregate of identified peptides frequently takes the form of false discovery rates (FDR) or *q*-values [21]. Assessment of the underlying LC separation may, however, de-emphasize the identifications and focus on MS signals instead [22]. While the appropriate course must be determined by the intended application of the quality metrics, certain simple metrics can almost always be expected to be useful in characterizing an LC-MS/MS experiment. These include typical mass accuracies of both precursor and fragment ions, the amount of time over which the bulk of the peptides are eluted during the LC gradient, and the typical peak width of a peptide in the time domain.

(ii) At the level of the overall experiment, the stability of spectral counts among technical and biological replicates is significant [23]. Attempts to establish a set of ''biomarker'' candidate proteins require statistical models of differences [24], with each yielding a metric to characterize the degree of change observed. Experiments of relative quantitation using labeled tags, such as iTRAQ [25, 26] or SILAC have their own requirements for quality assessment. When multiple tags are assigned to each cohort, quality measurement may correlate quantities derived from redundant tags.

These are all considered relevant metrics of quality; however, their relevance depends on the proposed application and utilization of the data. When a bioinformatics researcher applies new algorithms to instrument data, he or she may begin by replacing previous identifications with new ones. For this application, metrics are needed to confirm the quality of underlying MS and MS/MS scans. However, a cancer researcher may wish to use a list of differentially expressed proteins from a data set and need a more comprehensive set of metrics to evaluate the quality of the initial data analysis.

2.3 Selected reaction monitoring

SRM mass spectrometry has recently been adopted in proteomics as a targeted method for detecting and quantifying peptides. Until recently, publications featuring SRM methods have been greatly outnumbered by those enumerating proteome inventories. However, the number of researchers and publications employing SRM-based methods continues to increase. As a result, establishing data quality metrics would help to legitimize a field that would otherwise encounter challenges similar to those faced by proteomic discovery.

In considering how to develop data quality metrics, the workshop participants developed several broad questions that the proteomics community needs to answer regarding the quality of SRM data, including the following:

(i) Are the technologies and software analytic tools available today sufficiently well developed to compute accurate quantitative measures, along with FDRs for SRM experiments? In SRM experiments, the analyzed ion-chromatograms may have a substantial probability of measuring an ion other than the intended target. This becomes more prevalent as lower abundance proteins are targeted and the background interference is stronger. Software packages, such as mProphet [27] and AuDIT [28], as well as instrument-specific software developed by vendors do provide some quality-check computations, but each of these packages has limitations and few are applicable across multiple instrument vendors. Furthermore, it may be difficult to compare results generated by different software packages without some standard metrics for data quality or standard data sets by which to assess the various software packages.

(ii) In the absence of robust probabilities, what minimal data and metadata should be provided in order for the reported results to be objectively assessed? Should original chromatograms in an open format such as mzML [10] be required for publication? Should an annotated depiction of all transitions used to make an identification be required for publication? Should attributes of the peaks, such as peak signal-to-noise and peak asymmetry, be reported? A key concern here will be finding the proper balance between the need for sufficient data and metadata to reproduce a given set of experiments and the desire to see all the data generated by those experiments. Another issue is that it may be too early to set stringent standards because of the rapid rate of technology and software innovation.

(iii) What are the minimal guidelines on how many peptides per protein and how many transitions per peptide are needed to ensure accurate quantitation? How should data be treated that are derived from a single peptide ion for the detection of a protein, even if only a single peptide transition is available for identification? The answer to these questions may lie with the availability of computational models that show promise in predicting how many peptides and transitions are needed to provide reliable quantitation.

(iv) How can standard formats, minimal information specifications, and data repositories help promote data quality, and where should these requirements be implemented? While there are standard formats for sharing SRM experiment data, such as the Proteomic Standards Initiative (PSI) mzML [10] format for experimental output formats and the PSI TraML [29] format for transition lists, there is still a need to develop a standard format for reporting chromatogram analysis results.

The consensus among the stakeholder participants concerned the need for defined quality metrics addressing the above questions. While over-standardization could prevent innovation, a lack of quality metrics and standardized annotation could cripple the comparison, portability, reproducibility, and ultimate adoption of proteomic findings and techniques.

3 Principles for data quality metrics

The Sydney Workshop agreed upon the following principles to develop useful and successful proteomic data quality metrics for open access data. These principles addressing the challenges outlined above include:

3.1 Metadata inclusion

3.1.1 Reproducibility

All proteomics data must be accompanied with appropriate metadata that describe the biological sample, experimental procedures, instrumentation, and any software or algorithms (including version numbers) used for post-processing. Scientific reproducibility should guide the reporting of metadata, such that recipients should be able to recapitulate the analysis reported in the publication. Assessing compliance to such guidelines remains a role of journal reviewers, since they are the gatekeepers to public data in conjunction with publications.

3.1.2 Formats

Preventing metadata requirements from becoming too onerous will require the continued adoption of standard data formats by instrument manufacturers, data repositories, and journals. In addition to mzML [10, 30] and TraML [29], participants at the workshop called for an additional format to report analysis of chromatographic data. HUPO PSI, HUPO New Technologies Committee, and other domain-specific working groups have the charge to continue to develop and improve such formats.

3.1.3 Repository

Standardizing the content presentation of metadata will require a sustainable repository for proteomics. It may be necessary to establish a specialized data repository, perhaps a variant of Tranche [12, 31, 32], PRIDE [33], GPMdb [34] or PeptideAtlas [35] for quantitative data produced by SRM experiments. The need for metadata is even greater for targeted experiments since peptide sequences cannot be inferred from limited numbers of transitions. Considerations for housing such a repository should take into account an organization's experience to develop and promote standards for databases, data deposition and exchange, and the sustainable funding model for maintenance of such biomedical databases.

3.2 Recognition for data production

Quality metrics should provide added value to both generators and users of proteomic data. For investigators, quality metrics can protect against publishing mistakes and provide validation of experimental procedures that could prove useful in seeking follow-on grants. Additionally, searchable Digital Object Identifiers (DOI) should be associated with data sets to allow for stable references over time. These links will allow investigators to reference data collections with high specificity in research papers and proposals. High-quality data may be featured by the community, as in the GPMdb ''Data set of the week'' [36], or by citations in papers, creating an incentive for authors to share data. Documentation of data use would provide a "value metric" for investigators, funding agencies, and data repositories.

For journal editors, metrics can enhance the quality and reliability of the manuscript review process. For repository managers, quality metrics can increase the value of stored data for use in metaanalyses and other retrospective data-mining operations. Investigators in large, multi-site studies would have added confidence that data generated at one site can and should be aggregated and compared with the data generated by other sites. As noted at the Amsterdam Summit, all parties producing, using, and funding proteomic data have an inherent responsibility to cooperate in ensuring both open access and high-quality data [11].

3.3 Reference materials and reference data for benchmarking

Reference materials and reference data are necessary for benchmarking and comparing methods in proteomics. These methods include both physical measurement platforms and analysis software. Besides providing a benchmarking process for method development, the availability of reference materials and data quality software would aid the interpretation of biological measurements by helping to answer the question, ''*Are the results due to noise in the measurement or sample preparation processes or due to real biological signal?*''

Reference materials and software metrics have already played a role in standardizing LC-MS systems in the field of shotgun proteomics, including the UPS1 defined protein mix developed by Sigma through the 2006 sPRG study from ABRF and a yeast lysate developed by NCI's Clinical Proteomic Technology Assessment for Cancer and NIST, with accompanying data set [22, 37, 38]. Other groups of researchers have developed workflows for greater reprodu-

cibility in SRM-MS experiments [39]. Another lab has begun to develop the SRM Atlas which contains highly qualified spectral libraries of proteotypic peptide data of numerous organisms [40]. In terms of reference data sets, the ISB18 data set [41] and the Aurum data set for MALDI-TOF/TOF [42] were both created for the purpose of benchmarking new software tools. The ABRF Proteome Informatics Research Group (iPRG) has released two data sets specifically designed for benchmarking subtractive analysis and phosphopeptide identification [43, 44]. The HUPO Plasma Proteome Project released large volumes of data that demonstrated performance for a wide variety of instruments and protocols [45]. To provide a means of comparison that exceeds a basic metric such as the number of peptides identified, software that provides metrics of the measurement process, such as NISTMSQC [22], must be accessible. Other resources are available that identify low-quality peptide identifications, providing a quantitative assessment of data quality, such as: Census [46], Colander [47], Debunker [48], DTASelect 2 [18], IDPicker [49], PeptideProphet [50], Percolator [16, 51], and ProteinProphet [52].

The workshop did not call for adoption of a standard method or platform for every experiment but rather recognized the need for formal comparison of methods on equal footing. With no shortage of reference materials and tools for evaluation of data quality, it falls to journal editors and reviewers to ensure that published data were produced using methods assessed by appropriate reference materials and data.

3.4 Education and training

The proteomics community needs to improve education and training in order to improve the quality of published methods and data. The field should develop tutorials at scientific meetings, webinars, and vendor-sponsored hands-on workshops for students and postdoctoral fellows to ensure that instruments are used correctly. These tutorials might address topics such as, ''LOD and LOQ determination for SRM,'' ''A journal's (or repository's) guide on how to properly release a proteomic data set,'' ''Protecting against false positives in PTM identification,'' and ''Quality control in proteomics.'' Senior academic investigators in the field should establish proteomics courses at their institutions. Journals likewise can present timely tutorials and have a critical educational role through enforcement of their guidelines. The American Society of Mass Spectrometry (ASMS), the Association of Biomolecular Resource Facilities (ABRF), and the U.S. Human Proteome Organization (US HUPO) already present a number of excellent tutorials in conjunction with their meetings. In particular, ABRF Research Groups regularly conduct community-wide experiments to facilitate method comparison [38]. The Sydney Workshop called for advancing this effort by formalizing method comparison and publishing the results.

Additionally, proteome informatics needs a community resource (such as a wiki) in which software developers and expert users can communicate standard workflows for

Table 1. Recommendations for enhancing quality metrics for proteomic data

Metadata	
General	Guidelines for metadata requirements should be based on enabling scientific reproducibility and properly evaluating of the analysis protocol
Targeted quantitation	Standard formats need to be developed for reporting transitions and results of chromatogram analyses
	Minimum information guidelines need to be developed to ensure that sufficient information is reported about the analysis protocol so that the results may be properly evaluated
Reference materials and data	
	Reference materials and reference data should allow for formal benchmarking of new and existing methods
Quality metrics	
Protein survey	Fundamentally, quality metrics should confirm the quality of each underlying MS and MS/MS scan. For a data set of multiple spectra, quality metrics are needed to evaluate the quality of the initial data analysis.
Targeted quantitation	Quality metrics are needed to convey the confidence that peak measurements represent the abundance of the target and are not distorted by interferences.
Education	
	Journals, trade associations, and academic institutions have a responsibility to see that education in proteomics includes training in QC of proteomics measurements.
	Proteome informatics needs a community resource (such as a wiki) in which software developers and expert users can communicate standard workflows for optimal performance
Data deposition	
	Proteomics is in need of a sustainable repository that would work closely with journals to ensure long-term access to proteomics data.
	Formal recognition of data deposition with DOI is a key incentive to advance access of proteomic data.

optimal performance. Open public discussions on strategies for selecting a FASTA database for database search or choosing a ''semi-tryptic'' versus a ''fully-tryptic'' search will help reduce the competing factions within this field. Conducting comparisons of old and new revisions of software or comparing among different tools will help to characterize inevitable software change.

4 Concluding remarks and next steps

In summary, the Sydney Workshop identified challenges and principles in the following areas (Table 1)

Addressing and implementing these will require the combined efforts of journal editors, funding agencies, trade associations, instrument manufacturers, software developers, and the community of researchers. With regard to metadata, a group of workshop participants agreed to begin drafting guidelines for SRM experiments. This group, along with the HUPO PSI and the HUPO New Technologies Committee, are working together on a draft set of guidelines for evaluation by the community. The HUPO PSI is also developing the XML-based mzQuantML format that is intended to capture the metadata and results of quantitative proteomics assays, including SRM assays. There is no shortage of reference materials; however, journals, funding agencies, and investigators have a shared responsibility to ensure their routine usage and comparisons. Use of reference materials may only increase as the field advances toward commercial applications subject to regulatory agencies, such as the FDA. The quality metrics themselves will be developed in individual labs, but software developers and instrument manufacturers who incorporate them into their products will offer a competitive advantage to the researchers who use them. Additionally, journals have a role to continue to develop and enforce policies defining data quality for the data in manuscripts for publication. As such, journal reviewers and editorial boards are the ultimate gatekeepers of data quality. Education about quality metrics will be fostered by journals, trade associations, and academic institutions as they promote proper usage of quality metrics. Finally, for funding agencies to support a proteomics data repository, all members of the proteomics community must convey the value of proteomics research to key stakeholders who would benefit from funding such a repository.

Over the past decade, the research community has made remarkable progress in developing the technologies and experimental protocols needed to make proteomics a viable science. The technologies are now poised to build on the successes of the genomics community in furthering our understanding of diseases at the molecular level. Like the genomics field before it, the proteomics field has now established sensible guidelines, increasingly adopted by its practitioners, which encourage easy public access to data. The Sydney Workshop provided questions and frameworks to develop the necessary metrics and conditions to ensure the quality of proteomics data. The successful implementation of quality metrics will require processes to protect against errors in data interpretation, incorporate ''complete'' metadata with each data set, and reward the teams that deposit data. Much of the burden for data quality will fall on bioinformatics researchers tasked with the development and adoption of long-term and scalable data storage solutions, metric formulation, and comparative testing strategies.

Key to the development of these data quality metrics is an understanding that over-standardization of the still evolving field of mass-spectrometry-based proteomics could hinder innovation and potential discovery. However, it is clear that basic standardization and the application of quality controls will be necessary to ensure reproducibility, reliability, and reuse of data. It is now imperative for the field to develop the metrics and methodologies to ensure that data are of the highest quality without also impeding research and innovation.

While challenges remain in defining the policies and metrics needed to assess data quality, the Sydney Workshop reaffirmed that the proteomics community has a clear interest in addressing these questions among funding agencies, journals, standards working groups, international societies, data repositories, and above all the research community. The Sydney Workshop represents an initial step toward the inclusion of more detailed metadata, the adoption of reference materials and data, the development of data quality metrics for method comparison and quality control, the education of users in the areas of proteomics and bioinformatics, and the recognition of data depositors.

The authors have declared no conflict of interest.

5 References

[1] Policies on Release of Human Genomic Sequence Data. US Department of Energy Human Genome Project, Washington.

[2] Sharing Data from Large-sclae Biological Research Projects: A System of Tripartite Responsibility. Wellcome Trust, London 2003.

[3] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G. et al., Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 2001, *29*, 365–371.

[4] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. et al., The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[5] Bradshaw, R. A., Revised draft guidelines for proteomic data publication. *Mol. Cell. Proteomics* 2005, *4*, 1223–1225.

[6] Bradshaw, R. A., Burlingame, A. L., Carr, S., Aebersold, R., Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics* 2006, *5*, 787–788.

[7] Wilkins, M. R., Appel, R. D., Van Eyk, J. E., Chung, M. C. M. et al., Guidelines for the next 10 years of proteomics. *Proteomics* 2006, *6*, 4–8.

[8] Celis, J. E., Carr, S. A., Bradshaw, R. A., New guidelines for clinical proteomics manuscripts. *Mol. Cell. Proteomics* 2008, *7*, 2071–2072.

[9] Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A. et al., The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 2007, *25*, 887–893.

[10] Martens, L., Chambers, M., Sturm, M., Kessner, D. et al., mzML – a community standard for mass spectrometry data. *Mol. Cell. Proteomics* 2011, *10*, R110 000133.

[11] Rodriguez, H., Snyder, M., Uhlén, M., Andrews, P. et al., Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles. *J. Proteome Res.* 2009, *8*, 3689–3692.

[12] Hill, J. A., Smith, B. E., Papoulias, P. G., Andrews, P. C., ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository. *J. Proteome Res.* 2010, *9*, 2809–2811.

[13] Hermjakob, H., Apweiler, R., The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev. Proteomics* 2006, *3*, 1–3.

[14] Legrain, P., Aebersold, R., Archakov, A., Bairoch, A. et al., The human proteome project: current state and future direction. *Mol. Cell. Proteomics* 2011, *10*, M111.009993.

[15] Boja, E., Hiltke, T., Rivers, R., Kinsinger, C. et al., Evolution of clinical proteomics and its role in medicine. *J. Proteome Res.* 2011, *10*, 66–84.

[16] Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* 2007, *4*, 923–925.

[17] Farrah, T., Deutsch, E. W., Omenn, G. S., Campbell, D. S. et al., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* 2011.

[18] Tabb, D. L., McDonald, W. H., Yates, J. R., 3rd DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 2002, *1*, 21–26.

[19] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, *74*, 5383–5392.

[20] Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M. et al., Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* 2006, *5*, 652–670.

[21] Kall, L., Storey, J. D., MacCoss, M. J., Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* 2008, *7*, 29–34.

[22] Rudnick, P. A., Clauser, K. R., Kilpatrick, L. E., Tchekhovskoi, D. V. et al., Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics* 2010, *9*, 225–241.

[23] Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M. et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* 2010, *9*, 761–776.

[24] Whiteaker, J. R., Zhang, H., Zhao, L., Wang, P. et al., Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J. Proteome Res.* 2007, *6*, 3962–3975.

[25] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B. et al., Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 2004, *3*, 1154–1169.

[26] Liu, H., Sadygov, R. G., Yates, J. R., A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[27] Reiter, L., Rinner, O., Picotti, P., Huttenhain, R. et al., mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* 2011, *8*, 430–435.

[28] Abbatiello, S. E., Mani, D. R., Keshishian, H., Carr, S. A., Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. *Clin. Chem.* 2009, *56*, 291–305.

[29] Orchard, S., Albar, J. P., Deutsch, E. W., Eisenacher, M. et al., implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada. *Proteomics* 2010, *10*, 1895–1898.

[30] Deutsch, E., mzML: A single, unifying data format for mass spectrometer output. *Proteomics* 2008, *8*, 2776–2777.

[31] Falkner, J. A., Andrews, P. C., Tranche: secure decentralized data storage for the proteomics community. *J. Biomol. Tech.* 2007, *18*, 3.

[32] Falkner, J. A., Hill, J. A., Andrews, P. C., Proteomics FASTA archive and reference resource. *Proteomics* 2008, *8*, 1756–1757.

[33] Vizcaino, J. A., Cote, R., Reisinger, F., Barsnes, H. et al., The Proteomics Identifications database: 2010 update. *Nucleic Acids Res.* 2010, *38*, D736–742.

[34] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* 2004, *3*, 1234–1242.

[35] Deutsch, E. W., Lam, H., Aebersold, R., PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* 2008, *9*, 429–434.

[36] Beavis, R., The GPM Data Set of the Week, 2011. The GPMDB contains tens of thousands of data sets contributed by researchers around the world. Every week, a data set is selected because of its technical excellence, biological interest or simply because of its general interest to the proteomics community, 2011.

[37] Paulovich, A. G., Billheimer, D., Ham, A. J., Vega-Montoto, L. et al., Interlaboratory study characterizing a yeast

performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* 2010, *9*, 242–254.

[38] Turck, C. W., Falick, A. M., Kowalak, J. A., Lane, W. S. et al., The association of biomolecular resource facilities proteomics research group 2006 study: relative protein quantitation. *Mol. Cell. Proteomics* 2007, *6*, 1291–1298.

[39] Addona, T. A., Abbatiello, S. E., Schilling, B., Skates, S. J. et al., Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* 2009, *27*, 633–641.

[40] Picotti, P., Rinner, O., Stallmach, R., Dautel, F. et al., High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* 2010, *7*, 43–46.

[41] Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J. et al., The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* 2007, *7*, 96–103.

[42] Falkner, J., Kachman, M., Veine, D., Walker, A. et al., Validated MALDI-TOF/TOF mass spectra for protein standards. *J. Am. Soc. Mass Spectrom.* 2007, *18*, 850–855.

[43] Askenazi, M., Falkner, J., Kowalak, J. A., Lane, W. S. et al., *ABRF iPRG 2009 E. coli Subtractive Analysis*. In: Tabb, D. L., (ed.), Proteome Commons, Nashville, TN 2009.

[44] Askenazi, M., Clauser, K. R., Martens, L., McDonald, W. H. et al., *Study Materials for Phosphopeptide Identification*. In: Clauser, K., (ed.), Proteome Commons, Ann Arbor, MI 2010.

[45] Omenn, G. S., The human proteome organization plasma proteome project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. *Proteomics* 2004, *4*, 1235–1240.

[46] Park, S. K., Yates, J. R., 3rd, Census for proteome quantification. *Curr. Protoc. Bioinformatics Chapter* 2010, *13*, Unit 13 12 11-11.

[47] Lu, B., Ruse, C. I., Yates, J. R., 3rd, Colander: A probability-based support vector machine algorithm for automatic screening for CID spectra of phosphopeptides prior to database search. *J. Proteome Res.* 2008, *7*, 3628–3634.

[48] Lu, B., Ruse, C., Xu, T., Park, S. K., Yates, J., 3rd, Automatic validation of phosphopeptide identifications from tandem mass spectra. *Anal. Chem.* 2007, *79*, 1301–1310.

[49] Ma, Z. Q., Dasari, S., Chambers, M. C., Litton, M. D. et al., IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* 2009, *8*, 3872–3881.

[50] Keller, A., Nesvizhskii, A., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, *74*, 5383.

[51] Klammer, A. A., Park, C. Y., Noble, W. S., Statistical Calibration of the SEQUEST XCorr Function. *J. Proteome Res.* 2009, *8*, 2106–2113.

[52] Nesvizhskii, A., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646.

6 Addendum

[2] Agilent Research Laboratories, Santa Clara, CA, USA
[3] Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia
[4] Center for Bioinformatics and Information Technology, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[5] Genome BC Proteomics Centre, University of Victoria, Victoria BC, Canada
[6] Mass Spectrometry Facility, University of California, San Francisco, CA, USA
[7] Institute of Systems Biology, Seattle, WA, USA
[8] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[9] Institute for Systems Biology, Seattle, WA, USA
[10] Luxembourg Clinical Proteomics Center, CRP-Sante, Strassen, Luxembourg, Europe
[11] Protein Discovery Centre, Queensland Institute of Medical Research, Herston, Queensland, Australia
[12] Agilent Technologies, Santa Clara, CA, USA
[13] Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA
[14] Proteomics Services, European Bioinformatics Institute, Cambridge, UK
[15] Proteomics Software Strategic Marketing, Thermo Fisher Scientific, San Jose, CA, USA
[16] AB SCIEX, Foster City, CA, USA
[17] Directorate-General for Research, European Commission, Brussels, Belgium
[18] Wiley-VCH, Weinheim, Germany
[19] Exploratory Biomarkers, Hoffmann-La Roche, Basel, Switzerland
[20] The Technical University of Denmark (DTU), Cellular Signal Integration Group (C-SIG), Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Lyngby, Denmark
[21] Cellular and Molecular Logic Unit, Institute of Systems Biology, Seattle, WA, USA
[22] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[23] Complex Carbohydrate Research Center, University of Georgia, Athens, GA, USA

[24] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA
[25] David Geffen School of Medicine, University of California, Los Angeles, CA, USA
[26] Small Business Development Center, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[27] La Trobe institute for Molecular Science, L Trobe University, Bundoora, Victoria, Australia
[28] Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA
[29] Pacific Northwest National Laboratory, Richland, WA, USA
[30] Chemical Reference Data Group, National Institute of Standards and Technology, Gaithersburg, MD, USA
[31] Vanderbilt-Ingram Cancer Center, Nashville, TN, USA
[32] National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
[33] The Scripps Research Institute, The Scripps Research Institute, La Jolla, CA, USA

