Published in final edited form as:

Clin Biochem. 2013 April ; 46(6): 411–420. doi:10.1016/j.clinbiochem.2012.12.003.

Quality Assessment for Clinical Proteomics

David L. Tabb

Department of Biomedical Informatics, Vanderbilt University, (615)936-0380 voice, (615) 343-8372 fax, U9211 Learned Lab/MRB III, 465 21stAve S, Nashville, TN 37232-8575

David L. Tabb: david.l.tabb@vanderbilt.edu

Abstract

Proteomics has emerged from the labs of technologists to enter widespread application in clinical contexts. This transition, however, has been hindered by overstated early claims of accuracy, concerns about reproducibility, and the challenges of handling batch effects properly. New efforts have produced sets of performance metrics and measurements of variability that establish sound expectations for experiments in clinical proteomics. As researchers begin incorporating these metrics in a quality by design paradigm, the variability of individual steps in experimental pipelines will be reduced, regularizing overall outcomes. This review discusses the evolution of quality assessment in 2D gel electrophoresis, mass spectrometry-based proteomic profiling, tandem mass spectrometry-based protein inventories, and proteomic quantitation. Taken together, the advances in each of these technologies are establishing databases that will be increasingly useful for decision-making in clinical experimentation.

Keywords

Quality Assessment; Quality Control; Quality by Design; Repeatability; Reproducibility; Proteomics; Selected Reaction Monitoring; Mass Spectrometry Profiling; Polyacrylamide Gel Electrophoresis

Objectives for proteomic quality control

During the 1990s and 2000s, proteomics technologies were continuously in flux, with new instrument methods and separation techniques coming to the fore with each publication. Since contemporary technologies can yield considerable volumes of information, the emphasis in proteomics has begun shifting to applications, while further technology development continues in specialist laboratories. Growing numbers of researchers have turned to proteomics for clinical research, particularly in the area of biomarkers. The complexity of clinical proteomics has led to research in quality control to ensure that findings from these experiments are meaningful.

The chief traits sought through quality control include repeatability and reproducibility. The IUPAC Gold Book defines repeatability as “the closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of

© 2012 The Canadian Society of Clinical Chemists. Published by Elsevier Inc. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

time)” (1). In practice, repeatability describes the variability contributed by the analytical technique among technical replicates. Reproducibility measures “the closeness of agreement between independent results obtained with the same method on identical test material but under different conditions (different operators, different apparatus, different laboratories and/or after different intervals of time).” Reproducibility has posed significant challenges in the field of proteomics, in part because each laboratory uses subtly different instrumentation, protocols, and reagents.

Like many industries, pharmaceutical development has begun shifting to a “quality by design” paradigm (2). Instead of monitoring manufactured pharmaceuticals for those which fail to pass criteria, quality by design means “designing and developing formulations and manufacturing processes to ensure predefined product quality” (3). The FDA recommended quality by design as part of pharmaceutical development in their Center for Drug Evaluation and Research in 2006. These recommendations have begun impacting the field of proteomics through laboratories involved in the production of therapeutic monoclonal antibodies (4).

The inability to reproduce results for molecular assays has been cited as a principal point of weakness in applying biotechnology for clinical biomarker identification (5). Ransohoff argued that rules of evidence for assessing molecular marker validity are poorly established, and many studies emphasize marker discovery to a far greater extent than validation in independent sets. Even if instrumentation is operating with a minimum of variance, these designs are subject to considerable overfitting error.

Discovering biomarkers, then, can also benefit from quality by design, with an emphasis on incorporating all necessary experiments with appropriate statistical power and on minimizing variability in individual technologies. McGuire investigated an idealized workflow for biomarker discovery, building from a relevant clinical question to a planning phase, discovery experiments, data analysis and replication in independent data (6). A 2007 review by Karp and Lilley introduced the determinants of good experimental design, explaining the need for biological and technical replicates, formal hypothesis testing, multiple testing correction, and orthogonal validation (7).

A viewpoint spanning 26 laboratories in the second issue of *Proteomics: Clinical Applications* sought to define this field and set standards for its practice (8). Mischak delineated measurements to control quality for experimentation, specifying that researchers need to characterize intra-assay variation, inter-assay variation, and consistency of results across different sample concentrations. They targeted these evaluations as a way to comply with the Good Clinical Laboratory Practice (GCLP) standard, which was first published by the British Association of Research Quality Assurance (BARQA) to bridge Good Laboratory Practice and International Conference on Harmonization recommendations and then expanded by the National Institute of Allergy and Infectious Diseases to incorporate Clinical Laboratory Improvement Amendments (CLIA) and International Organization for Standardization (ISO) content (9). The expanded Good Clinical Laboratory Practice standard¹ incorporates quality control content ubiquitously, both within the 40-page standard definition and in the 60-page appendices guiding implementation of the standard.

A yawning gulf separates the documentation and design of GCLP experiments from the typical practice of proteomics. Lack of reproducibility has vexed studies that span many laboratories. Meta-analysis of protein and peptide lists from the HUPO Plasma Proteome Project, for example, found frustratingly low overlap among protein identifications from

¹<http://www.niaid.nih.gov/LabsAndResources/resources/DAIDSClinRsrch/Pages/Laboratories.aspx>

different laboratories, though data from similar techniques clustered within these data (10). Early studies such as the 2004 Plasma Proteome Project generally made no effort to control variation among laboratories, allowing the use of disparate digestion strategies, pre-fractionation techniques, liquid chromatographic apparatus, mass spectrometry platforms, and bioinformatics. More recently, HUPO published the result of a 27-laboratory study to identify the components of a 20 protein mixture, finding that the groups differed considerably in the identifications produced from the mix (11).

Since 2002, the Proteomics Research Group (PRG) and its subgroups in the Association for Biomolecular Resource Facilities have issued yearly challenges to the proteomics community to evaluate the capacity of anonymous participants in fielding particular challenges (12). Their findings have shown a wide range of performance for tasks such as protein identification and quantitation. Their most recent challenge, for 2012, evaluates the performance of instruments in a quality control test sample over the span of nine months.

Mann argued in a 2009 commentary that the different protocols, instruments, and informatics employed by laboratories posed a challenge to comparing data sets among sites (13); correspondingly, the differences among proteomic inventories were extreme. The HUPO Plasma Proteome Project also encountered challenges due to the dynamic range of plasma. Their experience was important in guiding biomarker discovery away from blood; Lescuyer argued that biofluids are inappropriate for biomarker discovery due to their high dynamic range, finding that few studies have ever suggested that tissue leakage is discoverable from this sample type (14).

As was true for other biotechnologies, proteomics for cancer biomarkers suffered from exaggerated promises in its early stages. A 2002 *Lancet* article by Petricoin announced a SELDI-TOF profile-based proteomic pattern to recognize ovarian cancer from serum with high sensitivity and specificity (15). Subsequent examination by Baggerly revealed that random collections of peaks from these data replicated the reported signature, suggesting that cohort differences were almost entirely attributable to batch effects (16). Limited reproducibility led to skepticism on the part of funding agencies that proteomics could deliver clinically meaningful results. The National Cancer Institute launched the Clinical Proteomic Technology Assessment for Cancer (CPTAC) as an effort to audit the capabilities of these methods as a guide to further funding. This and companion efforts from the Human Proteomics Organization have generated a more nuanced appraisal of what proteomics can offer the clinic.

In 2005, Robert Gentleman argued that data and supporting software be provided as part of electronic publication of research articles to enable further exploration (17). Genomics adjusted at a relatively early stage to the release of sequence data to public repositories. Many proteomics investigators, however, resisted public data sharing at first. Martens and Hermjakob detailed the variability contributed by data analysis in proteomic inventories to explain the need for raw data deposition in repositories (18). Representatives of the National Cancer Institute and others stressed the need for public release of raw instrument data in the 2008 Amsterdam Principles (19). In 2009, the journal *Molecular and Cellular Proteomics* mandated public access to raw instrument data as a requirement for publication (20); this requirement, however, has been held in abeyance due to stability challenges in one of the most widely-used proteomics repositories. A 2010 NCI workshop at the Human Proteomics Organization conference in Sydney began grappling with the challenge of assessing the quality of data sets held in public repositories (21).

If data are stored in proprietary file formats, their public availability may not be sufficient to allow subsequent re-use. The HUPO-PSI community effort has developed several standard

file formats to enable broader data access. The mzML standard (22) enables researchers to supply data in a near “raw” form with open access. The mzIdentML standard (23) complements these underlying data with the peptide and protein identifications. Other formats, such as TraML (24), enable researchers to express their mass spectrometry procedures in standardized forms. As the field accommodates these advances, research data will more frequently be supplied in openly accessible formats.

Roles and perspectives

Each individual taking part in a clinical proteomic experiment may have a different perspective on the meaning of quality measurement. The technician operating the instrument is frequently envisioned as the recipient of quality measurements. Typically, an instrument is the focus of a tension; frequently a queue of samples awaits available instrument time, but losing the information of precious samples to a malfunctioning instrument is undesirable. Quality assessment can assist a technician in deciding whether or not the instrument is ready for the initiation of a new sample queue, determining if instrument performance has drifted too much for the continuation of the current queue, and diagnosing which components are contributing the greatest variability in insufficient experiments. Because most analytical chemists are not bioinformatics specialists, though, reaching this critical audience may require decision support tools for interpretation of metrics derived from data.

For principal investigators who oversee laboratories, the role of quality assessment is somewhat different. The researcher needs a strategy for determining which data sets are sufficiently sturdy for publication; trying to build a manuscript atop a data set from suboptimal experimental pipelines adds unnecessary complexity to the task. If the data are produced as part of a core service or collaboration, meeting goals for sensitivity of identification or other high-level criteria is no less essential. Monitoring performance across a laboratory of instruments can assist managers to recognize cases where major maintenance is necessary, rather than leaving an instrument to limp along in weak performance.

Characterizing data in public and private repositories is useful for their maintainers and also for the individuals, such as program officers, who hold responsibility for the data collections. Bioinformatics users of repository data may apply altogether different algorithms to data than those for which their quality was assessed, so quality assessment needs to include information that is generic to data analysis. Frequently repository data are stored without other data that might give them context. For example, a repository might store experimental data that lack information about how recently the instrument was tuned or how it was calibrated, and QC sample data that preceded and succeeded the experiment have been omitted from the upload. Complicating repository characterization further, the data are generally produced by different types of instrumentation using widely variable protocols. Generating comparable metrics from experimental files from disparate methods and instruments that measure different sample types is a non-trivial challenge.

Measuring contributions to variability

Quality analysis corresponds to the type of sample employed. Many laboratories, for example, use a defined peptide or mix of peptides to tune and calibrate instruments. QA based on these data can establish the level of variability for mass accuracy, but so simple a mixture has little to contribute in discerning sensitivity of identification. For sites that intersperse QC samples among experimental queues, slightly more complex samples are typical (25); a digest of bovine serum albumin, a predigested defined protein mix from a vendor, or even a model proteome such as a yeast lysate might be employed. Several standard samples have been commercialized, such as the Sigma UPS1 (developed in collaboration with the ABRF sPRG) and UPS2 mixes of forty-eight human proteins, the

HUPO-Invitrogen Joint Proteomics Standard, the Bruker-Michrom 6 Bovine Protein Digest Mix, or the Agilent Complex Proteomics Standard (a whole cell lysate of *Pyrococcus furiosus*). These QC samples frequently play a double role, characterizing performance during normal operation and measuring improvement during instrument maintenance. In some cases, laboratories spike a small amount of a QC standard into experimental samples, enabling defined characterization of experimental data sets. Analyzing the quality of an unaltered experimental sample introduces a host of challenges, with many parallels to the problems described above for analyzing data quality across repositories.

A 2009 consensus review by Apweiler (26) observed that “before proteomic analysis can be introduced at a broader level into the clinical setting, standardization of the preanalytical phase including patient preparation, sample collection, sample preparation, sample storage, measurement and data analysis needs to be improved.” Continuous quality assessment plays a role in ensuring that each of these techniques is performing as expected. As demonstrated in a viewpoint by Köcher, laboratories should consciously plan for quality measurement at each step (25) (see Figure 1).

The impacts of preanalytical steps may pervade the remainder of the experiment. Robinson evaluated the proteolysis of high molecular weight proteins in response to initial sample preservation steps (27), showing that different tissues may have different requirements for preservation methods. Evaluating sample degradation requires benchmarks of stability; Sköld evaluated the use of stathmin 2–20 as a proxy for degradation of other cellular proteins (28). Preanalytical variability can be contributed by protein degradation during IEF (29) or due to improper protease inactivation (30). Though trypsin digestion is part of almost every peptide-centric proteomic methodology, techniques for this step vary in reagents and timings. Picotti evaluated the peptides generated through trypsin digestion of standard proteins, discovering that hundreds of peptides, covering a wide dynamic range, are generated through these digests, increasing the space of peptides sampled during an LC-MS/MS experiment (31). This finding paralleled the finding by Tabb that protein repeatability is far higher than peptide repeatability for LC-MS/MS experiments (32). Alternative techniques for trypsinization can greatly increase the speed of digestion (33), leaving less time for non-specific cleavage. Comprehensive quality control implies measuring the impacts of these preanalytical and digestive steps.

Measurement of Variability in Proteomic Technologies

Technological differences between MALDI-based protein profiling experiments and liquid chromatography-electrospray ionization-tandem mass spectrometry experiments imply different classes of metrics to assess each. By omitting liquid chromatography, MALDI-TOF profiling is not subject to its fluctuations. Both electrospray and MALDI ionization methods, however, can introduce experimental variation. MALDI depends upon co-crystallization of samples with matrix. The success of this crystallization, however, requires the proper choice of matrix, solvent composition, and sample handling (34). Once spectra have been produced, informatics steps to calibrate, smooth, align, and normalize data are needed for comparability (35); these steps must be applied uniformly to avoid introducing artifacts. Although MALDI-TOF profiling is simpler, it is not free of variation.

Electrospray ionization-tandem mass spectrometry data sets provide information arrayed on multiple axes (see Figure 2). Retention times of peptides reflect their hydrophobicities in the reversed phase liquid chromatography that is nearly universal for electrospray ionization-based proteomics. For a particular retention time, mass analysis may be conducted to produce inventories of intact peptides (mass spectrometry) or to produce inventories of fragments produced from ions of a particular peptide (tandem mass spectrometry). For spectra from both levels of mass spectrometry, the intensity of a peak reflects its abundance,

though intensities should not be compared among chemically distinct ions. The positions of these peaks along the mass-to-charge (m/z) axis is the most precise information afforded by mass spectrometry, with modern Time of Flight and Fourier Transform-based instruments yielding mass accuracies measured in ppm. Metrics for quality assessment frequently combine information from several axes.

Extracted ion chromatograms (XICs) are one of the most common abstractions of data for quality assessment. Like all chromatograms, an XIC shows signal intensity over time for a data set, but the XIC limits its scope to the signal produced for a particular analyte (or the m/z value associated with that analyte). Building an XIC in a “shotgun” data set requires knowing both a target m/z value and the m/z tolerance associated with the mass analyzer that produced MS data; one may allow only 10 ppm tolerance for highly accurate FT-based mass analyzers while permitting a far looser 0.5 m/z for a quadrupole ion trap. Particularly when examining XICs with loose tolerances or those produced in complex mixtures, an XIC trace across an entire experiment will contain peaks for both the analyte of interest and for other ions that fall within the tolerance of the target m/z value. When data from Selected Reaction Monitoring (SRM) traces are considered, extracting a chromatogram is no longer necessary because data from each transition are recorded as chromatograms. Instead, quality metrics may sum together the chromatograms corresponding to different transitions for the same peptide ion. Common metrics for XIC evaluation include maximum peak height, signal-to-noise ratio, and full width at half maximum intensity (see Figure 3).

Mass spectrometry generates data sets that multiplex many signals in a single volume. The digest of a single, purified protein may contain evidence for more than one hundred precursor ions due to the presence of incomplete digestion products, unanticipated peptide cleavages due to other proteases and in-source digestion, oxidation and other sample processing artifacts, and charge state variation for individual peptides. When a sample contains thousands of proteins, tens or hundreds of thousands of peptides are potentially identifiable in its resulting data. The data-dependent sampling employed in most shotgun identification platforms will generate tandem mass spectra for different collections of peptides each time a sample is run, drawing from this extended pool of potential precursors. Only a small fraction of most abundant peptides produce tandem mass spectra in every experiment for a sample.

Because peptides are not distributed uniformly in hydrophobicity, peptide density changes throughout the duration of reversed phase liquid chromatography. At the most common elution times, far more peptides are eluting at once than can be comprehensively sampled in tandem mass spectrometry. The probability of interference in an SRM transition or of diminished precursor intensity due to many peptides competing for charge is greatest at this peak. If a larger number of SRM transitions are being monitored at this moment, the produced chromatograms will be sampled at lower frequency as the instrument shares its attention among more analytes. Chromatographic separation introduces variability through many channels.

Quality assessment emphasizes both transient and time-dependent variability of differing scales. Fluctuations in electrospray, for example, may lead to a temporary gap in ion production for mass spectrometry; while the data for the rest of the experiment appear normal, this gap could produce strong effects for any peptides that would normally produce data during the gap interval. At the scale of a full LC-MS/MS experiment, an instrument may show time-dependent drift, such as a slow change in time-of-flight mass accuracy during a multi-hour experiment. When instrument variability is monitored over the course of months, any number of external factors may come into play, from building maintenance

(cleaning products or vapors from painting) to seasonal temperature variation. Achieving experimental reproducibility may involve consideration of all of these time scales.

Many quality assessment approaches emphasize the information yielded from mass spectrometry rather than the mass spectral data themselves. Evaluating shotgun experiments on the basis of identified proteins and peptides is commonplace; presumably, an instrument identifying more proteins now than at another time is operating more reliably. Similarly, researchers may evaluate SRM instruments by the peptide enrichment ratios found for peptides known to differ between a pair of samples. Evaluations from derived information are frequently argued to be more relevant to evaluating the products from an instrument workflow, but they also tend to offer little resolution for diagnosing the source of variation when performance dips.

QC of 2D Gel Electrophoresis

Technologies for gel-based proteomics constitute the original “top-down” proteomics. Polyacrylamide gel electrophoresis (PAGE) separates denatured intact proteins on the basis of their ability to sediment through the pores of the gel (essentially a function of molecular mass). 2D PAGE combines a first dimension of protein separation by isoelectric focusing (IEF) with follow-on PAGE analysis, resolving proteins by both isoelectric point and molecular mass. Difference gel electrophoresis (DIGE) enables the comparison of protein content from multiple samples within a single gel, affixing fluorescent dyes to the proteins to measure the contribution of each sample to each spot appearing in the gel. Whether one has produced a 1D PAGE analysis, a 2D PAGE analysis, or a DIGE comparison, tandem mass spectrometry is generally the next step in the pipeline to translate a protein spot with estimated mass into an identification, frequently discovering that the spot contains multiple proteins.

Variation in gel electrophoresis can begin at stages that precede the gel itself. Thongboonkerd evaluated several sources of variability in the context of urinary 2D PAGE (36). Their evaluations of precipitation protocols, evaluating acetic acid and ethanol composition, showed protein yield differences of greater than an order of magnitude. The time of day for sample contribution, as well as type of liquid ingestion, complement the gender of sample contributors to modify results. These phenomena represent pre-analytical variation in the model enunciated by Apweiler (26).

Fuxius conducted a study evaluating technical replicates of mouse brain regions to determine factors that lead to false positive apparent differences (37). Their first experiment demonstrated that all samples in a comparison should be conjointly equilibrated and transferred to PAGE from IEF in order to forestall batch effects. A subsequent experiment revealed that IEF strips from different packages produce separations that differ substantially. They examined the gel casting process and found that casting all gels in a single device introduced much less variation than casting half the gels and then casting the other half. In bioinformatics, the team found that gel quantification must employ normalization within small regions because the gel is not uniformly covered with protein spots. They also recognized that mapping spots between gels requires that all gels be mapped to the same reference gel. Their study identified major sources of variation throughout analytical and bioinformatic processes for 2D PAGE-based proteomics.

Molloy evaluated coefficients of variation for 2-D gel electrophoresis experiments with and without sample variation (38). Horgan explained the relationship between the extent of differential expression, the random variation among replicates, and the sample size for comparative proteomics (39). Hunt also evaluated variation, with an eye to discerning the

statistical power of experimental designs (40). These three papers probed the relationship between intragroup variability and intergroup variability for difference testing.

Karp examined technical replicates of DIGE experiments to evaluate the extent to which variation led to false biomarker candidates by randomly producing significant p-values (41). The study compared two-color and three-color gels, finding that designs that paired a single experimental sample with a pooled control in two-color analysis were superior in resisting false positive differences. Jackson sought to build differentiation power through the use of two plasma samples taken seven days apart in 2-D DIGE (42), producing a 44% increase in apparently significant differences. Both studies offer suggestions for quality-by-design in 2D gel comparative proteomics.

Using 2D gel electrophoresis for comparative proteomics requires mapping data for a spot observed in one gel to a comparable spot in another gel. Sometimes, though, this mapping process fails. A protein may be found in experimental samples but not in controls, for example. Pedreschi described techniques for fielding missing values for comparative 2-D gel proteomics (43), favoring Bayesian principal component analysis over iterative partial least squares for its more consistent performance in both “classical” 2D gels and DIGE data.

Auditing 2D gel techniques for biomarkers produced some surprising conclusions. Evaluating publications from three years of a peer-reviewed journal, Petrak observed that 2-D gel biomarker studies frequently report proteins from a “usual suspects” list, regardless of the disease or condition evaluated (44). As the authors noted, the frequent selection of these proteins may reflect that cells respond to many different stresses by these proteins, or 2D gels may highlight them as a product of technical limitations.

The reference above separate into those intended to establish standard operating procedures (SOPs) and those assessing variability in response to changes in experimental protocol. These two efforts are essential first steps in establishing quality by design. A common theme linking these publications is the generation of multiple 2D gels from a shared sample, testing for observed differences where none should exist or evaluating the distribution of p-values against the expected uniform distribution.

QC of proteomic profiles

Proteomic profiling generates a mass spectrum or small set of mass spectra to represent the content of a complex sample. Where a 2D gel employs two orthogonal low-resolution separations, profiling relies on a high-resolution separation of intact proteins by mass-to-charge ratio (m/z), typically in a Time-of-Flight (TOF) mass analyzer. If the instrument uses Matrix-Assisted Laser Desorption Ionization (MALDI), the TOF represents the only separation. If TOF is instead paired with Surface-Enhanced Laser Desorption Ionization (SELDI), a different spectrum will be produced for each type of surface employed, though such instruments typically use TOF analyzers of lower resolution. The use of laser desorption techniques implies that the ions are typically singly-charged, producing an m/z ratio that is one proton heavier than the neutral molecule. Because data collection for a given sample lasts only a few seconds and many samples can be embedded in matrix on a single sample plate, profiling represents a very high-throughput strategy for proteomics. Like 2D gels, though, supplying only an intact mass of a protein is not sufficient for certain identification.

A 2003 examination of SELDI profiling sought both to measure variability and to describe criteria for rejection of data (45). Coombes detailed the algorithms by which peak finding, signal-to-noise determination, baseline subtraction, and normalization are conducted. Like several of the 2D gel variability analyses above, the authors began their investigation with

24 replicate mass spectra produced from the same sample on four successive days, asking whether the peaks produced from these replicates coincided with each other. The authors then explored scaling algorithms to yield comparable variance for peaks of different intensity, ANOVA examination of different factors' contributions to "batch effects," and principal components analysis (PCA) and Mahalanobis distance evaluation to recognize spectra that did not conform to the other technical replicates. The authors evaluated prior attempts at QC in SELDI profiling but drew the conclusion that prior efforts to conduct QC by enumerating the presence calls for a small set of ions were insufficiently discriminating. By moving beyond variability metrics to systems that quarantine poorly reproduced data, this work marked a significant early advance in quality control for proteomics.

A 2007 evaluation of MALDI-TOF profiling in lung cancer illustrates the techniques applied for quality assessment of proteomic profiles in experimental data (35). The authors evaluated the reproducibility and variability of peak intensities from quadruplicate measurements of the same spot on a sample plate and from four spots representing the same sample. They evaluated the impacts of freeze and thaw cycles as well as differences induced by spotting samples on separate days. As a result, the investigators were able to determine the extremity of intensity differences needed to statistically differentiate biomarkers in their plasma data.

A 2007 review of reproducibility for protein profiling by Albrethsen enumerated three challenges to the clinical use of this technology (46). First, profiling is limited to observation of the most abundant proteins in the sample. Second, the range of masses for a TOF analyzer may limit sensitivity to high mass, and peak intensities vary considerably by mass. Third, peak intensities vary for a host of reasons, limiting reproducibility. Sample crystallization in MALDI may be highly idiosyncratic for proteins, the presence of some proteins may mask the presence of others, and alterations in sample drying time on plates may alter recovery. In his review of eight recent studies, Albrethsen found the mean coefficient of variation for peak intensity to range from 4% to 26%. Seeking statistically meaningful differences becomes much harder when intrasample variability is high.

Two extensions of profiling technology increase the volume and dimension of data produced by these experiments. Imaging mass spectrometry augments profiling by producing a mass spectrum for each point in a raster that is superimposed on a 2D slice of tissue (47). Since even a small grid of 100 pixels by 100 pixels would produce a raster of 10,000 mass spectra, each of these data sets can be of substantial size. Producing replicate samples may require too much instrument time for routine practice. Variability measurements may instead benefit from the assumption that mass spectra collected at neighboring locations are more likely to be similar than mass spectra that are separated by considerable distance in the raster. Research in variability or quality for imaging mass spectrometry is still in its infancy.

A second extension of profiling technology employs peptide liquid chromatography (LC) and electrospray ionization (ESI) rather than protein-centric MALDI. Repeated sampling of high resolution mass spectra (frequently by Fourier Transform mass analyzers rather than TOF) produces a time-series of MS scans in which peptide features have both retention time and m/z coordinates. Like MALDI profiling, LC-MS profiling may be used to seek differential signatures, though more time per experiment is needed for completion of LC gradients(48). A 2006 article from Piening sought to construct quality control metrics for characterizing these data sets without the benefit of identifications (49). As described next, researchers still frequently assume that the majority of information in LC-MS experiments is to be found in MS/MS-based identifications, not the MS scans that supply masses of peptides.

QC of proteomic inventories

For many researchers in proteomics, the production of large inventories of proteins through “shotgun” or “bottom-up” techniques is the only technology they employ. These techniques frequently couple a first dimension of separation (typically strong cation exchange chromatography, isoelectric focusing, or basic-pH LC for peptides or PAGE for proteins) with a second dimension reversed phase separation of peptides. The mass spectrometer then alternates the collection of mass spectra with the acquisition of tandem mass spectra, each of which represents fragments of an isolated peptide ion. Bioinformatics pipelines then match peptide sequences to tandem mass spectra, filter these identifications to achieve a given False Discovery Rate, and infer lists of proteins for the samples. Some of the features in bioinformatics that can modify identification results are shown in Figure 4. The information yielded from such experiments can be voluminous, but the complexity of the technology leaves plenty of room for variability.

Several groups have attempted to measure the variability of the first separation in these experiments. Fang investigated a replicated comparison among isoelectric focusing, PAGE, and peptide strong cation exchange for identification sensitivity in proteomic inventories of honeybees (50). Their findings supported the use of PAGE upstream of LC-MS/MS, a technique frequently referred to as “GeLC-MS.” An examination of strong cation exchange versus a combination of peptide IEF and LC-MS/MS found the former to be more sensitive but the latter to be more reproducible and better resolved (51), illustrating that sensitivity of identification is one of several factors to be included in selecting an initial separation for shotgun proteomics.

The NCI Clinical Proteomics Technology Assessment for Cancer (CPTAC) measured the repeatability and reproducibility of identification data in LC-MS/MS, when no prior separation was employed (32); because peptide ions differed in their order of elution, their MS intensities varied, resulting in different peptides being selected for fragmentation. The CPTAC studies revealed that repeat experiments for yeast lysates produced peptide identification overlaps of 36–59%. Protein overlap percentages were consistently higher by approximately 20% in these data, causing the authors to conclude that replicate experiments sample different sets of peptides from a given protein each time an LC-MS/MS experiment is run. Peptides that were most likely to be repeated were those matching to trypsin specificity on both termini, appearing at high intensity in MS scans, and coming from abundant proteins.

Liquid chromatography appears to be a significant contributor to variability. As noted in the above article, “If peptides from a single digestion are separated on the same HPLC column twice, variations in retention times for peptides will alter the particular mix of peptides eluting from that column at a given time. These differences, in turn, impact the observed intensities for peptide ions in MS scans and influence which peptide ions will be selected for fragmentation and tandem mass spectrum collection.” Relating the ion elution times between a pair of experiments can be particularly essential when combining high-resolution LC-MS data with tandem mass spectra from another instrument (52). In this attempt, researchers have generated algorithms that predict a retention time for a peptide in one experiment based on its observation in a second.

Researchers at the National Institute of Standards and Technology working with CPTAC sought to design metrics that could capture the variability taking place at each stage of a shotgun proteomics experiment (53). They designed a tool that produced metrics for evaluating the chromatography, ion source, mass spectrometry, precursor sampling, tandem mass spectrometry, and peptide identification steps of these experiments. The software

accepts a set of identified spectra to recognize precursor ions of importance in MS data and to highlight tandem mass spectra that are known to represent peptides. In total, more than forty metrics are computed for each LC-MS/MS experiment, with other metrics representing comparisons between pairs of experiments. A subsequent paper developed software to compute these metrics for instruments from a variety of instrument manufacturers (54). The “NIST Metrics” have been widely publicized, but determining how best to employ them in a quality control process has been less clear, as is discussed in the decision-making section below.

Most of the quality assessment tools described above have evaluated the variability of whole experiments. A related effort, however, has sought to examine tandem mass spectra for shotgun proteomics experiments to discern which ones are most likely to be identified successfully and which are of lowest quality. ScanRanker (55), msmsEval (56), and QualScore (57) each pursue this goal. ScanRanker and QualScore attempt to infer partial sequences from the fragments of the tandem mass spectrum, using this information as a basis for spectrum quality estimation. QualScore and msmsEval rely upon a set of metrics computed from the fragment ions of each spectrum, as well. Each of these tools can then be used to screen out tandem mass spectra that are unlikely to contribute identifications prior to peptide matching or to identify sets of spectra that may profit from additional scrutiny by more computationally intensive techniques. Alternately, one may recognize spectra that are repeated across many experiments, even without the benefit of their identification (58). The use of such tools may enable the quality assessment of LC-MS/MS data sets without requiring identifications, enabling the rapid generation of metrics from data without regard for the species of the employed sample.

QC of proteomic quantities

Proteomic quantitation can take on several different forms. The data produced for shotgun inventories can quantify their contents in two principal strategies. The first approach, frequently termed SILAC, employs isotopic labels to differentiate light and heavy variants of the same peptides. The second approach, iTRAQ or TMT, relies upon induced isobaric peptide modifications that produce different reporter ion fragments. Each of these techniques enables a relative quantitation for hundreds or thousands of proteins that can be identified in a sample. Selected reaction monitoring (SRM), on the other hand, is a targeted approach that is intended to produce measurements for a preselected set of protein constituents. For each protein to be measured, a set of peptides act as a proxy; “transitions” are generated for these peptides that specify the peptide’s m/z value along with the m/z values of fragments expected to be produced from that peptide. SRM can then generate a chromatogram for each transition, with an integrated area that can be compared to an isotopically labeled standard peptide. While SRM variability has been recently analyzed in systematic, multi-laboratory experiments, variability has been less explored for SILAC and iTRAQ experiments.

By mixing an unlabeled sample with one that has been isotopically labeled, SILAC enables a direct comparison of intensities for each identified peptide in LC-MS/MS data sets. Pan employed media containing only ^{15}N to produce a labeled variant of *R. palustris* bacteria (59). In this context, they evaluated the variability of peptide intensity ratios as a function of signal to noise. By comparing the intensity of light and heavy isotopologues at retention times spanning the elution peak of each peptide, they employed a principal components analysis to measure the ratio of intensities as well as a signal-to-noise metric for each ratio. Their findings showed that peak intensities at low signal-to-noise values were far more likely to report 1:1 ratios than were peak intensities at high signal-to-noise values; a protein that varies 5-fold between two samples is likely to produce apparent 1:1 ratios in its low-

intensity peptides. This study provides an important cautionary note for data analysis variability in SILAC studies.

Hill considered variability of a very different kind in their 2008 article analyzing variability for iTRAQ ratio computation (60). Their approach incorporated biological and experimental factors to evaluate differential protein expression and to estimate protein fold change from the fitted ANOVA model parameter estimates. A companion paper from Oberg demonstrated these techniques in applying normalization to span complex data sets (61). Mahoney conducted an analysis of bias, reproducibility and variance in iTRAQ data sets from two instrument platforms (62). Their findings related variance among multiple sources, and their evaluation of fold changes echoed the conclusion of Pan in finding a bias toward no apparent change. Although analyzing iTRAQ data may seem a relatively simple matter of comparing two reported intensities for each peptide, careful analysis can adjust for the variability of reporter ion intensities.

As SRM experiments are becoming more widespread, measuring the site-to-site reproducibility of this technology has become more significant. Researchers in the CPTAC network clinical proteomics conducted three types of experiments on eight instruments from two vendors (63) to evaluate the contribution to variability of centralized and federated trypsin digestion and spiking of reference peptides across a wide dynamic range of concentrations. Their data established that even in experimental designs where variability is highest, CV values for peptide concentrations generally fell below 30%. A subsequent examination by variance component analysis attempted to trace variability to its sources (64). The variability was heavily influenced by the identity of the peptide or transition to be measured; some peptides were monitored more effectively than others. In addition, the three types of experiments were each associated with a different characteristic variability. With these factors taken into account, the reproducibility of SRM was quite consistent.

A similar study, associated with instrument vendor Thermo Fisher Scientific, analyzed 51 peptides in digested plasma in four independent laboratories using the same model of mass spectrometer (65). As in the CPTAC study, the authors found that SRM measurement variability was associated with peptide identity. Twelve peptides were removed from results because they were either too hydrophilic (eluting from columns too early with low signal) or too hydrophobic (sticking to the plastic sample tubes). In the most complex matrix employed and a 400 fmol spike, 25 of the remaining 39 peptides produced CVs of 5–10%, with the remaining peptides spanning 15–20% in CV. The team produced a second study, this time focusing on the use of initial enrichment strategies, coupled with SRM (66). Their examination of nanoparticle, glycopeptide, and immunoaffinity enrichment techniques showed that even with the added complexity (and consequent contribution to variability) of enrichment, CV values of below 30% were feasible. Since the most sensitive techniques for mass spectrometry-based quantitation depend upon enrichment, this study of variability for these strategies is significant.

Challenges in decision-making from QC metrics

As new sets of quality control metrics have been defined among the disciplines of proteomics, researchers have begun asking what features would make one set of metrics superior to another or which particular metrics from a set are most worthy of attention. The following properties are desirable in a quality control metric set:

- Process-spanning: evaluates many aspects of the analytical chemistry methodology
- Comprehensible: associates metrics explicitly with performance properties
- Vendor-neutral: accepts data from varied instruments

- Robust: produces reasonable metrics across very diverse experimental configurations
- Simple: relies upon straightforward computations where possible to produce metrics
- Discriminating: separates normal from anomalous results through metric change

The last two of these properties bear additional scrutiny. Any metric that depends upon the interoperation of multiple software packages is likely to be less stable and reproducible than one that can be directly computed. For example, using the sequence coverage of identified proteins requires both peptide identification and protein assembly, but version or configuration changes in either of these toolsets could lead to irreproducibility of the values. Similarly, metrics that depend heavily on features that can be defined many different ways, such as chromatographic peak width or signal-to-noise, may disagree with these values as computed by other software.

Evaluating the discrimination of a metric set, of course, depends heavily on the intended purpose. If a proteomics core facility wants to recognize when an instrument is due for tuning based on twice-daily digests of albumin, it will have different metric needs than a program officer who seeks to ensure that generated data for experimental samples are of sufficient quality. Determining whether or not a set of metrics is sufficient for a task is not straightforward, as the approaches below attest.

Apweiler (26) noted that quality control determines the acceptable deviation of values for an analyte across different samples. Stead profiled the accumulation of variability throughout the experimental process, from factors in experiment design to data analysis (67). These two publications set the stage for a fundamental question of quality control; having produced metrics of variability for each stage of an experimental pipeline, what are the best options for decision-making on the basis of those metrics?

A univariate approach highlights extreme values for a particular metric. For example, a researcher might flag an experiment as a failure if the collection of tandem mass spectra falls to an excessively low rate. Accomplishing this goal can be supported by a variety of tools. Several teams have produced tools that visualize long-term trends for metrics from shotgun proteomics data. The company Proteome Software created MassQC to visualize the NIST metrics described in Rudnick (53). Under their model, LC-MS/MS experiments for quality control samples would be uploaded to a server where their metrics would be computed and added to the accumulated population of metrics. This service, however, was discontinued in early 2012. The “Metriculator” project from Ryan Taylor in the Prince Laboratory will enable research groups to set up their own web servers for visualizing the NIST metrics from quality control samples (68). These techniques enable researchers to compare a metric value from newly produced data against the distribution of that metric from historical collections of files.

Two other tools for local monitoring of QC metrics have been developed in academic laboratories. SIMPATIQCO, created by Michael Mazanek in the Mechtler Laboratory, fields collections of QC data from Thermo instruments (69) (<http://ms.imp.ac.at/?goto=simpatiqco>). Multiple fields of information are extracted directly from the RAW files, such as ion injection times, while protein identifications are produced through automated Mascot searches (identification software from Matrix Science). Long-term trends for these fields can be visualized via the software interface. Spotfire, from TIBCO Software, serves as the analytic engine for the SWIFT pipeline, produced by Roman enka at the Mayo Clinic (70). The SWIFT pipeline incorporates identifications from Scaffold (produced by Proteome Software) along with MS/MS quality metrics generated in msmsEval, plus a set of values

extracted directly from Thermo RAW data. Tools like these can be very useful to laboratories that already make use of the Mascot or Scaffold peptide identification platforms.

These systems can also be useful in the context of comparative proteomics. Vast Scientific released RawMeat (<http://vastscientific.com/rawmeat/>) to improve the results that users achieved with the SIEVE algorithm (71). Comparative experiments are subject to the variations in the proteomic inventories upon which they are based. Michael Athanas, the author of RawMeat, noted that enabling users to produce individual LC-MS/MS experiments of greater quality improved the differentiation of peptides based on MS peptide intensities (personal communication).

Looking for outliers among the many metrics generated for each experiment produces a high rate of false quality alarms. If twenty independent metrics were generated from each experiment and the quality threshold were set at two standard deviations for each metric, on average one should expect that one of the metrics will fall beyond the threshold for each experiment (20 metrics multiplied by a threshold of 5%). To simply reject any experiment that “fails” on a single metric is excessively cautious. This model of quality, however, presupposes that the metrics are independent of each other. In fact, each metric frequently depends upon others. For example, the number of MS/MS scans collected in an experiment is related to the total ion current observed in MS scans because a more intense signal enables faster collection of an MS/MS spectrum than one that is half as intense. As a result, metrics are generally correlated with each other, containing mutual information.

Multivariate approaches may provide an answer to fielding the covariance of quality metrics. Principal component analysis (PCA) attempts to reduce a set of potentially correlated metrics to linearly uncorrelated features. Other techniques, such as the Hotelling T-squared metric, can summarize across a set of metrics to measure the distance of a new experiment from the set of previously observed experiments. Multi-dimensional scaling has already been applied for this purpose. Chaorder was introduced by Prakash to enable the recognition of relationships among collected data from LC-MS profiling experiments (72). A set of metrics characterized the signals as a function of time, and these metrics were then used to project the LC-MS experiments on two axes by multidimensional scaling. As the use of quality metrics becomes more widespread in proteomics, multivariate approaches for interpreting them are sure to follow.

Deciding whether or not to accept newly generated experimental data is just one aspect of quality control. The quality by design paradigm calls for each step of these pipelines to be continuously monitored, with changes in these pipelines governed by the search for better reliability and accuracy. Translating a set of metrics into a single value that represents the compliance of a new data set with prior expectations addresses just one piece of this challenge. As the quality by design paradigm advances, metrics will be decomposed into sets that enable the monitoring of proteolysis, liquid chromatography, ionization, and other facets of each experiment. These measurements, in turn, will help guide technicians and researchers to an accurate diagnosis of the elements that require maintenance or replacement. As proteomic quality research advances, improved reproducibility for this technology will broaden the clinical applications that it can service.

Acknowledgments

DLT was supported by NIH/NCI U24 CA159988. I appreciate suggestions and early draft commentary from Erin H. Seeley and Lorenzo Vega-Montoto.

Abbreviations

QA	Quality Assessment
QC	Quality Control
LC	Liquid Chromatography
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
SRM	Selected Reaction Monitoring
XIC	Extracted ion chromatogram

References

1. Ni , M.; Jiráť, J.; Kořata, B.; Jenkins, A.; McNaught, A., editors. IUPAC Compendium of Chemical Terminology [Internet]. 2.1.0. Research Triangle Park, NC: IUPAC; repeatability. Available from: <http://goldbook.iupac.org/R05293.html>
2. Rathore AS, Winkle H. Quality by design for biopharmaceuticals. *Nat Biotechnol*. 2009 Jan; 27(1): 26–34. [PubMed: 19131992]
3. Yu LX. Pharmaceutical quality by design: product and process development, understanding, and control. *Pharm Res*. 2008 Apr; 25(4):781–91. [PubMed: 18185986]
4. Del Val IJ, Kontoravdi C, Nagy JM. Towards the implementation of quality by design to the production of therapeutic monoclonal antibodies with desired glycosylation patterns. *Biotechnol Prog*. 2010 Dec; 26(6):1505–27. [PubMed: 20665659]
5. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer*. 2004 Apr; 4(4):309–14. [PubMed: 15057290]
6. McGuire JN, Overgaard J, Pociot F. Mass spectrometry is only one piece of the puzzle in clinical proteomics. *Brief Funct Genomic Proteomic*. 2008 Jan; 7(1):74–83. [PubMed: 18308835]
7. Karp NA, Lilley KS. Design and analysis issues in quantitative proteomics studies. *Proteomics*. 2007 Sep; 7(Suppl 1):42–50. [PubMed: 17893850]
8. Mischak H, Apweiler R, Banks RE, Conaway M, Coon J, Dominiczak A, et al. Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics Clin Appl*. 2007 Feb; 1(2):148–56. [PubMed: 21136664]
9. Sarzotti-Kelsoe M, Cox J, Cleland N, Denny T, Hural J, Needham L, et al. Evaluation and recommendations on good clinical laboratory practice guidelines for phase I-III clinical trials. *PLoS Med*. 2009 May 26; 6(5):e1000067. [PubMed: 19536325]
10. Klie S, Martens L, Vizcaíno JA, Côté R, Jones P, Apweiler R, et al. Analyzing large-scale proteomics projects with latent semantic indexing. *J Proteome Res*. 2008 Jan; 7(1):182–91. [PubMed: 18047271]
11. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat Methods*. 2009 Jun; 6(6): 423–30. [PubMed: 19448641]
12. Friedman DB, Andacht TM, Bunger MK, Chien AS, Hawke DH, Krijgsveld J, et al. The ABRF Proteomics Research Group studies: educational exercises for qualitative and quantitative proteomic analyses. *Proteomics*. 2011 Apr; 11(8):1371–81. [PubMed: 21394914]
13. Mann M. Comparative analysis to guide quality improvements in proteomics. *Nat Methods*. 2009 Oct; 6(10):717–9. [PubMed: 19953682]
14. Lescuyer P, Hochstrasser D, Rabilloud T. How shall we use the proteomics toolbox for biomarker discovery? *J Proteome Res*. 2007 Sep; 6(9):3371–6. [PubMed: 17655344]
15. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002 Feb 16; 359(9306):572–7. [PubMed: 11867112]

16. Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J Natl Cancer Inst.* 2005 Feb 16; 97(4):307–9. [PubMed: 15713966]
17. Gentleman R. Reproducible research: a bioinformatics case study. *Stat Appl Genet Mol Biol.* 2005; 4:Article2. [PubMed: 16646837]
18. Martens L, Hermjakob H. Proteomics data validation: why all must provide data. *Mol Biosyst.* 2007 Aug; 3(8):518–22. [PubMed: 17639125]
19. Rodriguez H, Snyder M, Uhlén M, Andrews P, Beavis R, Borchers C, et al. Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principles. *J Proteome Res.* 2009 Jul; 8(7):3689–92. [PubMed: 19344107]
20. Cottingham K. MCP ups the ante by mandating raw-data deposition. *J Proteome Res.* 2009 Nov; 8(11):4887–8. [PubMed: 19891508]
21. Kinsinger CR, Apffel J, Baker M, Bian X, Borchers CH, Bradshaw R, et al. Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles). *Mol Cell Proteomics.* 2011 Dec.10(12):O111.015446. [PubMed: 22052993]
22. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics.* 2011 Jan. 10(1):R110.000133. [PubMed: 20716697]
23. Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics.* 2012 Jul. 11(7):M111.014381. [PubMed: 22375074]
24. Deutsch EW, Chambers M, Neumann S, Levander F, Binz P-A, Shofstahl J, et al. TraML--a standard format for exchange of selected reaction monitoring transition lists. *Mol Cell Proteomics.* 2012 Apr.11(4):R111.015040. [PubMed: 22159873]
25. Köcher T, Pichler P, Swart R, Mechtler K. Quality control in LC-MS/MS. *Proteomics.* 2011 Mar; 11(6):1026–30. [PubMed: 21360669]
26. Apweiler R, Aslanidis C, Deufel T, Gerstner A, Hansen J, Hochstrasser D, et al. Approaching clinical proteomics: current state and future fields of application in fluid proteomics. *Clin Chem Lab Med.* 2009; 47(6):724–44. [PubMed: 19527139]
27. Robinson AA, Westbrook JA, English JA, Borén M, Dunn MJ. Assessing the use of thermal treatment to preserve the intact proteomes of post-mortem heart and brain tissue. *Proteomics.* 2009 Oct; 9(19):4433–44. [PubMed: 19688732]
28. Sköld K, Svensson M, Norrman M, Sjögren B, Svenningsson P, Andrén PE. The significance of biochemical and molecular sample integrity in brain proteomics and peptidomics: stathmin 2–20 and peptides as sample quality indicators. *Proteomics.* 2007 Dec; 7(24):4445–56. [PubMed: 18072205]
29. Finnie C, Svensson B. Proteolysis during the isoelectric focusing step of two-dimensional gel electrophoresis may be a common problem. *Anal Biochem.* 2002 Dec 15; 311(2):182–6. [PubMed: 12470679]
30. Grassl J, Westbrook JA, Robinson A, Borén M, Dunn MJ, Clyne RK. Preserving the yeast proteome from sample degradation. *Proteomics.* 2009 Oct; 9(20):4616–26. [PubMed: 19824011]
31. Picotti P, Aebersold R, Domon B. The implications of proteolytic background for shotgun proteomics. *Mol Cell Proteomics.* 2007 Sep; 6(9):1589–98. [PubMed: 17533221]
32. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham A-JL, Bunk DM, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res.* 2010 Feb 5; 9(2):761–76. [PubMed: 19921851]
33. Russell WK, Park ZY, Russell DH. Proteolysis in mixed organic-aqueous solvent systems: applications for peptide mass mapping using mass spectrometry. *Anal Chem.* 2001 Jun 1; 73(11):2682–5. [PubMed: 11403317]
34. Schwartz SA, Reyzer ML, Caprioli RM. Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: practical aspects of sample preparation. *J Mass Spectrom.* 2003 Jul; 38(7):699–708. [PubMed: 12898649]

35. Yildiz PB, Shyr Y, Rahman JSM, Wardwell NR, Zimmerman LJ, Shakhtour B, et al. Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. *J Thorac Oncol*. 2007 Oct; 2(10):893–901. [PubMed: 17909350]
36. Thongboonkerd V, Chutipongtanate S, Kanlaya R. Systematic evaluation of sample preparation methods for gel-based human urinary proteomics: quantity, quality, and variability. *J Proteome Res*. 2006 Jan; 5(1):183–91. [PubMed: 16396510]
37. Fuxius S, Eravci M, Broedel O, Weist S, Mansmann U, Eravci S, et al. Technical strategies to reduce the amount of “false significant” results in quantitative proteomics. *Proteomics*. 2008 May; 8(9):1780–4. [PubMed: 18398880]
38. Molloy MP, Brzezinski EE, Hang J, McDowell MT, VanBogelen RA. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics*. 2003 Oct; 3(10):1912–9. [PubMed: 14625853]
39. Horgan GW. Sample size and replication in 2D gel electrophoresis studies. *J Proteome Res*. 2007 Jul; 6(7):2884–7. [PubMed: 17550277]
40. Hunt SMN, Thomas MR, Sebastian LT, Pedersen SK, Harcourt RL, Sloane AJ, et al. Optimal replication and the importance of experimental design for gel-based quantitative proteomics. *J Proteome Res*. 2005 Jun; 4(3):809–19. [PubMed: 15952727]
41. Karp NA, McCormick PS, Russell MR, Lilley KS. Experimental and statistical considerations to avoid false conclusions in proteomics studies using differential in-gel electrophoresis. *Mol Cell Proteomics*. 2007 Aug; 6(8):1354–64. [PubMed: 17513293]
42. Jackson D, Herath A, Swinton J, Bramwell D, Chopra R, Hughes A, et al. Considerations for powering a clinical proteomics study: Normal variability in the human plasma proteome. *Proteomics Clin Appl*. 2009 Mar; 3(3):394–407.
43. Pedreschi R, Hertog MLATM, Carpentier SC, Lammertyn J, Robben J, Noben J-P, et al. Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. *Proteomics*. 2008 Apr; 8(7):1371–83. [PubMed: 18383008]
44. Petrak J, Ivanek R, Toman O, Cmejla R, Cmejlova J, Vyoral D, et al. Déjà vu in proteomics. A hit parade of repeatedly identified differentially expressed proteins. *Proteomics*. 2008 May; 8(9):1744–9. [PubMed: 18442176]
45. Coombes KR, Fritsche HA Jr, Clarke C, Chen J-N, Baggerly KA, Morris JS, et al. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem*. 2003 Oct; 49(10):1615–23. [PubMed: 14500586]
46. Albrechtsen J. Reproducibility in protein profiling by MALDI-TOF mass spectrometry. *Clin Chem*. 2007 May; 53(5):852–8. [PubMed: 17395711]
47. Stoeckli M, Chaurand P, Hallahan DE, Caprioli RM. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nat Med*. 2001 Apr; 7(4):493–6. [PubMed: 11283679]
48. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*. 2006 Aug 1; 22(15):1902–9. [PubMed: 16766559]
49. Piening BD, Wang P, Bangur CS, Whiteaker J, Zhang H, Feng L-C, et al. Quality control metrics for LC-MS feature detection tools demonstrated on *Saccharomyces cerevisiae* proteomic profiles. *J Proteome Res*. 2006 Jul; 5(7):1527–34. [PubMed: 16823959]
50. Fang Y, Robinson DP, Foster LJ. Quantitative analysis of proteome coverage and recovery rates for upstream fractionation methods in proteomics. *J Proteome Res*. 2010 Apr 5; 9(4):1902–12. [PubMed: 20078137]
51. Slebos RJC, Brock JWC, Winters NF, Stuart SR, Martinez MA, Li M, et al. Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. *J Proteome Res*. 2008 Dec; 7(12):5286–94. [PubMed: 18939861]
52. Jaitly N, Monroe ME, Petyuk VA, Clauss TRW, Adkins JN, Smith RD. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal Chem*. 2006 Nov 1; 78(21):7397–409. [PubMed: 17073405]

53. Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, et al. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol Cell Proteomics*. 2010 Feb; 9(2):225–41. [PubMed: 19837981]
54. Ma Z-Q, Polzin KO, Dasari S, Chambers MC, Schilling B, Gibson BW, et al. QuaMeter: Multivendor Performance Metrics for LC-MS/MS Proteomics Instrumentation. *Anal Chem*. 2012 Jul 17; 84(14):5845–50. [PubMed: 22697456]
55. Ma Z-Q, Chambers MC, Ham A-JL, Cheek KL, Whitwell CW, Aerni H-R, et al. ScanRanker: Quality assessment of tandem mass spectra via sequence tagging. *J Proteome Res*. 2011 Jul 1; 10(7):2896–904. [PubMed: 21520941]
56. Wong JWH, Sullivan MJ, Cartwright HM, Cagney G. msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics*. 2007 Feb 9; 8:51. [PubMed: 17291342]
57. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Grissem W, 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 Apr; 5(4):652–70. [PubMed: 16352522]
58. Beer I, Barnea E, Ziv T, Admon A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*. 2004 Apr; 4(4):950–60. [PubMed: 15048977]
59. Pan C, Kora G, Tabb DL, Pelletier DA, McDonald WH, Hurst GB, et al. Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Anal Chem*. 2006 Oct 15; 78(20):7110–20. [PubMed: 17037910]
60. Hill EG, Schwacke JH, Comte-Walters S, Slate EH, Oberg AL, Eckel-Passow JE, et al. A statistical model for iTRAQ data analysis. *J Proteome Res*. 2008 Aug; 7(8):3091–101. [PubMed: 18578521]
61. Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, et al. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J Proteome Res*. 2008 Jan; 7(1):225–33. [PubMed: 18173221]
62. Mahoney DW, Therneau TM, Heppelmann CJ, Higgins L, Benson LM, Zenka RM, et al. Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides. *J Proteome Res*. 2011 Sep 2; 10(9):4325–33. [PubMed: 21755926]
63. Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, Bunk DM, et al. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol*. 2009 Jul; 27(7):633–41. [PubMed: 19561596]
64. Xia JQ, Sedransk N, Feng X. Variance component analysis of a multi-site study for the reproducibility of multiple reaction monitoring measurements of peptides in human plasma. *PLoS ONE*. 2011 Jan 26; 6(1):e14590. [PubMed: 21298095]
65. Prakash A, Rezai T, Krastins B, Sarracino D, Athanas M, Russo P, et al. Platform for establishing interlaboratory reproducibility of selected reaction monitoring-based mass spectrometry peptide assays. *J Proteome Res*. 2010 Dec 3; 9(12):6678–88. [PubMed: 20945832]
66. Prakash A, Rezai T, Krastins B, Sarracino D, Athanas M, Russo P, et al. Interlaboratory reproducibility of selective reaction monitoring assays using multiple upfront analyte enrichment strategies. *J Proteome Res*. 2012 Aug 3; 11(8):3986–95. [PubMed: 22639787]
67. Stead DA, Paton NW, Missier P, Embury SM, Hedeler C, Jin B, et al. Information quality in proteomics. *Brief Bioinformatics*. 2008 Mar; 9(2):174–88. [PubMed: 18281347]
68. Taylor, R.; Dance, J.; Prince, J. Metriculator: Quality Assessment for Mass Spectrometry Based Proteomics. Career Development: Technologies; Proceedings of the 8th Annual Conference of the United States Human Proteomics Organization [Internet]; San Francisco, CA. 2012. Available from: <http://www.hupo.org/flipbook/files/inc/923710546.pdf>
69. Pichler P, Mazanek M, Dusberger F, Weinböck L, Huber CG, Stingl C, et al. SIMPATIQC: A Server-Based Software Suite Which Facilitates Monitoring the Time Course of LC-MS Performance Metrics on Orbitrap Instruments. *J Proteome Res*. 2012 Nov 2; 11(11):5540–7. [PubMed: 23088386]

70. Zenka, RM.; Johnson, KL.; Bergen, HRI. Exploring Proteomics Metadata using Spotfire and a Companion User Interface. Proceedings of the 59th ASMS Conference on Mass Spectrometry and Allied Topics [Internet]; Denver, CO. 2011. p. ThP22: 2745 Available from: http://informatics.mayo.edu/svn/trunk/mprc/swift/spotfire_companion_asms_2011.pdf
71. Sutton J, Richmond T, Shi X, Athanas M, Ptak C, Gerszten R, et al. Performance characteristics of an FT MS-based workflow for label-free differential MS analysis of human plasma: standards, reproducibility, targeted feature investigation, and application to a model of controlled myocardial infarction. *Proteomics Clin Appl*. 2008 Jun; 2(6):862–81. [PubMed: 21136885]
72. Prakash A, Piening B, Whiteaker J, Zhang H, Shaffer SA, Martin D, et al. Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics. *Mol Cell Proteomics*. 2007 Oct; 6(10):1741–8. [PubMed: 17617667]

Highlights

- Quality control for proteomics is a necessary step toward heightened reproducibility.
- Different technologies for proteomics each present their own challenges for QC.
- Tools to compute QC metrics are coming, but decision-making remains a challenge.

Figure 1. Clinical proteomics workflows include diverse methods, any of which can contribute variability. Quality by design tunes and monitors each of these steps to control variation.

Figure 2.

Liquid chromatography-mass spectrometry experiments measure thousands of mass spectra over the space of an hour or more. Each mass spectrum is a list of m/z values associated with observed intensities. In an LC-MS/MS experiment, these mass spectra are interspersed with tandem mass spectra that enumerate the fragment ions produced by the dissociation of a particular peptide.

Figure 3.

An extracted ion chromatogram (XIC) visualizes the intensity associated with a particular m/z value over a range of retention times from liquid chromatography. An XIC is typically integrated to produce a peak area that measures the relative abundance of an ion.

Chromatographic resolution for an XIC is frequently expressed as the full width at half maximum (FWHM), the peak width in retention time where the shoulders of the peak are half the maximum observed. Signal-to-noise evaluates the extent to which a peak stands out from local intensity fluctuations.

Figure 4.

Many configuration choices during the bioinformatics steps of protein identification can significantly alter the peptides and proteins identified from an experimental data set. This Ishikawa diagram relates some of the most prominent sources of variation. For example, converting peak profiles from a Time-of-Flight mass analyzer to a list of peaks will result in very different lists when different tools are employed. The file format to which these peak lists are written may contain many types of metadata (as in the mzML format) or almost none (as in PKL or DTA format).