

COMMENTARY

Whole proteomes as internal standards in quantitative proteomics

Shao-En Ong*

Abstract

As mass-spectrometry-based quantitative proteomics approaches become increasingly powerful, researchers are taking advantage of well established methodologies and improving instrumentation to pioneer new protein expression profiling methods. For example, pooling several proteomes labeled using the stable isotope labeling by amino acids in cell culture (SILAC) method yields a whole-proteome stable isotope-labeled internal standard that can be mixed with a tissue-derived proteome for quantification. By increasing quantitative accuracy in the analysis of tissue proteomes, such methods should improve integration of protein expression profiling data with transcriptomic data and enhance downstream bioinformatic analyses. An accurate and scalable quantitative method to analyze tumor proteomes at the depth of several thousand proteins provides a powerful tool for global protein quantification of tissue samples and promises to redefine our understanding of tumor biology.

Introduction

Mass spectrometry (MS)-based proteomics is a uniquely powerful and versatile tool in biology as it allows unbiased, comprehensive and sensitive detection of proteins and post-translational protein modifications in complex mixtures. With the ability to identify thousands of proteins in a single experiment, MS-based proteomics makes it easy to generate lengthy protein catalogs, but qualitative comparisons of lists of proteins is less informative. Instead, the ability to quantify abundances of whole proteomes and to observe these changing over time or in response to a defined perturbation would be very powerful. Such information can be obtained with

quantitative proteomics, which greatly enhances the power and utility of MS-based methods [1,2].

MS measures and distinguishes analytes by their masses. The more robust and accurate quantification methods use stable isotopes such as ^{13}C , ^{15}N and ^{18}O to introduce a detectable increase in mass. Except for the increased mass from the additional neutrons, the stable isotope labeled (SIL) internal standard and the analyte are essentially indistinguishable. Comparing MS peak signal intensities from samples containing unlabeled 'light' and SIL 'heavy' peptides quantifies relative protein abundance. Minimizing physicochemical differences between the analyte and the internal standard allows analytical workflows to be combined and reduces experimental errors in quantification.

The toolbox for quantitative proteomics continues to expand, providing many options for researchers. Recently, Mann and co-workers described an approach based on stable isotope labeling by amino acids in cell culture (SILAC) [3] that combines multiple cellular proteomes to obtain whole proteome SIL standards suitable for the quantification of the complex tissue proteomes that are typical in clinical proteomics [4].

Pooling proteomes as internal standards

For over two decades, researchers have spiked peptides stably labeled with isotopes into samples and quantified these reference standards against their endogenous counterparts to measure protein levels. This approach to quantifying small numbers of analytes from complex peptide mixtures with targeted MS assays has grown in popularity for studying specific protein classes, such as kinases [5], and especially as a platform for the validation of candidate biomarkers in clinical samples (Figure 1a) [6,7]. Alternatively, faster peptide sequencing capabilities in modern MS instruments enable approaches combining peptide identification and quantification to provide whole-proteome analysis of differential protein expression. Stable isotope labels are introduced in entire proteomes through chemical derivatization with SIL tags [8,9] or metabolic labeling with essential metabolites such as SIL amino acids [3]. The latter approach, requiring living cells, is often thought to be incompatible with tissue proteomics.

*Correspondence: song@broadinstitute.org
Proteomics and Biomarker Discovery Platform, The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA

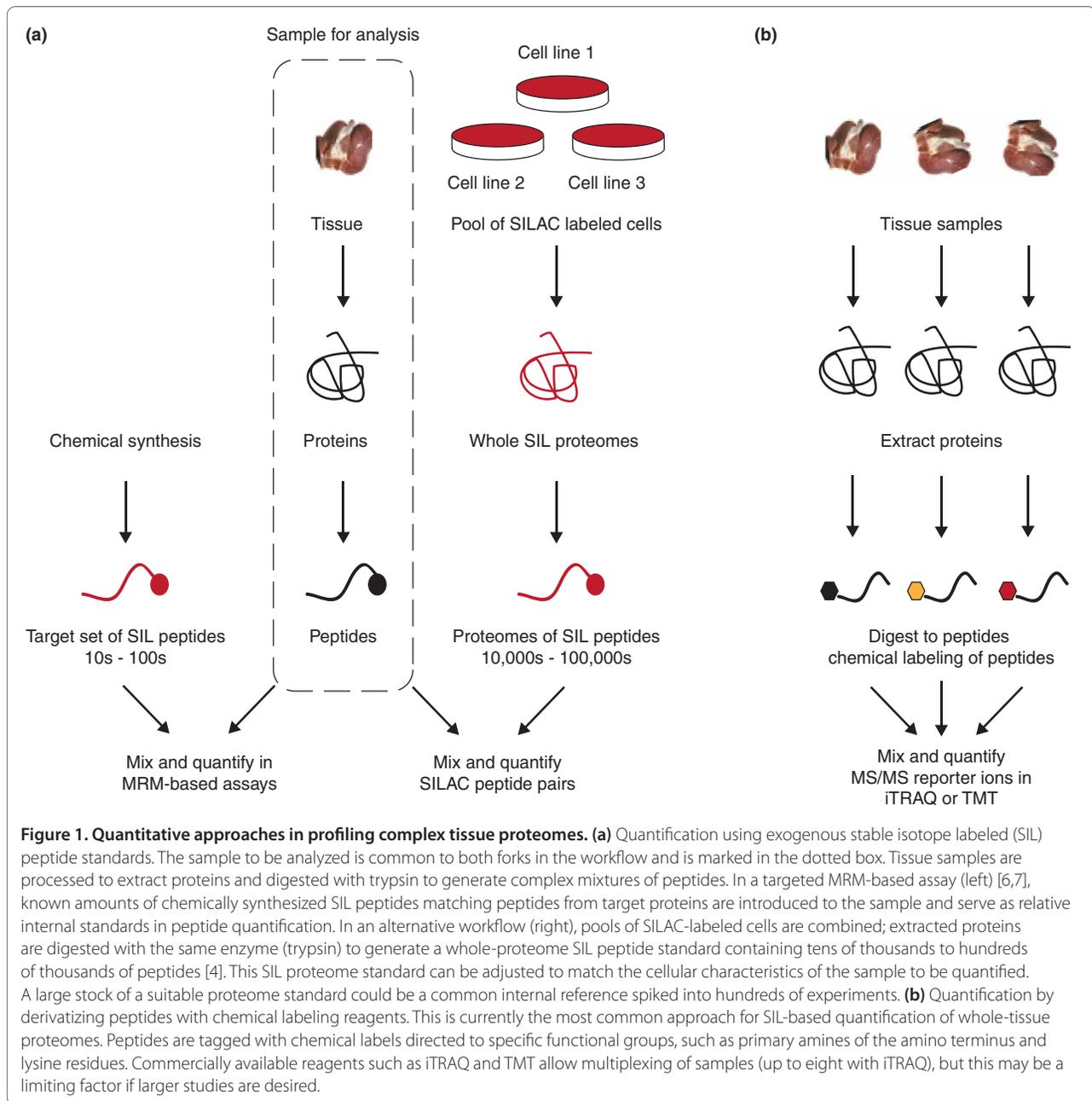


Figure 1. Quantitative approaches in profiling complex tissue proteomes. (a) Quantification using exogenous stable isotope labeled (SIL) peptide standards. The sample to be analyzed is common to both forks in the workflow and is marked in the dotted box. Tissue samples are processed to extract proteins and digested with trypsin to generate complex mixtures of peptides. In a targeted MRM-based assay (left) [6,7], known amounts of chemically synthesized SIL peptides matching peptides from target proteins are introduced to the sample and serve as relative internal standards in peptide quantification. In an alternative workflow (right), pools of SILAC-labeled cells are combined; extracted proteins are digested with the same enzyme (trypsin) to generate a whole-proteome SIL peptide standard containing tens of thousands to hundreds of thousands of peptides [4]. This SIL proteome standard can be adjusted to match the cellular characteristics of the sample to be quantified. A large stock of a suitable proteome standard could be a common internal reference spiked into hundreds of experiments. **(b)** Quantification by derivatizing peptides with chemical labeling reagents. This is currently the most common approach for SIL-based quantification of whole-tissue proteomes. Peptides are tagged with chemical labels directed to specific functional groups, such as primary amines of the amino terminus and lysine residues. Commercially available reagents such as iTRAQ and TMT allow multiplexing of samples (up to eight with iTRAQ), but this may be a limiting factor if larger studies are desired.

The heterogeneity of tissue has always complicated the analysis of its molecular components and is probably the central challenge in comprehensive analyses of tissue proteomes. Despite the difficulties, our understanding of disease biology could be greatly enhanced by improved methods to accurately profile global protein expression in tissue samples, such as patient tumor biopsies. Clinical tissue proteomics currently lags behind proteomics in other areas, such as model organisms or cell culture-based systems, particularly in quantitative comparisons of protein abundance between tissue samples. An important application in clinical proteomics is the identification

of protein biomarkers in samples from diseased versus unaffected people [7]. These clinical samples may be from tumor tissue or biological fluids near affected sites. Biomarker studies commonly apply a staged approach: initial discovery of highly differentially expressed proteins followed by more careful validation with spiked SIL internal standards to quantify specific proteins. In the discovery phase, it is possible to use chemical labeling strategies (Figure 1b) to compare six or up to eight tissue samples simultaneously with the commercial reagents tandem mass tags (TMT) [9] or the isobaric tag for relative and absolute quantification (iTRAQ) [8], respectively.

More commonly, however, researchers use semi-quantitative measures such as spectral counts [10] or total peptide signal intensity from identified peptides to determine differential expression [11,12]. Because of the larger variances in these semi-quantitative measurements, only very differentially expressed proteins are selected for downstream validation experiments, such as quantitative multiple reaction monitoring (MRM)-MS assays.

The approach of Mann and coworkers [4] may bridge the gap between the stages of initial discovery and MRM-MS validation of candidate biomarkers. They pooled five different SILAC-labeled breast cancer cell lines to generate a superset of SIL peptides derived from their combined proteomes. The large collection of peptides in the super-SILAC mix was then applied as internal standards to quantify proteins in breast and brain tumor samples. Their work [4] builds on earlier work from Ishihama *et al.* [13] in which a single SILAC-labeled neuroblastoma cell line was used to quantify protein expression in mouse brain. Because the whole-proteome SIL standard is derived from multiple cell lines, it provides a diverse pool of proteins that can be adjusted to more accurately represent the heterogeneous cell populations of a particular tumor sample, thus increasing the likelihood that a tumor-derived peptide will have a heavy SIL counterpart for accurate quantification. Geiger *et al.* [4] achieved high quantitative coverage, quantifying over 70% of identified proteins in both tumor samples and improving overall quantitative accuracy through the use of the pooled SILAC cell lines when compared with a single labeled cell line.

There are several practical advantages: SILAC labeling is inexpensive and several million cells can yield milligrams of SIL internal standards, material sufficient for hundreds of experiments. Although the authors [4] pooled only carcinoma cell lines, combining a more diverse collection of SILAC labeled cell lines and mixing these at different levels might better mimic the heterogeneity of cell types in a tumor. Quantitative accuracy would then be substantially better, as a greater number of SIL peptides would serve as internal standards for quantification or be available as 'landmarks' in normalization and sample matching [13,14]. The super-SILAC approach is scalable and flexible, allowing the generation of reference libraries of SIL peptides that can be applied over the duration of a lengthy biomarker discovery campaign, spanning different tissue types and sample sources. Improved quantification of complex tissue proteomic samples in the discovery phase could substantially improve confidence in the identification of differentially expressed proteins, effectively triaging the long lists of candidate biomarkers requiring validation.

Not surprisingly, spiking in a whole proteome's worth of SIL peptides brings new analytical challenges. The

combined super-SILAC and tumor proteome mixture will have at least doubled in complexity, and the dynamic range of accurate peptide quantification may not span the full range of analytes of interest. Indeed, the whole-proteome SIL standard is unlikely to be useful in the validation phase of biomarker discovery. Interfering signals from unrelated peptide species compromise MRM-MS assays, requiring the monitoring of multiple peptide precursor-fragment transitions to increase specificity when quantifying a particular peptide analyte. Adding hundreds of thousands of SIL peptides for MRM assays is unnecessary because experiments target specific peptides and doing so will have only a negative impact on quantitative accuracy and specificity.

Conclusions

There is relatively little collective experience in defining protein expression profiles from biomarker studies. There are few published biomarker discovery datasets and even fewer in public data repositories, in stark contrast to widely available microarray and next-generation high-throughput genomic data. We do not yet have common protocols for processing protein samples similar to those well established in transcript profiling experiments. Proteins cannot be amplified with powerful PCR-based methods and, compared with mRNA, proteins are less homogeneous and require more care in handling and extraction. Many current datasets of biomarker protein expression profiles use semi-quantitative measures of protein abundance; large variations in these profiles complicate attempts to extract meaningful hypotheses and limit their overall utility. The researcher has little choice but to attribute quantitative variation to biological noise and sample variability and only select proteins with the most significant expression differences for downstream validation experiments.

The complexities of tumor biology may well turn out to be the limiting factor in our attempts to make molecular profiles of cancer, but it is certainly harder to argue against better analytical tools. Greater quantitative accuracy, afforded by the use of a super-SILAC proteome standard or other means, will undoubtedly improve the quality of tissue protein expression profiles and our ability to confidently identify subtle changes in protein expression. Widespread use of whole-proteome SIL standards may provide a framework, similar to approaches commonly used in gene expression profiling [15], to standardize quantitative analyses of complex tissue samples in clinical proteomics. The ability to robustly compare different clinical proteomics datasets would facilitate the integration of datasets from proteomics and genomics and transform the field of clinical proteomics.

Abbreviations

iTRAQ, isobaric tag for relative and absolute quantification; MRM, multiple reaction monitoring; MS, mass spectrometry; SIL, stable isotope labeled/labeling; SILAC, stable isotope labeling by amino acids in cell culture; TMT, tandem mass tag.

Competing interests

The author declares that they have no competing interests.

Published: 30 July 2010

References

1. Gstaiger M, Aebersold R: **Applying mass spectrometry-based proteomics to genetics, genomics and network biology.** *Nat Rev Genet* 2009, **10**:617-627.
2. Ong SE, Mann M: **Mass spectrometry-based proteomics turns quantitative.** *Nat Chem Biol* 2005, **1**:252-262.
3. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.** *Mol Cell Proteomics* 2002, **1**:376-386.
4. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M: **Super-SILAC mix for quantitative proteomics of human tumor tissue.** *Nat Methods* 2010, **7**:383-385.
5. Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, Wenschuh H, Aebersold R: **High-throughput generation of selected reaction-monitoring assays for proteins and proteomes.** *Nat Methods* 2010, **7**:43-46.
6. Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, Bunk DM, Spiegelman CH, Zimmerman LJ, Ham AJ, Keshishian H, Hall SC, Allen S, Blackman RK, Borchers CH, Buck C, Cardasis HL, Cusack MP, Dodder NG, Gibson BW, Held JM, Hiltke T, Jackson A, Johansen EB, Kinsinger CR, Li J, Mesri M, Neubert TA, Niles RK, Pulsipher TC, Ransohoff D, 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.
7. Rifai N, Gillette MA, Carr SA: **Protein biomarker discovery and validation: the long and uncertain path to clinical utility.** *Nat Biotechnol* 2006, **24**:971-983.
8. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlett-Jones M, He F, Jacobson A, Pappin DJ: **Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents.** *Mol Cell Proteomics* 2004, **3**:1154-1169.
9. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C: **Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS.** *Anal Chem* 2003, **75**:1895-1904.
10. Liu H, Sadygov RG, Yates JR 3rd: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics.** *Anal Chem* 2004, **76**:4193-4201.
11. Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, Koziol JA, Schnitzer JE: **Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis.** *Nat Biotechnol* 2010, **28**:83-89.
12. Negishi A, Ono M, Handa Y, Kato H, Yamashita K, Honda K, Shitashige M, Satow R, Sakuma T, Kuwabara H, Omura K, Hirohashi S, Yamada T: **Large-scale quantitative clinical proteomics by label-free liquid chromatography and mass spectrometry.** *Cancer Sci* 2009, **100**:514-519.
13. Ishihama Y, Sato T, Tabata T, Miyamoto N, Sagane K, Nagasu T, Oda Y: **Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards.** *Nat Biotechnol* 2005, **23**:617-621.
14. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, Vitek O, Aebersold R, Muller M: **SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling.** *Proteomics* 2007, **7**:3470-3480.
15. Dozmorov I, Lefkovits I: **Internal standard-based analysis of microarray data. Part 1: analysis of differential gene expressions.** *Nucleic Acids Res* 2009, **37**:6323-6339.

doi:10.1186/gm170

Cite this article as: Ong S-E: Whole proteomes as internal standards in quantitative proteomics. *Genome Medicine* 2010, **2**:49.