

SYMPOSIUM

Proteomics as a Tool for Clinically Relevant Biomarker Discovery and Validation

MARK W. DUNCAN*,†¹ AND STEPHEN W. HUNSUCKER*

**Department of Pediatrics, Section of Pulmonary Medicine, and †Department of Medicine, Division of Endocrinology, Diabetes and Metabolism, University of Colorado at Denver and Health Sciences Center, Aurora, Colorado 80045*

The excitement associated with clinical applications of proteomics was initially focused on its potential to serve as a vehicle for both biomarker discovery and drug discovery and routine clinical sample analysis. Some approaches were thought to be able to "identify" mass spectral characteristics that distinguished between control and disease samples, and thereafter it was believed that the same tool could be employed to screen samples in a high-throughput clinical setting. However, this has been difficult to achieve, and the early promise is yet to be fully realized. While we see an important place for mass spectrometry in drug and biomarker discovery, we believe that alternative strategies will prove more fruitful for routine analysis. Here we discuss the power and versatility of 2D gels and mass spectrometry in the discovery phase of biomarker work but argue that it is better to rely on immunochemical methods for high-throughput validation and routine assay applications. *Exp Biol Med* 230:808–817, 2005

Key words: DIGE; difference gel electrophoresis

Introduction

Proteomics has remarkable potential to enhance our understanding and practice of medicine. The Human Genome Project has revealed that there are many fewer protein-coding genes in the human genome than there are

proteins in the human proteome (~30,000 genes vs. ~400,000 proteins), and therefore proteomics, more than gene expression analysis, is increasingly seen as an essential tool in understanding the complexities of normal human physiology and disease (1). One of the areas where it has the greatest potential is in the discovery of new diagnostic and prognostic biomarkers² of disease (2).

In proteomics there are several distinct analytical platforms for the task of discovering new biomarkers. These approaches differ substantially in design, and this influences the speed (throughput), cost, complexity, and, most important, the fundamental aspects of the data they return. In this paper we will outline common proteomics approaches and discuss how proteomics methods are applied in biomarker studies by way of examples from our own work.

Major Paradigms in Biomarker Discovery

The three primary approaches to proteomics are outlined in the sections that follow. Table 1 summarizes the advantages and disadvantages of the three major paradigms.

1. Protein Profiling. Protein profiling is an approach that attempts to identify molecular signatures, or unique features within mass spectral profiles, that characterize biological samples. Profiling has most frequently been applied to human serum studies. While the details of the process can vary markedly, the common characteristic of this approach is that a plot of m/z versus intensity (a mass spectrum) is generated on a complex mixture; however, the components of the sample are not identified. Differentially abundant proteins are determined by comparing protein

This work was supported by grants from the Cystic Fibrosis Foundation (DUNCAN01U1) and the University of Colorado Cancer Center Lung Cancer SPORE Pilot Program (P50 CA058187).

¹ To whom correspondence should be addressed at Department of Pediatrics, Section of Pulmonary Medicine, University of Colorado at Denver and Health Sciences Center, Mail Stop 8119, P.O. Box 6511, Aurora, CO 80045. E-mail: Mark.Duncan@uchsc.edu

1535-3702/05/23011-0808\$15.00

Copyright © 2005 by the Society for Experimental Biology and Medicine

² **Biomarker**—A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention.

Table 1. Summary of the Relative Merits of the Three Major Paradigms in Proteomic Research

Paradigm	Advantages	Disadvantages
Protein profiling	<p>Rapid generation of protein profiles. Applicable to large sample numbers in a high-throughput fashion.</p> <p>Representation of low-molecular-weight polypeptides is good.</p> <p>Analysis can be performed directly from a crude sample.</p> <p>Generally requires only small sample amounts.</p> <p>Except for the cost of expensive consumables that may be specific to a single strategy, this approach is not labor intensive and can be fully automated.</p>	<p>Technique has not been proven to be reproducible. Reproducibility is reliant upon sample collection and handling methods.</p> <p>Proteins are not identified in profiling experiments. Other laborious and costly methods are required to gain any insight into the identity of the differentially abundant proteins. Even then, there is ambiguity regarding protein identification.</p> <p>Under-representation of molecular weight > ~30 kDa.</p> <p>High probability of overlapping protein signals due to the complexity of the proteome, especially on low-resolution systems.</p> <p>Coverage of the proteome is low (i.e., only tens to low hundreds of proteins represented).</p> <p>Mass resolution and mass accuracy are poor (although this is very instrument dependent).</p> <p>Software for pattern comparisons is suboptimal.</p> <p>Time consuming relative to alternative strategies.</p> <p>Poor representation of very acidic and very basic proteins.</p> <p>Underrepresentation of low- and high-mass proteins.</p> <p>Poor representation of hydrophobic proteins.</p> <p>Expensive.</p> <p>Difficult to automate relative to alternative methods.</p> <p>Labor intensive at some stages.</p>
2D gels/mass spectrometry	<p>Capable of generating highly reproducible results.</p> <p>Good coverage of proteins with isoelectric point (pI) 3–11 and molecular weight 10–200 kDa.</p> <p>Separates intact proteins so isoforms are individually detected and quantified.</p> <p>In conjunction with approaches such as DIGE, can provide precise (relative) quantitative information on intact proteins.</p> <p>In combination with mass spectrometry, proteins can be identified and fully characterized.</p> <p>Coverage of the proteome can be high if combined with additional fractionation steps.</p> <p>Protein resolution is unrivaled.</p>	<p>Information on multiple isoforms is lost.</p> <p>Coverage of the sequence of each protein is generally low.</p> <p>Quantification of proteins is based on quantification of peptides and is therefore subject to error.</p> <p>In general, precision of available quantification strategies is suboptimal.</p> <p>Complex data analysis routines are required.</p>
Shotgun strategies	<p>Can be used in a high-throughput mode.</p> <p>Good representation of all proteins, independent of pI and molecular weight.</p> <p>Offers good representation of lipophilic and membrane-bound proteins.</p> <p>Analysis can be performed directly on crude samples.</p> <p>Easily automated.</p>	

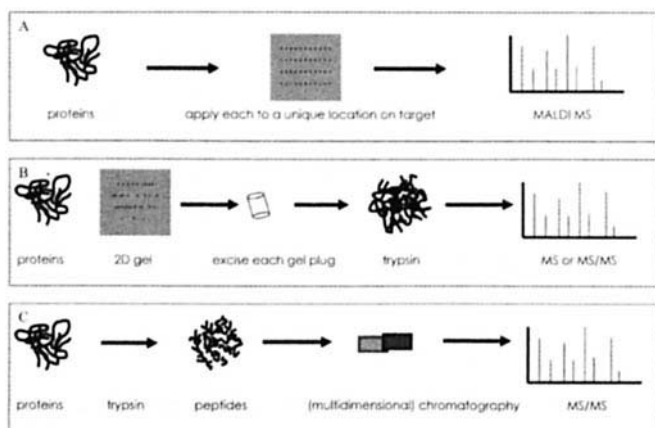


Figure 1. Schematic representations of the three major paradigms of proteomic approaches in biomarker research. (A) Protein profiling. (B) 2D gels combined with mass spectrometry. (C) The shotgun strategy.

profiles (peak intensities) across many samples (mass spectra). In its most common manifestation, the Ciphergen SELDI approach, a “chip” modified with chromatographic media or affinity capture surfaces, selectively retains a subset of the molecular species present in the sample. The crude biological sample is applied directly to the chip surface and washed to aid in the removal of components that do not specifically bind, a UV-absorbing matrix is added, and the sample is then analyzed in a MALDI mass spectrometer (3, 4). Other manifestations of profiling involve off-line sample fractionation followed by MALDI time-of-flight mass spectrometry (ToFMS) analysis (5, 6), direct application of sample to an unmodified target (7), and imaging of tissue samples (8, 9).

Profiling continues to attract considerable interest, in large part because of its potential to deliver high-throughput data. In particular, attention has been directed at its application in a clinical laboratory setting. However, increasingly this approach is viewed with skepticism. In part this is because different groups undertaking essentially the same analyses have come up with different diagnostic features (10). Further, it is increasingly apparent that when compared with alternative biomarker discovery strategies, the spectra are not rich in information, and the complexity of the proteome is substantially underestimated. However, perhaps the most serious concern is that it is difficult to move away from this platform to the next step and to identify the distinguishing “peaks” in the profiles. This is necessary if the tests are to be validated by independent methods, if mechanistic data are to flow from these investigations, or to develop platform-independent assays for these protein markers. Finally, and unfortunately, several studies based on a compromised experimental design and/or suboptimal data analysis have profoundly dampened enthusiasm for this approach (4, 10–15). These experiences have led to scrutiny of this approach and highlighted the need for due diligence in all phases of the biomarker discovery and validation process.

2. 2D Gels Combined with Mass Spectrometry. Another commonly adopted approach to discovery proteomics involves 2D gel electrophoresis (2DGE). Here a complex mixture of proteins is separated in the first dimension based on isoelectric point (pI) and then, orthogonally, in the second dimension based on molecular weight (SDS-polyacrylamide gel electrophoresis [SDS-PAGE]). Thousands of proteins can be resolved on a single gel, the gel stained to visualize them, and their relative amounts determined. The protein spots can be excised from the gel, digested with a protease (usually trypsin), and analyzed by mass spectrometry to identify the parent proteins. Protein identities are determined by comparison of experimental data to primary sequence databases. Gel-based approaches are most frequently coupled with MALDI-ToFMS to generate a mass fingerprint (or map), but liquid chromatography combined with tandem mass spectrometry (LC-MS/MS) is also used, especially to resolve ambiguities. Changes in the abundance and/or position of spots on the gel can be used to make detailed qualitative and quantitative comparisons between two or more conditions. The success of the 2D gel strategy is illustrated by over 1500 papers reporting the application of this approach.

Gels are a powerful approach to separation, but proteins at the extremes of either molecular weight or pI are underrepresented by this approach, and, in particular, hydrophobic proteins are difficult to analyze without resorting to specialized methods.

Enthusiasm for performing comparative proteomics studies based on 2D gel electrophoresis was dampened by the irreproducibility of the technique, and consequently, the problems associated with making comparisons across gels. Difference gel electrophoresis (DIGE) is an emerging technology that provides an increase in analytical precision, dynamic range, and sensitivity. DIGE facilitates repetitive measurements and multivariable analyses in a single, coordinated experiment because it allows for the incorporation of an internal standard composed of equal amounts of every sample.

In DIGE, two distinct protein samples (e.g., representing normal and disease) are separately labeled with two cyanine dyes, mixed, and then separated by 2DGE on a single gel (16). Fluorescent laser scanning of the gel at two distinct excitation and emission wavelengths generates two distinct images that can be superimposed on a pixel-to-pixel basis. Relative increases and decreases in the levels of the proteins can therefore be quantified precisely. Furthermore, a third dye is available that can be used as an internal standard to increase quantitative precision and improve protein spot matching when comparing multiple gel images (17). Without the benefits of the two-dye (or three-dye) strategy, high analytical variability makes it difficult to determine biological variability, and this has stymied previous attempts to obtain statistically meaningful results.

The capability to run *two samples* on the *same gel* eliminates this problem.

3. Shotgun Strategies: LC-MS/MS. Increasingly, proteomic investigations use the shotgun strategy, an approach based on proteolysis of proteins followed by analysis of the complex peptide mixture by LC-MS/MS (18). Sequences are then assigned to the MS/MS spectra by automated database searching algorithms. The approach is well suited to automated analysis and gives good coverage of the diverse array of proteins present in biological samples, and with the current generation of LC-MS/MS systems, this approach offers unrivaled sensitivity. Shotgun strategies are a powerful approach to gaining specific information regarding the peptide constituents of a complex mixture, but it is important to recognize that the experimental process begins with enzymatic digestion. Consequently, unambiguous reassembly of the peptide sequences into their precursor proteins is rarely, if ever, possible, especially in higher eukaryotic organisms, because a single peptide sequence can be represented in several distinct and biological active proteins (isoforms).

Shotgun strategies can also be combined with stable isotope labeling to allow for the quantification of changes in protein expression levels of hundreds to thousands of proteins in a single experiment. The most commonly adopted approaches include ICAT (19–21) and *i*TRAC (22) methods. However, quantification is again based on relative changes in the levels of labeled *peptides* that may be common to a family of proteins, with differential regulation/abundance, and therefore quantification experiments can lead to ambiguous or conflicting results.

Key Considerations in Biomarker Research

Biomarker research aims to identify, from the thousands of peptides and proteins present in a biological tissue or fluid, previously unrecognized compounds and/or associations that prove to be both specific and sensitive

for disease. Ultimately, proteomics promises to deliver panels of proteins that deliver much greater diagnostic power than any single analyte alone (23).

The Early Detection Research Network, established by the Division of Cancer Prevention, National Cancer Institute, has identified five separate phases in the development and testing of disease biomarkers (24). These phases are summarized in Table 2. Notably, they clearly distinguish between the tasks of identifying promising directions (or phase I, identification) and the subsequent careful characterization of these candidates (phases 2–5).

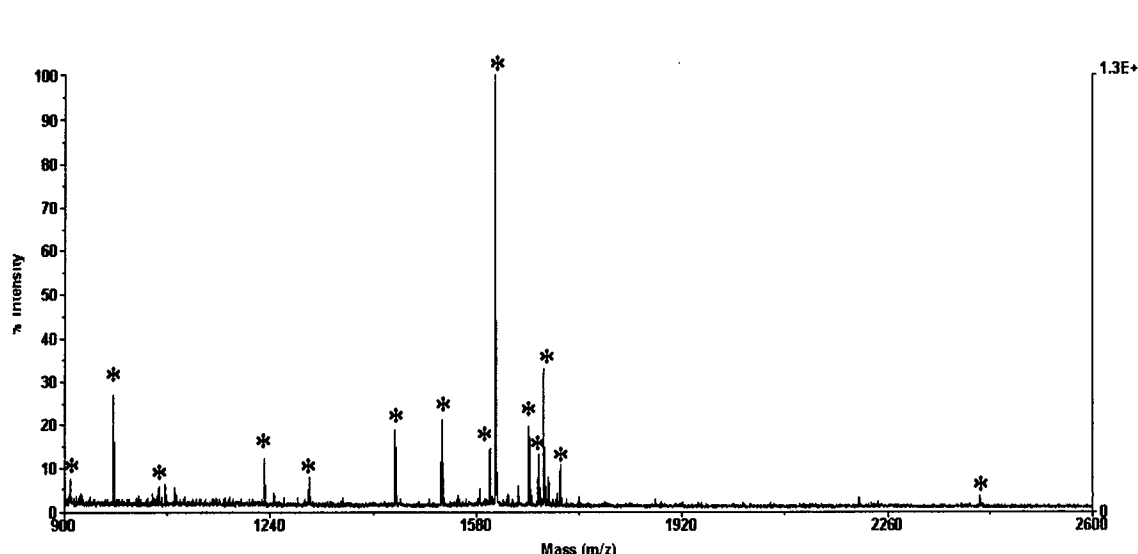
Although there is general agreement that several distinct stages are required to translate a newly discovered biomarker into a clinical assay, there is no consensus on how this can be best achieved. In part the controversy rages because, up until now, many groups have focused on single-technology solutions to the problem. Currently, protein profiling, 2D gel electrophoresis (2DGE), liquid chromatography coupled with mass spectrometry, and protein arrays are just a few of the approaches that vie for prominence on the biomarker landscape. However, the reality is that the complex and diverse requirements of biomarker investigations, from the process of discovery through clinical implementation, call for disparate technologies strategically adopted at each stage of the process. Increasingly, investigators will need to embrace several distinct processes, first to discover biomarkers and then later to validate them in substantial patient populations.

The three paradigms discussed here have been widely employed and over the past decade or more have delivered a continuous outpouring of data. However, of the thousands of putative biomarkers identified by these strategies, few have been validated, and even fewer have made their way into routine clinical use. In the following sections we highlight some important considerations in the area of new biomarker discovery and subsequent clinical implementation, and we illustrate these points with several examples from our own work.

The Discovery Phase. Qualitative Considerations. During the discovery phase of biomarker work, the most assiduous approach is one that allows comprehensive qualitative and precise quantitative analysis of biological samples with the objective of identifying the largest possible set of proteins that distinguish between the control and test populations. Currently, discovery is usually undertaken by employing either 2DGE or LC-MS/MS: two distinct strategies that return complementary data sets. For example, although 2D gels unveil some of the complexity inherent in each gene product, the technique usually fails to provide representative coverage of small, basic, and low-abundance proteins. Shotgun strategies deliver more comprehensive representation of the protein complement, but with this approach information on each distinct gene product, including truncations, alternative splicing, and other modifications is typically lost. Not surprisingly then, when both techniques have been applied to the analysis of the very

Table 2. The Five Phases Involved in the Development and Testing of Disease Biomarkers as Proposed by the Early Detection Research Network (EDRN)

Phase I	Exploratory studies to identify potentially useful biomarkers (i.e., the “discovery” phase).
Phase II	Biomarkers are studied to determine their capacity for distinguishing between people with cancer and those without (i.e., the “validation” phase).
Phase III	Studies to assess the capacity of a biomarker to detect preclinical disease by testing the marker against tissues collected longitudinally from research cohorts.
Phase IV	Prospective screening studies.
Phase V	Definitive large-scale population studies to determine the overall impact of screening on health outcomes in the target population.



3951	HELRDLVEFK	PLAPDGVLLF	SGGKSGPVED	FVSLAMVGGH	LEFRYELGSG
4001	LAVLRTAEP	ALGRWHRVSA	ERLNKDGLR	VNGGRPVLRS	SPGKSQGLNL
4051	HTLLYLGGVE	PSVPLSPATN	MSAHFRGCVG	EVSVNGKRLD	LTYSFLGSQG
4101	IGQCYDSSPC	ERQPCQHAT	CMPAGEYEFQ	CLCRDGIKGD	LCEHEENPCQ
4151	LREPCLHGGT	CQGTRCLCLP	GFSGPRCQQG	SGHGIAESDW	HLEGSGGNDA
4201	PGQYGAYFHD	DGFLAFPGHV	FSRSLPEVPE	TIELEVRTST	ASGLLLWQGV
4251	EVGEAGQGKD	FISLGLQDGH	LVFRYOLGSG	EARLVSEDPI	NDGEWHRVTA
4301	<u>LREGRRGSIO</u>	<u>VDGEELVSGB</u>	<u>SPGPNVAVNA</u>	<u>KGSIYIGGAP</u>	<u>DVATLTGGRE</u>
4351	<u>SSGITGCVKN</u>	<u>LVLHSARPGA</u>	<u>PPPQPLDLQH</u>	<u>RAQAGANTRP</u>	<u>CPS</u>

Figure 2. A peptide mass map generated by MALDI time-of-flight mass spectrometry that was identified by database searching as the C-terminal portion of heparan sulfate proteoglycan. The spectrum shows the mass values that correspond to tryptic peptides from this protein (denoted with asterisks). All major peaks in the spectrum were assigned. The C-terminal portion of the amino acid sequence is depicted below the spectrum with corresponding residue numbers. All underlined amino acids (singly or doubly underlined) were detected in the peptide mass map, and the doubly underlined peptides were confirmed by LC-MS/MS analysis.

same sample, the overlap between the data sets is modest (25, 26), and it therefore follows that the application of several tools is advisable if not essential. This applies whether the objective is to search for biomarkers or to unravel the complexities of a biological system.

We have employed multiple analytical strategies in most of our studies, and although some of these approaches are cumbersome, slow, costly, and quantitatively imprecise, the objective has always been to obtain the most comprehensive coverage of the proteome as is possible. Gel-based strategies are pivotal in our discovery platform because these allow separation of intact proteins. It is increasingly evident that this is important because post-translational and post-transcriptional variants are ubiquitous in eukaryotic systems and include phosphorylation, glycosylation, sulfation, ubiquitination, truncation, and alternative splicing (27). These modifications frequently transform a single gene product into multiple variants that are chemically and functionally distinct.

1. Analysis of Human Seminal Plasma. The need to adopt alternative strategies is well illustrated in our

studies of human seminal plasma (26). In this work the peptide and protein components of pooled human seminal fluid were identified by a combination of techniques including gel electrophoresis (1D and 2D), MALDI-ToFMS, and LC-MS/MS. Over 100 unique protein and peptide components of normal human seminal fluid were identified, including over 20 distinct forms of prostate-specific antigen (PSA). This, of course, raises questions about the suitability of existing tests aimed at only one epitope as a biomarker. The epitope that is targeted and the specificity of any PSA antibody may be critically important because the form of PSA that is biologically relevant in specific diseases, such as prostate cancer, is yet to be defined. Published PSA assays often involve the use of a cocktail of antibodies targeted at different forms of PSA, but in the absence of information about the diagnostic utility of each form, this strategy may be a serious compromise. Complete characterization of these different molecular forms is required so that investigators can establish the specific forms of PSA (or other proteins) that might serve as the best biomarkers of health and/or disease. This approach

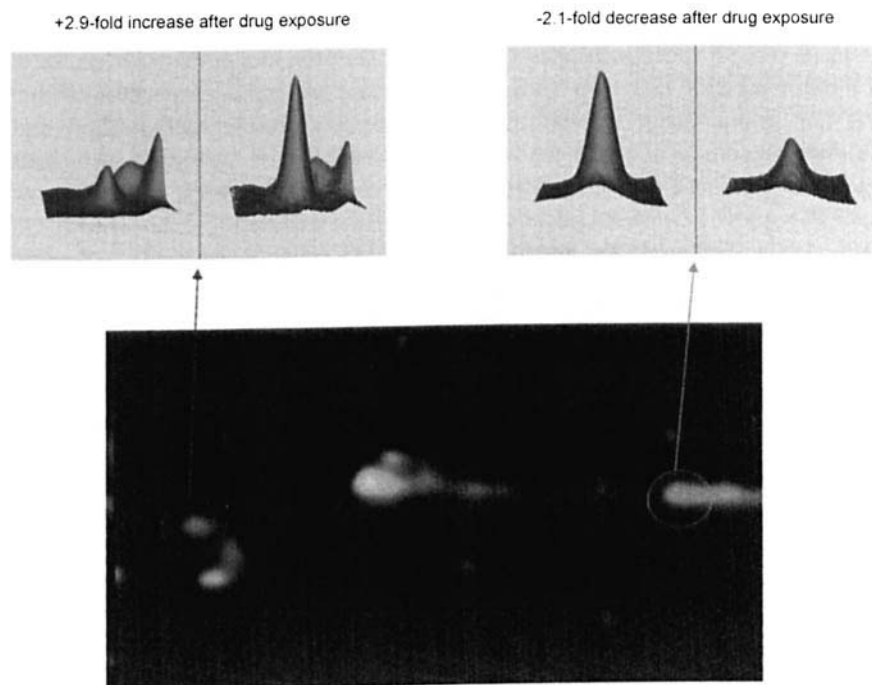


Figure 3. A pseudocolor difference gel electrophoresis image from the comparison of a non-small-cell lung cancer cell line before and after exposure to a therapeutic agent. Each of the circled spots was identified by mass spectrometry as triosephosphate isomerase. The more acidic isoform (left) was upregulated after exposure, while the more basic isoform (right) was downregulated after exposure.

has the potential to significantly enhance both the specificity and sensitivity of current diagnostic tests.

2. Analysis of Human Urine. In a study of the proteins in human urine, we identified a heparan sulfate proteoglycan fragment on 2D gel analysis at an apparent molecular weight of ~ 20 kDa (Hunsucker and Duncan, unpublished observations). Figure 2 shows the peptide mass map obtained by MALDI-ToFMS. This indicates almost complete sequence coverage of the C-terminal fragment. The identity of the spot on the gel was also confirmed by LC-MS/MS analysis of the same tryptic digest used for MALDI-ToFMS analysis. Although the intact protein has a molecular weight of 450 kDa, we could account for only AA 4224–4391 (4391 being the C-terminal amino acid). The predicted molecular weight of this region of the protein is 17.9 kDa, consistent with its mobility on the gel. By using multiple strategies (2D gels, MALDI-ToFMS, and LC-MS/MS), we were able to demonstrate that there is a heparan sulfate proteoglycan C-terminal fragment present as a component of normal human urine. LC-MS/MS alone (shotgun) would not have been able to determine the intact molecular weight of this species in the original urine sample. Several other groups have found this protein in normal urine, two by shotgun approaches (25, 28) and the other by 2DGE (29), but none of these reports made any mention of truncation.

3. Analysis of Human Tear Fluid. Detailed MALDI-ToFMS and LC-MS/MS analyses of human tear fluid demonstrated that lacrimal proline-rich protein (PRP) is present as its N-terminal portion (residues 18–121), together with a series of proteolytic fragments derived from

the C-terminus (residues 122–134) (30). These data are consistent with *in vivo* cleavage of the intact protein to yield several C-terminal peptides and formation of a truncated protein. The accuracy with which these masses was determined forced us to conclude that the N-terminal pyroglutamic acid of the protein is cleaved from the secreted form of lacrimal PRP. The complexity of this scenario would have been missed completely if a single analytical strategy had been employed.

4. Analysis of Human Cell Lines. A final example is taken from a recent study of human non-small-cell lung cancer (NSCLC) cell lines. In this study aimed at identifying novel markers of a patient's sensitivity or resistance to a specific line of therapeutic intervention, we compared the same cell line before and after exposure to the drug. We found several examples of differential regulation of protein isoforms after treatment. Figure 3 shows one such example where we picked and identified both of the circled spots as the same protein, triosephosphate isomerase. The more basic isoform was downregulated following treatment; the more acidic isoform was upregulated (Hunsucker, Solomon, and Duncan, unpublished observations).

These examples illustrate the benefits associated with adopting multiple strategies, and, in addition, they caution that shotgun approaches that incorporate proteolysis before protein separation do not deliver information about protein isoforms because they reduce the complexity of the proteome to that of the genome.

Quantitative Considerations. As a component of the discovery phase, it is necessary not only to identify proteins but also to be able to quantify the differences in levels

between two or more populations (e.g., control vs. disease). In our own work we have placed special emphasis on precise quantification of intact proteins and their variants. The primary quantitative tool in our discovery platform is DIGE. This approach allows separation of intact proteins, and, consequently, protein isoforms can be independently identified and quantified. This proves to be a critical step because of the propensity for post-translational modifications to convert a single gene product into multiple functionally and chemically distinct entities. In addition, DIGE offers precise relative quantitative comparisons between two samples, and, when combined with mass spectrometry, the differentially expressed proteins can be targeted and identified. DIGE alone can only provide quantitative information on proteins; spots have to be excised from the gels, digested with trypsin, and then analyzed by MALDI-ToFMS and/or LC-MS/MS for identifications to be made.

Protein Identification. MALDI Versus LC-MS/MS. Protein identification from simple mixtures or pure proteins (i.e., excised gel bands) based on MALDI-ToFMS (i.e., mass mapping or mass fingerprinting) is fast, reliable, and easily automated. We download raw mass spectrometry data, de-noise, baseline correct, calibrate, detect peaks based on signal-to-noise ratio, and match the resulting peak lists against theoretical mass maps, all in a batch processing format. This strategy allows hundreds of proteins to be identified in several hours without the need for operator intervention. Automated processing does not obviate the requirement for careful review of the data, including visual inspection of the spectra and the search results to ensure validity. The crucial elements required for data processing are a reliable primary sequence database, accurate calibration of the mass spectrum, accurate peak detection (i.e., defining peak signal-to-noise ratio), accurate monoisotopic mass assignment, and contaminant peak removal. Algorithms that set an intensity threshold for peak detection are not amenable to automation because spectra have variable background and noise levels. Although most protein identifications yield to MALDI-ToFMS analysis, in rare instances (<10% in our hands) there is insufficient information available to make a match, or confounding peaks are present in the peptide mass fingerprint. In these instances we adopt LC-MS/MS because it has the advantage that it employs chromatography to separate peptides before ionization, and, increasingly, LC-MS/MS offers sensitivity advantages over MALDI-ToFMS. Although more time consuming, determining the components of a mixture by LC-MS/MS returns more comprehensive information (i.e., peptide sequence coverage) on portions of the protein(s).

Protein identification from complex mixtures (e.g., *via* shotgun methods) is a much more challenging task, and currently there is no standard method for the analysis and validation of mass spectrometric data. Some investigators have begun to think about guidelines and standards, but these are yet to be put into universal practice (31). Further,

even when these guidelines are followed, the false-positive rate for protein identification can be high, particularly in the absence of expert inspection of the data (32). There are numerous protein identification algorithms and multiple approaches for assessing the statistical significance of protein identifications, but in the wrong hands these can generate misleading data (33–36). Strategies based on LC-MS/MS are unquestionably powerful, but the dual processes of data generation and interpretation are far from trivial.

There is a need for improved software that allows for reliable automated identification of proteins from mass spectrometric data. Automation will markedly reduce the need for operator intervention and, if properly implemented, will reduce user error and bias. Unfortunately, however, the analysis of data generated in a proteomics study remains the major cause of delay, frustration, and errors for many research groups.

The Role of Mass Spectrometry Post-Discovery. Most biomarker discovery work stops short of validation because the processes and tools used for discovery are very different to those required subsequently. However, there is an urgent need for *general strategies* for the quantification of proteins that can be applied in the biomarker validation phase, and this area has received relatively little attention. The current mind-set is that mass spectrometry has no place past the discovery phase, but this may be shortsighted.

We and others have been active in applying mass spectrometry to the quantification of peptides and proteins in biological tissues and fluids. These methods can prove to be precise and sensitive and are capable of providing absolute protein levels. The most promising and powerful strategies incorporate sample cleanup steps, digestion with trypsin, thoughtful selection of internal standards (either structural analogs or stable isotope-labeled standards), and then mass analysis of the peptide mixture. The approach requires the careful selection of an internal standard and a proteolytic fragment that can be cleaved, isolated reproducibly in high yield, and measured over a broad dynamic range.

Our own strategies have employed MALDI-ToFMS (37–39); others have focused on LC-MS/MS (40–43). Kuhn and colleagues adopted nanoflow chromatography–tandem MS on a triple quadrupole mass spectrometer (MRM mode), and they reported this to be a powerful approach to prescreening candidate protein biomarkers in human serum before antibody and immunoassay development (43). Muddiman and colleagues have adopted a similar strategy (protein cleavage–isotope dilution mass spectrometry [PC-IDMS]), and they also concluded that PC-IDMS is a promising technique for quantifying proteins. They specifically note the potential of this approach to standardizing immunoassays, monitoring post-translational modifications and quantifying newly discovered biomarkers before the development and implementation of an immunoassay (40). Gygi and colleagues have also used isotopically labeled internal standards and a similar LC-MS/MS strategy for the

precise determination of protein expression and post-translational modification levels in cell lysates (42). An LC-MS/MS approach that can independently confirm the diagnostic potential of a target protein identified during the discovery phase is particularly attractive because it is not based on antibodies of ill-defined specificity that can be both costly and time consuming to acquire. LC-MS/MS can be quickly implemented and offers specificity and sensitivity that is more than adequate for most circumstances.

The place for mass spectrometry further along the validation and implementation pipeline is less obvious. Although MALDI-ToFMS, 2DGE and LC-MS/MS are powerful tools to probe the complexity of a biological sample at the discovery phase, with few exceptions they are suboptimal for subsequent steps. During the discovery phase, generic conditions for protein isolation, separation, and mass analysis are employed in an attempt to gain comprehensive coverage of the proteome. However, under these conditions no single protein is measured cost effectively or with optimal speed, sensitivity, or precision. Once a target analyte (or set of analytes) is identified, higher-throughput, more cost-effective and precise analytical methods are required for routine testing. Some investigators have enthusiastically proposed the application of mass spectrometry throughout the whole process, but there are several fundamental reasons for why this is rarely, if ever, the best choice.

As the molecular weight of a compound increases, its physical properties begin to place limits on the utility of the mass spectrometer to serve as a sensitive and precise quantitative tool. As mass increases, the natural isotopic envelope distributes the total ion current for a given population of molecules across multiple species. (Or at low resolution, the outcome is exceedingly broad peaks.) Even at the mass of insulin (i.e., 5730 daltons), the natural isotopic distribution extends across 10 separate detectable species. At higher molecular weights (e.g., ~50,000 daltons), the natural isotope envelope is distributed across more than 25 distinct species. This is not a result of an instrumental limitation but is a fundamental property of biomolecules and is therefore insurmountable. The distribution of the signal across so many distinct species dramatically compromises the achievable limit of detection, confounds the potential to distinguish between closely related species of similar mass, and complicates the selection of an internal standard.

Contrary to common dogma, mass spectrometry is far from ideal for intact protein quantification and techniques based on immunorecognition can offer superior "selectivity," excellent precision, and high sensitivity. There are very few instances where mass spectrometry is being applied to the quantification of proteins, especially in a routine setting, and, where it has been employed, it invariably involves proteolysis and subsequent analysis of specific peptide fragments (39–42). In stark contrast, assays for a wide array of proteins are routinely and cost effectively performed by ELISA methods with both accuracy and precision, and,

further, the hardware necessary to perform these assays is easy to use, affordable, and ready available. In fact, immunoassays are the backbone of routine clinical pathology worldwide. The most significant advantage of mass spectrometric approaches is that they do not require the generation of specific antibodies and therefore can be developed expeditiously. Therefore, protein quantification by mass spectrometry might sometimes offer special advantages, such as enhanced specificity (39), but immunoassays deliver the most cost-effective, precise, accurate, sensitive, and transportable approach to protein determinations in almost all instances.

Conclusions

It is perhaps worth stating the obvious here: The process of biomarker discovery is complex and time consuming. Consequently, too much emphasis is often placed on "high throughput," as increased speed comes at a price. Dramatic decreases in the analysis time invariably involve a sacrifice in either (or both) qualitative accuracy or quantitative precision, thereby compromising the discovery process. Proteomics has delivered hundreds if not thousands of putative biomarkers over the past decade, but precious few have moved from discovery to development, to validation, and on to application. Unearthing one or two "needles" in the disease-specific protein "haystack" that can serve as sensitive and specific biomarkers of disease is inherently slow and costly and, as we have discussed, is best performed by adopting several complementary analytical strategies. While we might like it to be otherwise, gaining detailed (and accurate) qualitative and quantitative information on thousands of constituents comes at the expense of speed, and shorthand analysis during the discovery phase is likely to generate compromised data—and false leads—that ultimately impede the overall process.

We have discussed our strategy for biomarker discovery and presented some of our own data to illustrate key practical concerns. Currently there are biomarker discovery studies under way at many centers worldwide, and almost all of these are employing different approaches. While any specific discovery approach can be rationalized, there is no standard analytical strategy, nor should there be. Different approaches to protein isolation select for a distinct subset of the proteome (e.g., acid, basic, low molecular weight, high molecular weight), and different separation and detection methods offer unique advantages. Each combination of methods provides a different view of the proteome and has the potential to yield valuable insights and promising biomarkers. This diversity in experimental approaches during the discovery phase is highly desirable and will ultimately help to reveal the complexity of the proteome and provide the largest sample set of potential biomarkers for further investigation. Thereafter, carefully designed and rigorously controlled validation studies will establish the clinical utility of each candidate biomarker, either when employed alone or in combination with others. Despite a

decade or more of active research and considerable hype along the way, we are only now beginning to see candidate markers make their way through the whole process. Current indications are that proteomics is beginning to deliver on its promise to provide sensitive and specific disease biomarkers that can significantly improve our ability to diagnose and treat a broad spectrum of diseases.

The authors thank their colleagues at the University of Colorado at Denver and the Health Sciences Center for assistance with generating some of the data presented in this submission.

- Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 1:845–867, 2002.
- Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 69:89–95, 2001.
- Tang N, Tomatore P, Weinberger SR. Current developments in SELDI affinity technology. *Mass Spectrom Rev* 23:34–44, 2004.
- Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, Denoyer L, Fleisher M, Robbins RJ, Tempst P. Correcting common errors in identifying cancer-specific serum peptide signatures. *J Proteome Res* 4:1060–1072, 2005.
- Sheehan KM, Calvert VS, Kay EW, Lu Y, Fishman D, Espina V, Aquino J, Speer R, Araujo R, Mills GB, Liotta LA, Petricoin EF III, Wulfkuhle JD. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. *Mol Cell Proteomics* 4:346–355, 2005.
- Koomen JM, Shih LN, Coombes KR, Li D, Xiao LC, Fidler IJ, Abbruzzese JL, Kobayashi R. Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin Cancer Res* 11:1110–1118, 2005.
- Sidransky D, Irazary R, Califano JA, Li X, Ren H, Benoit N, Mao L. Serum protein MALDI profiling to distinguish upper aerodigestive tract cancer patients from control subjects. *J Natl Cancer Inst* 95:1711–1717, 2003.
- Caldwell RL, Caprioli RM. Tissue profiling by mass spectrometry: a review of methodology and applications. *Mol Cell Proteomics* 4:394–401, 2005.
- Chaurand P, Schwartz SA, Caprioli RM. Assessing protein patterns in disease using imaging mass spectrometry. *J Proteome Res* 3:245–252, 2004.
- Coombes KR, Morris JS, Hu J, Edmonson SR, Baggerly KA. Serum proteomics profiling—a young technology begins to mature. *Nat Biotechnol* 23:291–292, 2005.
- Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359:572–577, 2002.
- 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* 97:307–309, 2005.
- Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. *J Natl Cancer Inst* 97:315–319, 2005.
- Robbins RJ, Villanueva J, Tempst P. Distilling cancer biomarkers from the serum peptidome: high technology reading of tea leaves or an insight to clinical systems biology? *J Clin Oncol* 23:4835–4837, 2005.
- Villanueva J, Tempst P. OvaCheck: let's not dismiss the concept. *Nature* 430:611, 2004.
- Unlu M, Morgan ME, Minden JS. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* 18:2071–2077, 1997.
- Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, Lewis S, Currie I. A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics* 3:36–44, 2003.
- Wolters DA, Washburn MP, Yates JR. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 73:5683–5690, 2001.
- Griffin TJ, Han DK, Gygi SP, Rist B, Lee H, Aebersold R, Parker KC. Toward a high-throughput approach to quantitative proteomic analysis: expression-dependent protein identification by mass spectrometry. *J Am Soc Mass Spectrom* 12:1238–1246, 2001.
- Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* 19:946–951, 2001.
- Smolka MB, Zhou H, Purkayastha S, Aebersold R. Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. *Anal Biochem* 297:25–31, 2001.
- 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* 3:1154–1169, 2004.
- Cahill DJ. Protein and antibody arrays and their medical applications. *J Immunol Methods* 250:81–91, 2001.
- Sullivan Pepe M, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 93:1054–1061, 2001.
- Cutillas PR, Chalkley RJ, Hansen KC, Cramer R, Norden AG, Waterfield MD, Burlingame AL, Unwin RJ. The urinary proteome in Fanconi syndrome implies specificity in the reabsorption of proteins by renal proximal tubule cells. *Am J Physiol Renal Physiol* 287:F353–F364, 2004.
- Fung KY, Glode LM, Green S, Duncan MW. A comprehensive characterization of the peptide and protein constituents of human seminal fluid. *Prostate* 61:171–181, 2004.
- Kettman JR, Coleclough C, Frey JR, Lefkowitz I. Clonal proteomics: one gene—family of proteins. *Proteomics* 2:624–631, 2002.
- Spahr CS, Davis MT, McGinley MD, Robinson JH, Bures EJ, Beierle J, Mort J, Courchesne PL, Chen K, Wahl RC, Yu W, Luethy R, Patterson SD. Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionated tryptic digest. *Proteomics* 1:93–107, 2001.
- Smith G, Barratt D, Rowlinson R, Nickson J, Tonge R. Development of a high-throughput method for preparing human urine for two-dimensional electrophoresis. *Proteomics* 5:2315–2318, 2005.
- Fung KY, Morris C, Sathe S, Sack R, Duncan MW. Characterization of the in vivo forms of lacrimal-specific proline-rich proteins in human tear fluid. *Proteomics* 4:3953–3959, 2004.
- Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. 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* 3:531–533, 2004.
- Cargile BJ, Bundy JL, Stephenson JL Jr. Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res* 3:1082–1085, 2004.
- Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* 75:768–774, 2003.
- Sadygov RG, Cociorva D, Yates JR III. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 1:195–202, 2004.
- Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74:5383–5392, 2002.

36. Clauser KR, Baker P, Burlingame AL. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* 71:2871–2882, 1999.
37. Bucknall M, Fung KY, Duncan MW. Practical quantitative biomedical applications of MALDI-TOF mass spectrometry. *J Am Soc Mass Spectrom* 13:1015–1027, 2002.
38. Cerpa-Poljak A, Lahnstein J, Mason KE, Smythe GA, Duncan MW. Mass spectrometric identification and quantification of hemorphins extracted from human adrenal and pheochromocytoma tissue. *J Neurochem* 68:1712–1719, 1997.
39. Helmke SM, Yen CY, Cios KJ, Nunley K, Bristow MR, Duncan MW, Perryman MB. Simultaneous quantification of human cardiac alpha- and beta-myosin heavy chain proteins by MALDI-TOF mass spectrometry. *Anal Chem* 76:1683–1689, 2004.
40. Barnidge DR, Goodmanson MK, Klee GG, Muddiman DC. Absolute quantification of the model biomarker prostate-specific antigen in serum by LC-MS/MS using protein cleavage and isotope dilution mass spectrometry. *J Proteome Res* 3:644–652, 2004.
41. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* 100:6940–6945, 2003.
42. Kirkpatrick DS, Gerber SA, Gygi SP. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* 35:265–273, 2005.
43. Kuhn E, Wu J, Karl J, Liao H, Zolg W, Guild B. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and ^{13}C -labeled peptide standards. *Proteomics* 4:1175–1186, 2004.