Editorial

Cancer Proteomics: In Pursuit of ''True'' Biomarker Discovery

Zhen Zhang and Daniel W. Chan

Center for Biomarker Discovery, Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, Maryland

For the past several years, we have seen an increasing number of reports on the clinical application of proteomics research for risk assessment, diagnosis, prognosis, and management of cancer. These reports have generated a high level of excitement. At the same time, they have attracted critical reviews that have pointed out the shortcomings of the early studies. These shortcomings included technical limitations in the sensitivity of detecting low abundant biomarkers (1, 2) and possible systematic biases in the observed data (3-5), which in turn, cast doubts on the claimed clinical performances. Partly influenced by these reviews, the validity of the current approach of ''discovery-based'' research in clinical proteomics has come under question.

Proteomic analysis of clinical samples from actual patient populations for biomarker discovery involves a set of unique issues that are quite different from those in basic science research involving only controlled experiments and *in vitro* cultured samples or specimens from inbred animals. Here, we first review those issues that might have contributed to the problems in the early studies. This will allow us to better understand the limitations and capabilities of current technologies, and to devise constructive measures to alleviate the impact of these issues in a rational and scientifically sound manner. The ultimate goal of cancer proteomics and biomarker discovery is to discover truly disease-associated biomarkers with relevant clinical utilities and to gain insight into the biology of the disease process.

Issues in Cancer Proteomics and Biomarker Discovery

There are two major facts about cancer proteomics studies for biomarker discovery: (*a*) the necessary use of human samples from diverse populations and collected under varying clinical conditions; and (*b*) the predominant use of case-control study design with retrospective samples for discovery phase studies. The issues described below are the direct implications of these two facts.

Complexity of Clinical Specimen Proteome. The large number of proteins in clinical specimens such as serum and plasma is only one aspect of the complexity of human proteome. In a recent study by the Plasma Proteome initiative of the Human Proteome Organisation, the dynamic range of plasma protein concentration was found to be in the order of 10^{10} for many known proteins (6). Currently, it remains impractical for any single technology platform to consistently cover such a wide concentration range. Although improvement in detection sensitivity is important (and there has been steady progress by instrument manufacturers), it might be more important to develop methods and associated hardware for sample preprocessing/fractionation. This will allow the analysis of subproteomes that are less complex and with narrower dynamic ranges. Although such technologies have been in development, their ability to produce consistent results (i.e., a low coefficient of variance in repeated experiments) has not been sufficiently shown or documented. A low coefficient of variance in the preprocessing steps is a critical requirement for clinical proteomics where other sources of variability are already significant. The tradeoff between better sensitivity/resolution in protein detection and the introduction of additional analytical variability needs to be balanced in order to achieve the overall beneficial outcome.

The complexity of human specimens is also a reflection of the dynamic nature of the human proteome. The expression, modification, and processing of proteins could be altered significantly by many non–disease-associated events and change from one moment to another. For instance, a commonly used description of clinical specimens is that samples were ''collected prior to surgery.'' In fact, depending on the time between specimen collection to surgery, the patients' knowledge of impending surgery itself could change expression of many proteins, a change that will not be observed among the otherwise well-matched controls.

Biological Variability. Unlike samples from inbred animal models, the biological variability of protein expressions in samples from diverse human populations could be very significant. In addition, many current disease classifications are actually collections of multiple known and unknown subphenotypes with distinct disease pathways. The resultant disease-associated proteomic expression profiles could be very different from one another.

Preanalytical Variability. Retrospective use of clinical samples shortens the overall study duration and is often acceptable in discovery studies. For certain diseases such as ovarian cancer, due to the low prevalence in the general population and/or lack of current effective detection modality, the use of retrospective samples becomes necessary. However, such archived specimens are often collected under different protocols for different purposes. The sample handling and processing could be different for the disease and control samples.

Analytical Variability. The desire to look deeper into the complex proteome of clinical specimens often competes with the need to minimize variability introduced by the analytical procedures. Sample preprocessing techniques, such as those based on multidimensional fractionation and protein depletion improve sensitivity in detecting proteins that are otherwise undetectable. However, until the additional variability introduced by these sample preprocessing steps can be quantified and controlled at an acceptable level, the benefit of sample preprocessing has to be assessed against the possible problem of processing-introduced variability in data.

Without proper design, many of the variables that contribute to the preanalytical and analytical procedures could be confounded with sample group designations. Such systematic biases in proteomic expression data between disease and

Cancer Epidemiol Biomarkers Prev 2005;14(10):2283–6

Requests for reprints: Zhen Zhang, Center for Biomarker Discovery, Department of Pathology, Johns Hopkins Medical Institutions, 419 North Caroline Street, Room 200, Baltimore, MD, 21036.

doi:10.1158/1055-9965.EPI-editorial

control samples could, in some cases, render the final results unusable. Case-control study design and supervised methods are often used for biomarker discovery. The knowledge of the sample classes is used explicitly to search for markers with differential power; it guarantees that proteins whose concentration or even presence is affected by preanalytical or analytical variability will be selected incorrectly as the top candidate biomarkers with an extremely high discriminatory power.

Measures to Alleviate the Impact of Possible Biases

In Fig. 1, we outline a basic workflow of clinical proteomics research for biomarker discovery that we have been using for the past several years. Even though these steps are important in improving the overall odds of discovering truly disease-associated biomarkers, in the following discussion, we will emphasize some of these steps where practical measures may be taken to alleviate the impact of the abovementioned possible biases.

To Have A Clearly Defined Clinical Utility for the Target Biomarkers. It is without question that most researchers understand that ''garbage-in garbage-out'' also applies to the selection of clinical specimens for biomarker discovery. It is less clear what exactly constitutes a ''good'' clinical sample set. Too often, a clinical proteomic project begins with a convenient set of disease and control samples marked by a generic classification such as ''breast cancer,'' ''benign,'' and ''control.'' This is often viewed as less of an issue because the discoveries will have to be further validated using more rigorously constructed sample sets. However, the success rate of such a discovery-first-worrying-about-validation-later approach has been quite dismal, as evidenced by the number of biomarkers used clinically in comparison with the biomarkers reported in many issues of clinical and basic science journals.

Because the ultimate goal of biomarker discovery is to translate discoveries into clinical applications, it would be more cost-effective to design a biomarker discovery study by starting with a clearly defined utility for the biomarkers that is both clinically relevant and practical. This definition will then determine the appropriate composition of sample population.

As an example, with the introduction of the prostate-specific antigen test, the clinical presentation of prostate cancer patients: their initial serum prostate-specific antigen levels and tumor stage and grade, have significantly and forever changed. Thus, samples collected prior to the introduction of the prostate-specific antigen test would be much less relevant than those from more contemporary patient populations for a biomarker discovery study intended to detect prostate cancer in men with a modestly elevated prostate-specific antigen level.

The Power of Multicenter Study Design. To eliminate the systematic biases in samples, one would ideally like to have samples collected prospectively and, better yet, in such a way that the sample collection and processing procedures are done

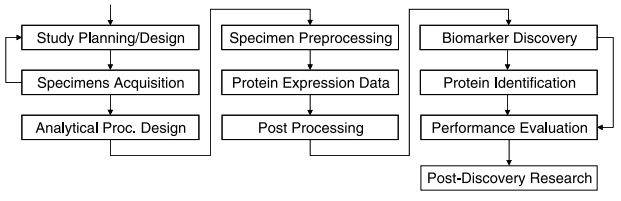

Figure 1. Workflow of cancer proteomics research for biomarker discovery.

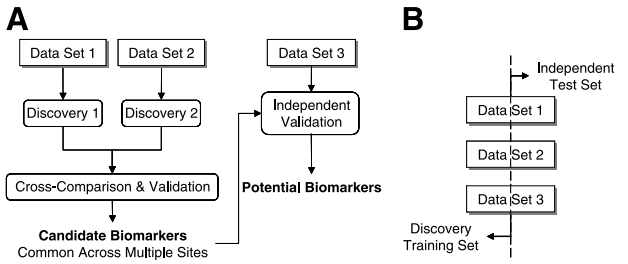

Figure 2. Two different designs for the use of multicenter sample sets for biomarker discovery and derivation of multivariate predictive models. **A,** a proposed design in which samples from independent sites are used separately and independent in discovery; the results are then cross-compared and validated to reach a common set for further validation using independent samples from additional sites. **B,** a traditional design in which samples from multiple sites are pooled first and then divided into a discovery/training set and a test/validation set.

without the knowledge of the disease status of the human subjects. However, for diseases such as ovarian cancer, which are of a relatively low prevalence and without an effective screening test, it is often impractical to expect to obtain such pristine sample sets. The design of many clinical proteomics and biomarker studies, therefore, has to deal with less than perfect conditions.

The observed expression level of a particular protein E_p is typically comprised of possible true (disease-associated) changes as well as the abovementioned added sources of variability. In a somewhat simplified way, it can be expressed as:

$$E_p = E_{disease} + E_{biological} + E_{preanalytical} + E_{analytical} + E_{err},$$

where E_{err} is random noise due to other sources of variability that have not been explicitly accounted for. In a single site sample set, if for some variables, their $E_{preanalytical}$ or $E_{analytical}$ components have systematic biases, it is impossible, from a statistical point of view, to separate them from variables with truly disease-associated changes ($E_{disease} \neq 0$).

In statistical inference which forms the basis of data analysis for biomarker discovery, a fundamental assumption is that the variables (potential biomarkers) satisfy the condition of being independently and identically distributed between the sample sets used for biomarker discovery and the populations to which the discovered biomarkers are targeted. In reality, many sample sets currently used for biomarker discovery do not strictly satisfy the independently and identically distributed condition. Variations in expression data due to site-specific differences in samples are commonly observed in both proteomics and genomics studies, as well as in the extension to future large populations. However, we believe that one should actually be able to take advantage of such site-to-site differences. We hypothesize that site-to-site variations affect only the non–disease-related components of the observed expression data and the disease-associated component $E_{disease}$ will, to a large degree, satisfy the independently and identically distributed condition across data sets from multiple sites. Of course, the assumption that a disease is the same from site to site is almost certainly not true under all circumstances, e.g., the pattern of molecular subtypes of colorectal cancer differs between Japan and the U.S. Nonetheless, under this hypothesis, one could design a multicenter biomarker discovery study in which data from different sites are triangulated to identify the true biomarkers against a noisy and often biased background of a large number of candidate biomarkers (Fig. 2A). In this design, sample sets from a minimum of two independent sites

are used separately and independently in the data analysis stage to generate individual lists that rank candidate biomarkers. Only those candidates that are common among lists from multiple sites will be further investigated and validated using samples from additional independent sites.

The multicenter design in Fig. 2A is quite different from the conventional approach where samples from multiple sites are pooled before being divided with randomization into a discovery set and a validation set (Fig. 2B). The conventional approach has both pros and cons. Its advantage is that the discovery set includes samples from all sites so that it is more diversified and representative. The disadvantage is that the discovery set and the validation set are artificially made to be identically distributed. It is possible that, with the sophisticated modern statistical learning algorithms, a multivariate predictive model can be derived based on non–disease-associated features in the data that works well in both the discovery/training set and the validation set.

Randomization in Analytical Procedures. In traditional biological experiments where a small number of samples from cultured cells or inbred animals are analyzed under well-controlled conditions, randomization is often not a major concern in study design. However, in clinical proteomics studies where uncontrolled and unknown variables could introduce biases to significantly affect the interpretation of comparisons, randomization becomes extremely important. In fact, a lack of appreciation of the need to randomize, to a large degree, contributed to the issues raised by critics of the early study results (3, 4). It should be pointed out that the effective use of randomization is more than randomizing the order in which samples are processed. It requires extensive knowledge of the analytical procedures and underlying physical processes to identify the factors over which to randomize—and then to make necessary compromises.

Incorporation of Knowledge of the Disease into Biomarker Discovery. It often takes knowledge to discover new knowledge. Clinical and epidemiologic knowledge about the disease of interest plays a critical role in defining the expected clinical use and performance criteria of the target biomarkers. The same knowledge should also be incorporated into the data analysis and discovery process. For example, prostate biopsy as the diagnostic gold standard for prostate cancer itself has a significant false-negative proportion (7). One should not expect a new biomarker for prostate cancer to have a perfect specificity among the biopsy-negative samples. In fact, perfect tracking of an imperfect target could be a sure indication only of the effect of non–disease-associated biases.

The Need to Separate Feature Selection (Biomarker Discovery) and Multivariate Predictive Model Derivation. Biomarker discovery (selection of the most informative features) from the volumes of raw expression data and the derivation of multivariate predictive models are traditionally done as two separate steps. For simplicity and because of the limitation imposed by sample size, feature selection is commonly done using univariate analysis based on some modified version of the *t* test. Results from such approaches tend to be fairly robust. However, the simplification ignores the fact that changes in a biological system are often multifactorial. This approach also runs the risk of overlooking candidate biomarkers that are useful in detecting less prevalent subphenotypes of a disease. Furthermore, a simplified up-front ''filter'' for informative features could limit the usefulness of a sophisticated nonlinear multivariate model.

In a number of recent reports, more complex nonlinear multivariate models were used for feature selection as well as the final predictive model. It needs to be pointed out that even though such ''one-step'' approaches may not have an explicit assumption of the data distribution, the type of biomarkers to be selected will be restricted by the type of multivariate models being fitted. For example, the so-called *n*-dimensional cluster analysis used in several early publications (8), because of its use of Euclidian distance in the feature space to form sphere-like clusters, the implicit assumption is that the samples followed multivariate Gaussian distribution in the selected feature space.

It is much more difficult in such an approach to control false discovery and model overfitting. In particular, the role of individual biomarkers in the model can be so convoluted that it may be difficult to explain in a biologically meaningful way. The scatter plot in Fig. 3 exemplifies this point with real data from one of our ovarian cancer studies. The data are from early stage ovarian cancer patients and healthy controls from two independent sites. The dashed line indicates that a simple quadratic classifier (e.g., a logistic regression with interaction terms) would be able to separate the cancer from the controls reasonably well. This would hold if we randomly divide the pooled samples into a training set and validation set. Upon closer examination, however, one would notice that whereas one of the markers (M1) was down-regulated among cancer samples for both sites, the second marker (M2) is up-regulated for one site and down-regulated for the other among the cancer samples. Without a clear biological explanation, one would have difficulty accepting results with such unexplained discrepancies.

We believe strongly that a pattern-based approach using multiple biomarkers will significantly improve the performance over that of individual biomarkers used alone. However, it is prudent to examine and validate the roles of individual biomarkers and biological relevancy using independent technological platforms and data set sites. This often requires protein identification of the selected biomarkers.

Postdiscovery Research: Why We Should Not Have to Choose between Profiling/Discovery and Hypothesis-Driven Research

Issues in clinical proteomics such as those discussed in this report, and the problems with some of the early studies, have been further fueling the longstanding debate on the merits of

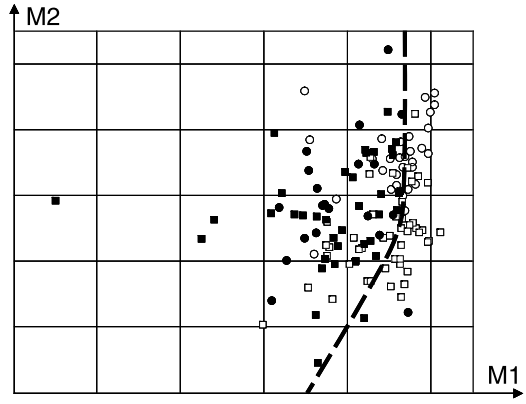

Figure 3. Scatter plot demonstrating the importance of examining the expression patterns of individual biomarkers across multiple sample sets. In the plot, stage I/II invasive ovarian cancer samples (■,●) and healthy controls (□,○) from two independent sites (●,○ versus ■,□) are plotted in two potential protein biomarkers, M1 and M2. A simple quadratic classifier can separate the cancer and control samples reasonably well (*dashed line*). Upon closer examination, M1 is down-regulated in cancer samples from both sites yet M2 is down-regulated in cancer samples from one site (●,○) and up-regulated in cancer samples from the other site (■,□).

high-throughput profiling and discovery research (9, 10). Much of the argument would be less contentious if one took the view that results from high-throughput discovery are not the end of research: they rather serve as potential leads for the beginning of new discovery research, much of it hypothesis-driven.

With protein identification of the discovered biomarkers, post-discovery research, such as immunoprecipitation/pull-down followed by mass spectrometry analysis, will allow the identification of interacting partners of the biomarkers and pathways in which the biomarkers may function. Through this process, in addition to gaining a better understanding of disease biology, we could also discover additional biomarkers that possess better characteristics for clinical applications than the initial discovery.

Conclusions

High-throughput measurement of protein expressions in complex clinical samples has only become possible very recently. Sophisticated bioinformatics tools help in the discovery of informative markers and their combination in multivariate predictive models. A critical point is that the information carried by these markers has to be truly disease-associated. This can only be achieved by incorporating biological, clinical, and epidemiologic knowledge of the disease into the total process of study design, analytical processing, and data analysis and discovery.

Currently, proteomic analysis of complex crude clinical samples is still technically challenging. The development of efficient and, much more importantly, consistent (i.e., low analytical variability) sample preprocessing and fractionation methods is urgently needed. Such technical limitations should encourage us to innovate and should not be used as the reason to reject the overall approach.

Finally, overwhelmed by the large number of reports and at the same time the sharp criticisms, ''does it work?'' has become a common question asked by many. Specific experiments and studies have been planned to answer this question. Such studies will undoubtedly provide invaluable information. However, one also needs to answer this question from a more historical perspective. For any emerging technology, whether it works or not at the beginning is not as important as understanding whether the physical principles behind the new technology are sound. If the answer is yes, the imperfections will eventually be resolved (one does not need to go back far to find such examples in medicine). Right now, we should try to answer the more relevant question of ''will it work?'' If the evidence supports it, then we should let scientific discovery and technological innovation have the opportunity to proceed.

References

1. Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst 2004;96:353–6.
2. Diamandis EP. Re: Serum proteomic patterns for detection of prostate cancer. J Natl Cancer Inst 2003;95:489–90.
3. Baggerly KA, Edmonson SR, Morris JS, Coombes KR. High-resolution serum proteomic patterns for ovarian cancer detection. Endocr Relat Cancer 2004;11:583–4.
4. 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;97:307–9.
5. Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. J Natl Cancer Inst 2005;97:315–9.
6. Omenn GS. The Human Proteome Organization Plasma Proteome Project pilot phase: reference specimens, technology platform comparisons, and standardized data submissions and analyses. Proteomics 2004;4:1235–40.
7. Fleshner NE, O'Sullivan M, Fair WR. Prevalence and predictors of a positive repeat transrectal ultrasound guided needle biopsy of the prostate [discussion 8–9]. J Urol 1997;158:505–8.
8. Rosenblatt KP, Bryant-Greenwood P, Killian JK, et al. Serum proteomics in cancer diagnosis and management. Annu Rev Med 2004;55:97–112.
9. Allen J. In silico veritas: data-mining and automated discovery: the truth is in there. EMBO Rep 2001;2:542–4.
10. Smalheiser NR. Informatics and hypothesis-driven research. EMBO Rep 2002;3:702.