

The “omics” Haystack: Defining Sources of Sample Bias in Expression Profiling

Great medical benefit may result from biomarker discovery, but the scarcity of useful biomarkers among myriad genes and proteins makes this task every bit as daunting as finding the needle in a haystack. At the moment we are not even certain what the needle or the haystack looks like (although many of us are certain we know one when we see one) or, more precisely, how to separate the signal from the noise. The need for better disease management tools has placed considerable demand on the scientific community to find appropriate clinical biomarkers, ushering in the “omics” era—the application of specific technologies such as proteomics, genomics, and metabolomics along with the mainstreaming of high-throughput, high-volume analytical approaches.

Clearly the development and implementation of novel technologies as well as innovative new application of “old” technologies is justified. However, this pushing of the technical envelope must be balanced with careful scientific evaluation of the performance characteristics of each new paradigm. The collision of these two imperatives has never been more apparent than in the current debate over protein expression profiling and pattern recognition-based diagnostics. The opposing forces of excitement associated with innovation (1, 2) and caution regarding bias, chance, and overgeneralization (3–5) must be balanced by the research community.

One case study for this issue is serum protein expression profiling. Following promising seminal work, many questions were raised after closer scrutiny of published data (3–5). Causes of concern included lack of analytical reproducibility, diminished robustness of discovered biomarkers during validation, and the fear that the prevalent detected serum proteins were produced by the liver. Indeed, the majority of these concerns can be attributed to bias, chance, and our rush to generalize results, a phenomenon nicely articulated in two recent articles (6, 7). The research community is called to strengthen its vigil over possible sources of bias, which can occur at many points along the discovery pathway. Although biostatisticians and epidemiologists bear the greatest responsibility for study design and data analysis, there are avoidable sources of experimental bias that must be recognized by laboratory scientists. Evaluation of analytical reproducibility and determination of sources of variability are essential steps in the biomarker discovery process. Identifying sources of sample bias introduced during clinical or laboratory processing allows for a greater understanding of the non-disease-related events that confound biomarker discovery.

Examining the influence of known sample processing variables on the spectral output after expression profiling analysis by mass spectrometry was the focus of the study by Rosamonde Banks and colleagues (8) in this issue. The authors introduced several changes in blood sample collection and processing and measured the resulting vari-

ability by use of surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS). Specifically, they assessed the impact of anticoagulant, types of serum collection tubes, and elapsed time between venipuncture and sample analysis. The test samples were processed by use of various affinity-activated surfaces immediately before mass spectral analysis. As has been suggested elsewhere (9), plasma types and serum separator tube choice appeared to have a profound impact on the spectra observed. This variability was further confounded by the selected affinity surface, confirming a phenomenon that has been observed before but never demonstrated under controlled conditions (10). Of particular interest is the observation that the time from blood collection to analysis is a critical period in which changes in protein profiles occur. Whereas most researchers would assume that the quicker the serum sample is processed the better, these authors observed that a sample stabilizes following a 30-min period. Thus, the recommendation for protein expression profiling of blood samples would include a provision that analysis should not be performed until after an initial 30-min period. Clearly the skeptical researcher must be aware that this study investigated only a small number of many possible variables. Thus, although analysis after a 30-min stabilizing period will avoid these reported blood changes, there is no guarantee that analysis between 30 min and 4 h will avoid artifacts in all features of the serum proteome. Indeed, the results presented by Banks and colleagues (8) serve to underscore the need for further and more extensive examination of the constituents of the “haystack”.

When we attempt to understand the possible sources of sample processing variability it is useful to examine known biological phenomena that might give rise to observed events. The authors approach this question by examining the impact of the clotting process on the sample-specific spectra. Their analysis of the spectral profile showed alterations of many fairly prominent peaks corresponding with *in vitro* manipulation of platelets. The authors list several *m/z* peaks that were altered in their system, providing a useful catalog for the research community, but the strongest take-home message here is that these sorts of perturbation studies should be routinely conducted as a means of pinpointing possible confounding events. The knowledge that biological phenomena such as platelet activation can profoundly influence observed mass spectral output demands that “omics” researchers examine study protocols for population characteristics that would affect the platelet activation process.

A benefit of studies designed to discover sources of experimental bias is the direct definition of confounders that lead to incorrect associations and erroneous conclusions. Such new knowledge will facilitate study design and data interpretation. Our efforts toward uncovering

the elusive protein biomarker in the proteome require a complete understanding of the proteomic haystack. The observed platelet-dominated changes seen in the unfractionated abundant proteins may or may not be relevant among the less abundant proteins, but selective removal of platelets before sample analysis might prove useful. Conceptually, one might reduce the impact of the haystack by targeting subproteomes. In this respect, biological perturbation studies may serve as nice systems for developing techniques that minimize the observation of non-disease-related events. The more we define the biological noise, the easier it will be for us to find the biologically relevant signal.

References

1. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res* 2002;62:3609–14.
2. 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;359:572–7.
3. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 2004;20:777–85.
4. Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. *J Natl Cancer Inst* 2005;97:315–9.
5. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003;4:24.
6. Hu J, Coombes KR, Morris JS, Baggerly KA. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic* 2005;3:322–31.
7. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005;5:142–9.
8. Banks RE, Stanley AJ, Cairns DA, Barrett JH, Clarke P, Thompson D, et al. Influences of sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. *Clin Chem* 2005;51:1637–49.
9. Drake RR, Cazares LH, Malik G, Corsica A, Schwegler E, Libby AE, et al. Quality control, preparation and protein stability issues for blood serum and plasma used in biomarker discovery and proteomic profiling assays. *Bioprocessing J* 2004;3:45–50.
10. Laronga C, Becker S, Watson P, Gregory B, Cazares L, Lynch H, et al. SELDI-TOF serum profiling for prognostic and diagnostic classification of breast cancers. *Dis Markers* 2003;19:229–38.

O. John Semmes

*Department of Microbiology and Molecular Cell Biology
Center for Biomedical Proteomics
Virginia Prostate Center
Eastern Virginia Medical School
700 W. Olney Road
Norfolk, VA 23508
E-mail semmesoj@evms.edu*

DOI: 10.1373/clinchem.2005.053405
