Journal of Molecular Diagnostics, Vol. 7, No. 2, May 2005

Commentary

Immunohistochemical Validation of Expression Microarray Results

Lawrence True* and Ziding Feng†

From the Department of Pathology, University of Washington, Seattle; and the Public Health Sciences Division,† Fred Hutchinson Cancer Research Center, Seattle, Washington*

Expression microarrays, which determine the level of expression of tens of thousands of mRNAs by a specific cell or tissue type, are powerful and increasingly more widely used investigative, diagnostic, and prognostic molecular biological tools.[1,2] However, there are technical aspects to using expression microarrays that can produce results erroneously representing either under- or overexpression of specific genes. Chuaqui et al[3] and Simon et al[4] have discussed some of these pitfalls. For example, false negativity can result from low expression levels, transcript drop-out (attributable to inefficient priming of specific mRNA(s)), poor adhesion of DNA to the slide, and splice variants with sequences not included on the array. Conversely, sources of false positivity include repetitive nucleotide elements, poly(A) tails, and sequence homology between functionally different transcripts, an inappropriately chosen reference standard, and high background levels due to nonspecific binding of nucleotides to the microarray slides. Ways to minimize the sources of error continue to be developed. For example, use of multiple different sequences of a given gene provides a way not to under-represent a given gene. Conversely, more rigorous attention to minimizing sources of background binding of detection nucleotides can minimize over-representation of highly expressed genes.

However, since these sources of error remain a potential source of confounding data, confirmation of expression microarray results before proceeding to undertake more elaborate, gene-specific experiments based on array results is important. A variety of different approaches have been used to validate expression microarray results, as discussed by Chuaqui and colleagues.[3] Confirmation of RNA levels can be based on precedent, ie, referral to already published work, or on experiments using independent methods of quantifying RNA and protein levels. Methods used to quantify RNA levels include real-time RT-PCR, Northern blot, ribonuclease protection assay, and *in situ* hybridization. Some of these techniques have disadvantages. For example, Northern blots require such large amounts of RNA, obtained from a large number of cells, eg, $>10^6$ cells, that one may question whether a Northern blot accurately represents the array results, eg, typically obtained from 10^3 cells, especially if the sample that is arrayed consists of a minor population of cells obtained by laser microdissection.

A method that is more likely to be relevant to the biology of a cell or tissue of interest is evaluation of protein expression levels. A variety of methods can be used to quantify protein content. These include Western blot, mass spectrometry, and immunohistochemistry. However, most of these techniques are limited to assaying fairly large amounts of material since proteins, in contrast to nucleic acids, cannot be amplified. The type of protein assay that is not limited by quantity of material is immunohistochemistry, which can be used to assess protein presence in a single cell. Furthermore, the label, either fluorochrome or optically dense reaction product, can be measured with great precision.[5] Given the great potential of immunostains to quickly assess protein expression in tissues, particularly when applied to such a high throughput platform as tissue microarrays,[6] defining a strategy to efficiently use immunostains to validate expression microarrays results is crucial. Although immunohistochemistry is widely used to validate expression microarray experiments, there has been little discussion about determining how many tissue samples and how many differentially expressed gene products should be assayed to confirm expression array results. The accompanying article by Betensky and colleagues[7] in this issue of *The Journal of Molecular Diagnostics* provides an imaginative, thoughtful approach to address the challenge of sample size when designing an immunohistochemical experiment. They provide a statistical model that, based on clearly stated assumptions regarding antibodies and immunostains, can be used to determine the minimum number of immunostains that would validate findings of an expression microarray at different levels of significance.

As readers of the Journal apply the methods of Betensky et al,[7] we would like to raise several points to consider when analyzing the application of their approach to

Accepted for publication February 22, 2005.

Address reprint requests to Lawrence True, Department of Pathology, Room EE110, 1959 NE Pacific St., Box 356100, University of Washington, Seattle, WA 98195. E-mail: ltrue@u.washington.edu.

the study of gliomas. The proposed model provides a range of scenarios of both significance and of strategies to quantify the immunostains. The model requires the investigator to make a series of assumptions regarding development of the immunohistochemical panel: 1) selection of genes, 2) optimization of the antibodies, 3) individual assay outcomes, and 4) comparability of subjects. Since assumptions 1 and 3 are crucial but less clear, it might be helpful to simulate more configurations for assumptions 1 and 3 so that readers understand the impacts on sample sizes and statistical powers when the underlying truth deviates from the assumed configuration. For example, if we accept 3+/4+ as positive immunohistochemistry values, assumption 3 indicates 90% sensitivity, which is a pretty strong association.

Biomarker discovery and validation is an iterative, progressive process. The authors clearly sketch a generally sound study strategy, illustrated in their Figure 1.[7] One consideration is that each time a new study is conducted using a model (classifier) from a previous study, the investigator should first validate the previous model. If the previous classifier holds, then there is more confidence in being able to refine that model using new data. If the previous model does not hold, there is less confidence in the performance of the newly built model because the model performance is examined using the same data set on which the model was built. Validation using an independent data set is a powerful tool to detect overfitting and bias from the previous study. Thus, validation with independent data should be performed whenever possible.

There should be caution in drawing conclusions from a simulation study. The simulation study design described by Betensky et al[7] specifies that half of the genes are differentially expressed. In the real world, such an assumption cannot be made with certainty. Therefore, selecting between 30 and 90 of the most differentially expressed genes, a simple ranking, may not be appropriate. The magnitude of the standardized differences, such as *t*-statistics and the receiver operating characteristic (ROC) curve, should also be considered in selecting candidate genes.[8] Beyond these considerations of the specific study described by Betensky et al,[7] this type of approach represents a valuable extension of rigorous statistical methods to the validation microarray experiments.

That said, there are several problems to using immunoperoxidase stains as quantitative tools that are rarely discussed. One significant limitation of immunohistochemical assays is that, basically, they are rarely stoichiometric.[9] Most immunohistochemical stains used to confirm RNA array results are three-step, enzyme-catalyzed reactions wherein an optically dense substrate is deposited on the cells expressing the antigen. Although there are multiple steps in this method at which antigen expression may be misrepresented, very few studies have examined the limits of stoichiometry.[10] Rarely is a standard curve run in the typical immunoperoxidase experiment, wherein the optical density of the reaction product is shown to reflect a given level of antigen concentration. Thus, a reaction product on specimen A that has twice the optical density (or intensity if the assay is an immunofluorescence stain) of a reaction product on specimen B does not necessarily mean that specimen A has a twofold greater level of protein than specimen B.

Another confounding aspect to immunoperoxidase studies arises from the fact that a wide variety of approaches are used to "quantify" immunoperoxidase stains. For example, one investigator may express the result of an immunoperoxidase study as a single numerical value calculated as IH stain intensity = Sum of ((% cancer cells staining intensely) $\times$ 3) + (% cancer cells staining moderately) $\times$ 2) + (% cancer cells staining faintly) $\times$ 1) + (% cancer cells not staining) $\times$ 0)), where the range of possible values is 0 to 300. Another investigator studying the same antigen in the same tissue may not consolidate the immunostain results into a single value. In addition, very few papers report interobserver variances in assessing immunoperoxidase stains. What investigator A interprets as "intense" immunoreactivity investigator B may interpret as "background" reaction product. Consequently, comparing results of different studies of expression of the same antigen is challenging. A consequence of observer variance in the visual assessment of optical density is that there is a high level of interobserver and interlaboratory variability in assessing immunostains.[11] Although digital cameras may decrease observer variance, other sources of variance, eg, setting the threshold of what is defined as "positive" and selecting the microscopic field to be analyzed, remain. We recommend reporting ROC curves to characterize the diagnostic performance of quantitative immunohistochemical biomarkers because, in addition to their relevance to clinical diagnosis and decision, different laboratories using different metrics will maintain the same ROC curve for the biomarker as long as they have consistent rankings.

In lieu of standardization of immunohistochemical methods and in lieu of attempts to control the reproducibility of immunoperoxidase stains confidence that immunostains accurately assess protein expression levels is muted. We can look forward to the day when more stoichiometric assays for protein expression in very small samples are developed. Until then, we will continue to strive for better validation of data from expression microarrays.

References

1. Nelson PS: Predicting prostate cancer behavior using transcript profiles. J Urol 2004, 172:S28–S33
2. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004, 351:2817–2826
3. Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, Flaig MJ, Hewitt SM, Phillips JL, Krizman DB, Tangrea MA, Ahram M, Linehan WM, Knezevic V, Emmert-Buck MR: Post-analysis follow-up and validation of microarray experiments. Nat Genet 2002, 32(Suppl):509–514
4. Simon R, Radmacher MD, Dobbin K, McShane LM: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003, 95:14–18

5. Rubin MA, Zerkowski MP, Camp RL, Kuefer R, Hofer MD, Chinnaiyan AM, Rimm DL: Quantitative determination of expression of the prostate cancer protein α-methylacyl-CoA racemase using automated quantitative analysis (AQUA): a novel paradigm for automated and continuous biomarker measurements. Am J Pathol 2004, 164:831–840
6. Bubendorf L, Nocito A, Moch H, Sauter G: Tissue microarray (TMA) technology: miniaturized pathology archives for high-throughput in situ studies. J Pathol 2001, 195:72–79
7. Betensky RA, Nutt CL, Batchelor TT, Louis DN: Statistical considerations for immunohistochemistry panel development following gene expression profiling of human cancers. J Mol Diagn 2005, 7:276-282
8. Pepe MS, Longton G, Anderson GL, Schummer M: Selecting differentially expressed genes from microarray experiments. Biometrics 2003, 59:133–142
9. True LD: Quantitative immunohistochemistry: a new tool for surgical pathology? Am J Clin Pathol 1988, 90:324–325
10. Gross, DS and JM Rothfeld: Quantitative immunocytochemistry of hypothalamic and pituitary hormones: validation of an automated, computerized image analysis system. J Histochem Cytochem 1985, 33:11–20
11. Paik S, Bryant J, Tan-Chiu E, Romond E, Hiller W, Park K, Brown A, Yothers G, Anderson S, Smith R, Wickerham DL, Wolmark N: Real-world performance of HER2 testing: National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer Inst 2002, 94:852–854