

What brown cannot do for you

David L Rimm

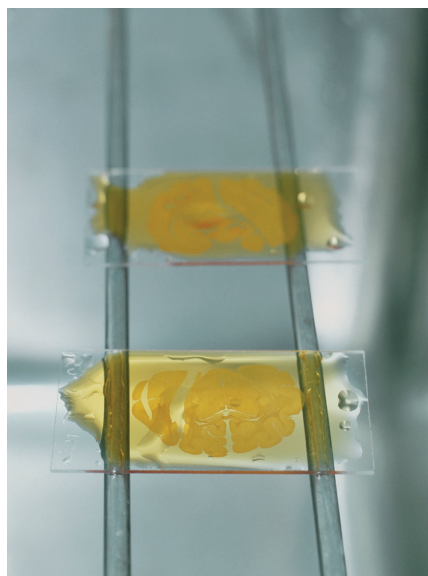
Chromogenic stains have long been used in immunodiagnostic assays, but fluorescence-based readouts could supplant them as emphasis shifts away from diagnosis to prediction by means of quantifiable results.

Chromogenic stains have been a mainstay as readouts for immunodiagnostics. 'Brown' staining—also known as diaminobenzidine (DAB) chromogenic staining or immunoperoxidase staining—was adopted very early in anatomic pathology and rapidly became the standard for immunohistochemistry in diagnostic pathology. Although immunofluorescent methods were developed before chromogen-based visualization, the DAB-based method was clearly better for pathologists because it was compatible with hematoxylin counterstains that allowed visualization of the context of a protein's expression pattern. The diagnostic features could easily be discerned and the expression pattern was valuable as much for its context or location as for its intensity. In fact, during the first 20 years of brown stain use (and for the vast majority of other antibody stains), the critical variable was presence or absence rather than intensity.

Gradually, however, the paradigm has expanded. The question is no longer binary (expressed or not); assays now query the amount of expression as the function of immunohistochemistry has evolved from diagnosis to prediction of response to therapy. And as the role of immunohistochemistry has evolved, its limitations have become more apparent^{1,2}.

Key problems

As molecularly targeted therapies have been developed, investigators have turned to immu-



Is 'brown' (DAB, or immunoperoxidase) staining on its way out? Source: James King-Holmes/ Science Photo Library

nohistochemistry as a predictive tool to assess the likelihood of response. The classic and one of the earliest molecularly targeted therapies is tamoxifen, a hormone analog that by blocking estrogen receptors inhibits tumorigenesis and is used in the treatment of breast cancer. As early as the 1970s, there was interest in predicting responses of patients to tamoxifen by measuring their estrogen receptor expression (for a review, see McGuire³). Initially, estrogen receptor expression was measured by a ligand binding biochemical assay⁴. The latter method showed a direct relationship between concentration of receptor in tumor biopsies and response to endocrine therapies^{5,6}. As tissue samples got smaller, however, it became impractical to grind up the tumors and carry out the biochemical assays. Soon immuno-

histochemistry became the standard and was even thought to be a better assay⁷.

Now immunohistochemistry is routine, but the linear relationship between estrogen receptor amount and predictive value has been lost. Recent studies describe a bimodal distribution of estrogen receptor in breast cancer patients⁸ that is almost certainly an artifact of the mechanism of measurement, because such bimodality is not seen in quantitative biochemical assays. Furthermore, the linear relationship between receptor amount and response has been lost. This observation has recently led to publications predicting that new assays that provide a more quantitative approach will improve the targeting of endocrine therapies (for example, see ref. 9).

Perhaps the highest profile brown stain test on the market is Dako's (Glostrup, Denmark) HercepTest for measuring HER2/neu in breast cancer biopsies as a means of predicting response to Genentech's (S. San Francisco, CA, USA) Herceptin (trastuzumab), a monoclonal antibody used to treat individuals with those breast cancers that overexpress the HER2/neu protein. Although this is a relatively simple assay where the need is only to tell very high HER2 expression from no HER2 expression, the inaccuracy of the assessment of the brown stain has led to the introduction of quantitative assays using fluorescence *in situ* hybridization (FISH) to detect gene amplification as an alternative to protein measurement¹⁰. This has resulted in the current recommendation that when immunohistochemistry is equivocal (2+ on a 0–3+ scale, where 0 is no staining, 1 is weak staining, 2 is moderate staining and 3 is strong staining), then FISH be done to determine therapy.

There are studies that claim immunohistochemistry as typically practiced is inferior to FISH¹¹ and there is evidence for differential practice in 'local' versus 'central' laboratories¹². Some investigators even recommend FISH

David L. Rimm is in the Department of Pathology, Yale University School of Medicine, 310 Cedar St., PO Box 208023, New Haven, Connecticut 06520-8023, USA. He is a founder, stockholder and consultant to HistoRx, the exclusive licensee of the Yale-owned AQUA patent.
e-mail: david.rimm@yale.edu

Box 1 Automated quantitative analysis (AQUA) to assay dynamic range

AQUA is an automated scoring system for assessing biomarker expression in tissue sections, using measurement of antibody-conjugated fluorophores within a specified subcellular compartment (typically including the nucleus, cytoplasm or plasma membrane), that relies on colocalization of the target and the compartment combined with a special image-flattening algorithm that results in accuracy comparable to that of an ELISA without the loss of spatial information¹⁸. My laboratory has studied the issue of dynamic range in measurement of protein expression *in situ* using a quantitative immunofluorescent method on a series of standards produced from cell lines both with and without amplification of the gene encoding HER-2 (ref. 16). We used ELISAs to determine the dynamic range for a standard curve using a cell line with genetically absent HER-2 (BAF3 cells, 0 pg per μg total protein) along with a series of cell lines, some of which have gene amplification, the extreme being BT474 with nearly 4,000 pg per microgram of total protein. This nearly 3.5-log dynamic range was not measurable with a single assay. We found that the standard antibody titration used in the clinical assay (1:8000 dilution) was sufficient to assess the amplified cell lines within the linear range of the assay, but all of the unamplified cell lines were below the threshold of detection (Fig. 1a). Those lines could be accurately quantitatively assessed, but to do so required a higher concentration of antibody (1:500 dilution). At the latter dilution, the concentrations in

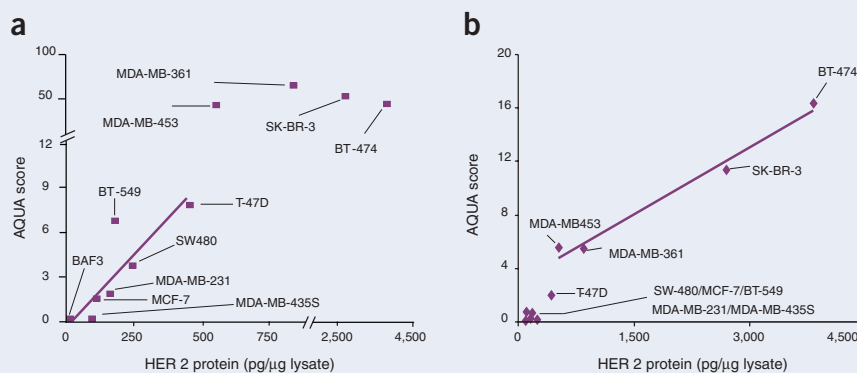


Figure 1 The relationship between AQUA score and absolute protein concentration as measured by ELISA assay as a function of antibody concentration. (a) When the antibody concentration is low (at a titer of 1:8,000), then protein concentrations in cell lines with amplification of the gene encoding HER2/neu fall in the linear range of the assay, whereas those in cell lines with normal amounts of HER2/neu are undetectable. (b) When the antibody concentration is high (at a titer of 1:500) then protein concentrations in unamplified lines fall in the linear range of the assay, whereas those in amplified lines have all saturated the assay and are indistinguishable. Line was fit by least squares method using only the cases within the range of the line. (Reprinted with permission from McCabe *et al.*, *J. Nat. Can. Inst.* **97**,1808–1815 (2005) Oxford University Press, Oxford, UK..

the unamplified lines could be read linearly with HER-2 protein concentration, but concentrations in the amplified lines could not be accurately measured because they saturated the assay (Fig. 1b).

This finding illustrates both the importance of standard curves (to be sure measurements are in the correct dynamic range) and the value of fluorescence (to extend that dynamic range). If it were important to detect low amounts of HER-2 to predict response, it would not have been achievable with the conventional assay. Furthermore, this subtle difference is not detectable by visual observation of the brown stain¹⁶.

as the preferred assay to predict response to Herceptin¹³. This is particularly ironic because Herceptin itself is an antibody, and its interaction with the target, HER2/neu, is ultimately required for response, regardless of the degree of amplification seen at the DNA level.

Dynamic range issues

Brown stains, as performed today, suffer from several deficiencies that may prevent their participation in the next-generation immunohistochemical assay. Perhaps the most significant issue is a biophysical property of the DAB substrate. The optimal visual brown stain (with the assay adjusted to generate sufficient DAB deposition to 'look good') has an absorbance of around 1–2 units. This means that up to 99% of the light is being blocked by the substrate and the eye or the automated reader must use only 1% of the total signal for analysis. This leads to problems in maximizing the dynamic range and in multiplexing. The dynamic range of any given detection device is a direct function of the signal-to-noise ratio;

the higher the signal and the lower the noise, the broader the dynamic range.

Although new charge-coupled device-camera detection instruments are very good at assessing very low amounts of light, the goal of increasing the signal-to-noise ratio is inherently contravened when using a method of analysis that decreases the total signal. Similarly, when multiplexing to identify more than one substrate, the total signal is progressively decreased as the number of light-absorbing substrates is increased. Thus, although multiple colored substrates are available and multiple species antibody probes can be used simultaneously¹⁴ by either spatial or spectral differentiation (referred to as spectral unmixing¹⁵), the core problem of decrease in signal proportional to the increase in substrates creates a barrier to maximizing signal-to-noise ratios.

Notably, many other biological assay systems have progressively moved from absorptive to emissive systems. For example, western blots are largely now done using luminescence,

and quantitative PCR and nanoparticle bead assays are done using fluorescence. This move has occurred as well for *in vivo* systems, where green fluorescent protein and similar fluorophores have replaced LacZ gene reporter systems, which use chromogenic substrates such as Xgal.

The issue of dynamic range is also complicated by the need for enzymatic amplification for signal detection. Protein concentrations *in vivo* generally span at least two logs (factors of ten), and in cases of gene amplification (as with the gene encoding HER2/neu in certain breast tumors), the protein expression ranges from a few hundred molecules per cell to more than 1,000,000 molecules (four logs). Thus, to accurately traverse this range, an assay with a broad dynamic range is required. Conventional DAB-based brown stains have a dynamic range of one to two logs at best. Claims of higher ranges are often not reproducible at the lower end of the scale. This limitation is probably owing to the combination of the range of the chromogen and the requirement for optimizing the

linear range of the enzymatic system used to deposit the chromogen. Although the inherently broader range (2–3 logs) of fluorescent probes is good enough for most proteins, it is insufficient for HER2/neu¹⁶. In this case, more than one antibody concentration is required to span the entire dynamic range (see **Box 1** and **Fig. 1**)¹⁷.

Ignoring issues of dynamic range and standardization can be disturbing. One investigator might use a high antibody concentration and find that low expression is associated with worse outcome, whereas another investigator might use very little antibody and find that high expression is associated with worse outcome. These opposite survival curves are possible because the limited dynamic range might allow each investigator to see only half the information. For example, the highly expressing cancers may not be resolved from the majority of moderately expressing cancers when using a high antibody concentration, owing to saturation of the assay. Using these types of observation, an investigator might resolve only the low expressors and define their worse outcome with respect to the larger remaining population. The reverse could be the case for an investigator who uses a very low concentration of antibody. A U-shaped relationship between receptor level and outcome exists in these situations—that is, small subclasses with either high or low expression are associated with poor outcome, compared with larger classes with moderate expression (**Fig. 2**).

Multiplexing assays and the future

The concept of multiplexing is becoming increasingly important and is likely to be critical for pathologists to retain their central role in tumor classification. The use of DNA-based arrays has shown the value of looking at mRNA expression levels from multiple genes, where best classification schemes for breast cancer and other diseases have used from 70 to 200 different markers¹⁸. Even minimizing the number of markers for optimal nucleic acid based analysis has yielded a cocktail of 16 genes (and 5 controls) in one marketed assay¹⁹, Genomic Health's (Redwood City, CA, USA) Oncotype, for predicting recurrence risk in breast cancer.

Proteins have the inherent advantage that they are the mechanism of function. Theoretically, therefore, fewer proteins than nucleic acid probes should be able to provide classificatory information for diagnostics. RNA concentrations are not tightly correlated with protein concentrations because of differential stability and to regulation of expression at the translational level²⁰, among other mech-

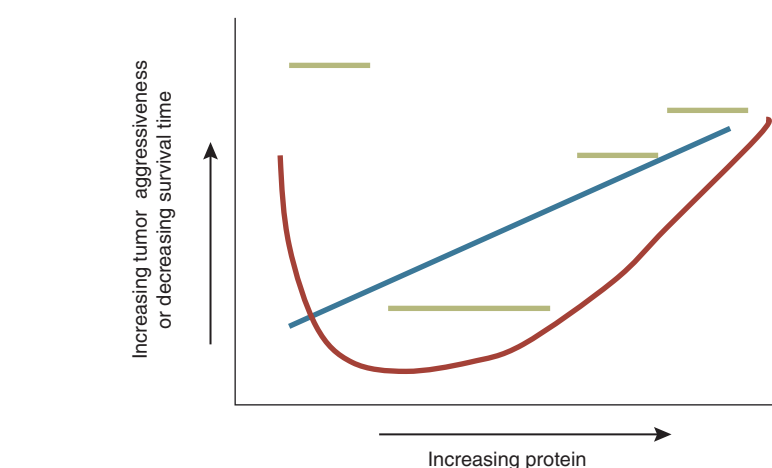


Figure 2 A theoretical series of curves showing relationships between expression and disease outcome. The straight line (blue) is an example of a linear relationship, in which increased expression of a given protein is associated with better outcome. The estrogen receptor marker in breast cancer is an example of this type of relationship. The U-shaped curve (red) is the relationship observed for markers, such as HER-2 and P53, in which both high and low expression are associated with worse outcomes compared with moderate expression. The green lines indicate biological classes, all with the same outcome (hence they are flat with respect to survival), but with a small, finite length with respect to protein concentration, suggesting a range of protein associated with that tumor class.

anisms. There are now a few studies that have appeared in abstract form and in the literature for breast cancer and melanoma^{21,22} suggesting that a smaller number of protein markers will be required for classification. However, this work is largely multiplexing by comparing serial sections. True multiplexing would allow 5–7 markers to be assessed on a single tissue sample. At present, two colors appear to be the multiplex limit of chromogenic methods; in comparison, fluorescence can easily cope with five colors and may soon be able to achieve seven or more. Thus, as pathologists move immunohistochemistry into the next generation, fluorescence appears to be a better mechanism to capture the advantages of multiplexing.

Although it is impossible to predict the future, the expanding landscape of molecularly targeted therapies and companion diagnostics will require quantitatively measured protein expression, without the destruction of spatial information inherent in ELISA-type assays or the subjectivity and lack of standardization of the conventional immunohistochemical assay. It seems likely that the next generation of protein-based predictive assays will have to be rigorously quantitative, internally and externally standardized, and objectively reproducible.

ACKNOWLEDGMENTS

Thanks to M. Dolled-Filhart, C. Kleer and C. Allred for critical review of this manuscript. Work in the Rimm laboratory is supported by grants R01 CA 114277, R33 CA 106709, R33 CA 110511 and an

Avon Patients for Progress Grant from the U.S. National Institutes of Health and the Breast Cancer Alliance of Greenwich, Connecticut, USA. Dr. Rimm is a founder, stockholder and consultant to HistoRx, the exclusive licensee of the Yale owned AQUA patent.

- McCormick, D., Yu, C., Hobbs, C. & Hall, P.A. *Histopathology* **22**, 543–547 (1993).
- Hall, P.A. & Lane, D.P. *J. Pathol.* **172**, 1–4 (1994).
- McGuire, W.L. *Annu. Rev. Med.* **26**, 353–363 (1975).
- Garola, R.E. & McGuire, W.L. *Cancer Res.* **38**, 2216–2220 (1978).
- Pertschuk, L.P. *et al. Cancer* **46**, 2896–2901 (1980).
- Bezвода, W.R., Esser, J.D., Dansey, R., Kessel, I. & Lange, M. *Cancer* **68**, 867–872 (1991).
- Harvey, J.M., Clark, G.M., Osborne, C.K. & Allred, D.C. *J. Clin. Oncol.* **17**, 1474–1481 (1999).
- Collins, L.C., Botero, M.L. & Schnitt, S. *J. Am. J. Clin. Pathol.* **123**, 16–20 (2005).
- Schnitt, S.J. *J. Clin. Oncol.* **24**, 1797–1799 (2006).
- Pauletti, G., Godolphin, W., Press, M.F. & Slamon, D.J. *Oncogene* **13**, 63–72 (1996).
- Press, M.F. *et al. J. Clin. Oncol.* **20**, 3095–3105 (2002).
- Press, M.F. *et al. Clin. Cancer Res.* **11**, 6598–6607 (2005).
- Mass, R.D. *et al. Clin. Breast Cancer* **6**, 240–246 (2005).
- Hasui, K. *et al. J. Histochem. Cytochem.* **51**, 1169–1176 (2003).
- Levenson, R.M. *Cytometry* (in the press).
- McCabe, A., Dolled-Filhart, M., Camp, R.L. & Rimm, D.L. *J. Natl. Cancer Inst.* **97**, 1808–1815 (2005).
- Camp, R.L., Chung, G.G. & Rimm, D.L. *Nat. Med.* **8**, 1323–1327 (2002).
- van de Vijver, M.J. *et al. N. Engl. J. Med.* **347**, 1999–2009 (2002).
- Paik, S. *et al. N. Engl. J. Med.* **351**, 2817–2826 (2004).
- Yoon, S.O., Shin, S. & Lipscomb, E.A. *Cancer Res.* **66**, 2732–2739 (2006).
- Makretsov, N.A. *et al. Clin. Cancer Res.* **10**, 6143–6151 (2004).
- Alonso, S.R. *et al. Am. J. Pathol.* **164**, 193–203 (2004).