Advances in analytic technologies have often driven progress in medicine. This is particularly true for diagnostics in which new imaging methods and novel molecular analytic tools such as DNA sequencing and PCR have all found their diagnostic niche and are now routinely used in the clinic. High-throughput microarray-based analytic methods were first described over a decade ago (1). It enabled researchers to simultaneously and semiquantitatively measure the expression of thousands of mRNA species from a biological specimen in a single experiment. Molecular biologists and clinical investigators quickly realized the potential of this technology (2). It was expected that new insights into cell biology would be gained through unbiased analysis of the transcriptome, and that combinations of genes would provide more accurate prognostic and predictive tests than any single gene alone. It may be timely to reflect on what has materialized from these initial hopes, keeping in mind that it often takes several decades for a new technology to mature and show its full effect on science or show its value in clinical medicine.

Are DNA Microarrays Suitable for Diagnostic Assay Development?

New techniques inevitably have their enthusiastic early proponents with optimistic, but mostly unproven, claims, and early skeptics who point out the numerous pitfalls that lay ahead before true value could be shown. This healthy debate motivates research whose results eventually settle the claims. Early critics of microarray results pointed out that DNA microarray–based measurements might not be reliable because different array platforms frequently provide discordant results for the same RNA (3, 4). Furthermore, when a platform contains several different probe sets that target the same gene at different sequence locations, each probe could yield different expression results. Also, different research groups who studied the same disease invariably reported different sets of genes that were predictive of the same clinical outcome. Not surprisingly, several investigators concluded that microarrays, at best, represent a discovery tool and that all microarray results must be validated with RT-PCR or other methods and that future diagnostic tests would be based on these more established techniques. Many of these seemingly controversial results reflect the unusual nature of the data that is generated by high-throughput analytic methods and do not necessarily indicate an unreliable or inferior technology.

The U.S. Food and Drug Administration launched the Microarray Quality Control Project in collaboration with 51 academic and industry collaborators to systematically examine the technical reproducibility of microarray measurements within and between laboratories, as well as to compare results across different microarray platforms. Four different RNA samples were profiled in five replicates on seven different array platforms, three different laboratories represented each platform. The same RNA samples were also profiled by three different RT-PCR methods. The most important finding of this collaborative effort was that the microarray measurements were highly reproducible within and across the platforms (5). The median coefficient of variation for within-laboratory replicates ranged from 5% to 15% for the various platforms, whereas it was 10% to 20% for between-laboratory replicates (Fig. 1). Most importantly, the concordance between the microarray-based mRNA measurements and RT-PCR results were also quite high, the correlation coefficients ranged from 0.79 to 0.92 for several hundred genes that were examined with both methods. Much of the between-platform variations in gene expression measurements were due to sequence variations in the probe sets that target the same gene at different locations. Important factors that contribute to the different signal intensities generated by distinct probes for the same gene include differing GC-nucleotide content, sequence length, intraplatform cross-match opportunities, and the location of the probe sequence in relation to the 3′-end of the target gene (6). Using a simple analogy, different probes could be considered similar to different antibodies that target distinct epitopes of the same protein (7). The concordance between signal intensities for probes that target the same gene on different platforms is directly related to sequence homology (8, 9). Probes with complete sequence matches yield concordant results across platforms (Fig. 2; ref. 10).

Microarray-based clinical prediction results are usually based on the combined expression values of many different probe sets, and small measurement errors in each could alter the final prediction score substantially. Few studies have systematically examined the within- and cross-platform reproducibility of multi-gene prediction scores. Two reports have indicated the high reproducibility of pharmacogenomic prediction results in replicate experiments from the same RNA when the same experimental procedures were used (11, 12). Greater caution must be exercised when trying to reproduce prediction results across different platforms and different data sets. There are multiple sources of variation that could
compromise reproducibility, including differences in probe sets (see discussion above), data normalization methods, tissue sampling, and differences among the study populations. Essentially all published results that attempted cross-platform testing of gene signatures reported diminished, but not completely lost, classification accuracies on data generated by platforms other than the original platform (13–15). This is not unique to microarray results, older analytic methods including immunohistochemistry, PCR, and other techniques are all subject to the same increased variability when the experimental details are changed (e.g., different antibodies or primers are used) and different study populations are compared.

There is one other feature of microarray results that has lead to substantial skepticism about the reliability of this methodology. Different groups who studied the same clinical problem invariably identified different sets of genes as predictors. This is not surprising, in fact, it is expected. In many fields of science, it is well understood that several different but equally good solutions to the same complex problem can exist. High-throughput analytic methods provide measurements of thousands of variables that are not independent of each other. Gene expression values are highly correlated, and therefore, if the expression of a particular gene is associated with clinical outcome, all other genes whose expression is closely correlated with that index gene will also correlate with the particular clinical outcome. The strength of correlation between the genes and clinical outcome varies from training set to training set, and therefore, the rank order of these informative genes is...
unstable. The important observation, however, is that all of these coexpressed genes carry information about the outcome of interest. A corollary of this is that many different statistically equally good predictors can be discovered from the same or similar data sets (12, 16). An important recent publication clearly illustrates this, investigators applied five distinct multi-gene prognostic predictors to a single breast cancer gene expression data set that represented independent validation for four of the five predictors. These prognostic gene sets had very few genes in common, yet four of the five have shown similar prognostic values (17).

Do Microarray-Based Prognostic and Response Prediction Signatures Provide Clinical Value?

MEDLINE literature search using combinations of “cancer” and “microarray” shows that in 1996, there were 161 publications on this subject, and by 2005, this number had increased to 1,635. It is not possible to review all clinically relevant microarray results; therefore, we will illustrate the progress of microarrays from the laboratory to the clinic by focusing on breast cancer research. The two perpetual diagnostic challenges in breast cancer are (a) to predict the prognosis of an individual with newly diagnosed breast cancer (i.e., estimate the likelihood of cure with surgery alone) and (b) to predict what treatment may work if additional systemic chemotherapeutic therapy is needed. Prognostic indices such as AdjuvantOnline1 that integrate clinical and pathologic variables into a single risk prediction score represent the current standard for prognostic prediction. Clinical variable–based prognostic models are useful but imperfect. Can gene expression profiling–based tests improve on prognostic prediction? At least two different multi-gene prognostic signatures for breast cancer have been developed that were also evaluated on independent cases. The first genomic prognostic index included 70 genes and was developed from 98 patients with lymph node–negative breast cancer (Mammaprint, Agendia, Inc., Amsterdam, the Netherlands; ref. 18). The prognostic value of this 70-gene signature was evaluated on a partially independent set of 295 patients and showed that those with a good prognosis signature had 95% distant metastasis–free survival at 5 years (85% at 10 years) compared with 60% in the poor prognosis signature group (19). A second independent validation (n = 307) of the same gene set also showed that patients with a good prognosis signature had 89% overall survival at 10 years compared with 69% in the poor prognosis group (20). Importantly, the performance of this gene signature was compared with predictions from AdjuvantOnline, and in discordant cases, the gene signature provided more accurate prognostic information than the clinicopathologic prediction model (Fig. 3). Other investigators identified genes that were associated with relapse separately for the estrogen receptor–negative and the estrogen receptor–positive breast cancers. The markers selected from each group were combined to form a single 76-gene prognostic signature (VDX2 gene chips, Veridex LLC, Warren, NJ; ref. 21). This genomic prognostic predictor also did well when independently tested on 180 lymph node–negative cases. The 5- and 10-year distant metastasis–free survival rates were 96% and 94% for the good prognosis group and 74% and 65% for the poor prognosis groups, respectively (22).

Are these results good enough for clinical use? Both of these microarray-based assays provide binary prognostic prediction (good versus bad prognosis) with moderately high accuracy. What constitutes low enough risk to forgo systemic chemotherapy, however, is influenced not only by the absolute risk of relapse but also by the risk of adverse events from therapy as well as by personal preferences (23). Some patients are willing to accept adjuvant chemotherapy (i.e., chemotherapy given after surgery to further improve the chance of long-term survival) for small gains in survival. Molecular prognostic markers may provide little clinical value for these individuals because no test is accurate enough to completely rule out the risk of relapse or benefit from adjuvant therapy. Many other patients are more reluctant, however, to accept the toxicities, inconvenience, and costs of chemotherapy for a small and uncertain benefit. For these individuals, a more precise prediction of risk of recurrence and sensitivity to adjuvant therapy with genomic tests can result in more informed decision making.

The clinical importance of predicting who will or will not benefit from a particular therapy is intuitively obvious. Those who are predicted to respond could be treated whereas others could be spared from the unnecessary toxicity. Similar to prognostic prediction, it is possible to combine clinical and pathologic variables (e.g., nuclear grade, tumor size, and estrogen receptor status) into a multivariable model to predict the probability of response to preoperative chemotherapy for patients with stage I to III breast cancer2 (24). However, this clinical variable–based prediction model lacks regimen specificity and cannot be used to select one treatment over another. Several small studies provided “proof-of-principle” that the gene expression profile of cancers which are highly sensitive to chemotherapy are different from tumors that are resistant to treatment (for review, see ref. 25). The largest study thus far included 133 patients with stage I to III breast cancer who all received weekly paclitaxel and 5-fluorouracil, doxorubicin, and cyclophosphamide (T/FAC) preoperative chemotherapy. The first 82 cases were used to develop a multi-gene predictor of complete response to treatment and the remaining 52 cases were used to test the accuracy of the predictor (12). The 30-gene predictor correctly identified all but one of the patients (n = 12/13) who achieved pathologic complete response, and all but one of those who had residual cancer (n = 27/28) in the validation set. The pharmacogenomic test showed significantly higher sensitivity (92% versus 61%) than the clinical variable–based response predictor, the overall accuracy was 76%, the positive predictive value was 52%, and the negative predictive value was 96%. To what extent this genomic predictor of drug sensitivity is specific to the T/FAC treatment regimen, rather than being a generic marker of chemotherapy sensitivity, however, is yet to be determined.

An interesting and promising alternative strategy to predictive marker discovery is to identify predictors in preclinical models and test these markers in the clinic. This approach was explored

1www.adjuvantonline.com.

by investigators at Duke University who compared the gene expression profiles of panels of sensitive and resistant cell lines for six different chemotherapy drugs. Drug-specific multi-gene response predictors were constructed for each drug and these were combined into multi-drug predictors. When this in vitro–defined T/FAC-predictive signature was tested on the human breast cancer data (see above), it showed 82% overall accuracy, 61% positive predictive value, and 94% negative predictive value (26).

There is considerable uncertainty regarding what level of predictive accuracy is clinically useful. In fact, different levels of predictive accuracy may be required for different clinical situations. For instance, the clinical usefulness of a chemotherapy response prediction test that has 60% positive predictive value (i.e., 60% chance of response if the test is positive) and 80% negative predictive value (20% chance of response if the test is negative) will depend not only on these test characteristics but also on the availability and efficacy of alternative treatment options, as well as the frequency and severity of adverse effects, and the risks of exposure to ineffective therapy (i.e., rapid disease progression with life-threatening complications). A test with the above performance characteristics may be of limited value in the palliative setting, wherein alternative treatment options are limited and generally ineffective. Patients and physicians may want to try a drug even if the expected response rate is only 10% to 15%, particularly if side effects are uncommon or tolerable. On the other hand, in the setting of potentially curative therapy, wherein multiple treatment options are available, a test with the same performance characteristics may be helpful to select the best regimen from several treatment options.

Conclusions

Much has been learned about the performance of DNA microarrays in the past 13 years. It is increasingly recognized that gene expression profiling experiments when done properly yield reliable results. Investigators are also increasingly aware of some unusual but natural features of high-throughput data, such as the innate instability of P values–based rank orders (which yield different gene sets every time new cases are added to the data set). Because of multiple large networks of coexpressed genes, several different gene sets can be discovered from the same data that each perform equally well in independent validations. The costs of these experiments have also decreased. The U.S. Food and Drug Administration has approved at least one microarray hardware system for clinical use together with a CYP450 genotyping chip. Several studies have established the validity of multivariable gene expression–based clinical outcome predictors. At least in some instances, the gene signature–based predictors were more accurate than clinical variable–based models (12, 20, 27). The first multi-gene prognostic test for estrogen receptor–positive, lymph node–negative breast cancer is now used clinically. This assay, Oncotype Dx (Genomic Health, Redwood City, CA) calculates a recurrence score based on the expression of 21 genes and can stratify patients who received 5 years of endocrine therapy into various risk categories for relapse (27). This assay can help some women make a more informed decision about chemotherapy which may be recommended in addition to endocrine treatment. The test uses multiplex RT-PCR to quantify gene expression. This investigator, however, does not see any fundamental flaw in DNA microarray technology that would preclude it from becoming a similar diagnostic test on its own. Indeed, microarray-based measurements of the same 21 genes can also stratify patients by risk of recurrence (17).

It is important to keep in mind that the many genomic tests that are currently under development are scattered across a spectrum of clinical development stages (Fig. 4). Many more gene signatures were proposed than validated. What is the minimum standard that a novel diagnostic test needs to meet

Fig. 3. Kaplan-Meier plots of time to distant metastasis in the absence of any systemic adjuvant therapy by the 70-gene prognostic signature and by the AdjuvantOnline risk categories. Error bars indicate 95% confidence intervals (adapted with permission from ref. 20).
before it could be considered for clinical use? Historically, the assay is expected to be technically robust and reproducible, and the performance characteristics of the test, including its predictive values (for a clinically relevant outcome), need to be defined with narrow confidence intervals from independent validation data. If these criteria are met, the test may be considered for clinical use (or at least for prospective clinical evaluation) because it predicts a relevant clinical outcome with a known degree of certainty. When considering this question, it is important to understand that even when a test is not indicated for every patient, it could provide clinical value for some. For example, magnetic resonance imaging is not done on all patients with newly diagnosed breast cancer or with a suspicious lump in the breast; however, it is a useful test for a subset of women with these conditions. Many of the emerging gene expression–based prognostic and predictive tests may fall into a similar category. Sometimes treatment decisions can be made relatively easily based on the available clinical information and the current moderately accurate molecular predictors may add little further value. For other patients who are undecided, however, molecular prognostic/predictive assays could result in a more informed decision making. Ultimately, one would also like to see proof that such “more informed” medical decision leads to improved patient outcome (e.g., increased survival and better quality of life). At least two large randomized studies, one in Europe (MINDACT trial) and one in the U.S. (PACT trial) will examine exactly this last but critical question for the 70-gene prognostic signature and for the Oncotype Dx recurrence score, respectively. Final results from these studies will not be available for many years, however, the development and systematic validation of these two gene expression profiling–based tests mark a clear and exemplary translational research path for many of the aspiring novel diagnostic tests.

References


Chips to Bedside: Incorporation of Microarray Data into Clinical Practice

Lajos Pusztai