ELSEVIER

Seminars in
HEMATOLOGY

Interpretation of Genomic Data: Questions and Answers

Richard Simon

Using a question and answer format we describe important aspects of using genomic technologies in cancer research. The main challenges are not managing the mass of data, but rather the design, analysis, and accurate reporting of studies that result in increased biological knowledge and medical utility. Many analysis issues address the use of expression microarrays but are also applicable to other whole genome assays. Microarray-based clinical investigations have generated both unrealistic hype and excessive skepticism. Genomic technologies are tremendously powerful and will play instrumental roles in elucidating the mechanisms of oncogenesis and in bringing on an era of predictive medicine in which treatments are tailored to individual tumors. Achieving these goals involves challenges in rethinking many paradigms for the conduct of basic and clinical cancer research and for the organization of interdisciplinary collaboration.

Semin Hematol 45:196-204 © 2008 Elsevier Inc. All rights reserved.

This chapter will address some key issues on the use of genomic technology in biomedicine. The focus will be on cancer therapeutics, although many issues have broader relevance. Study design for both developmental and validation studies will be addressed, as well as topics in the analysis of genomic data: matching analysis strategy to study objective, limitations of traditional statistical tools for whole genome assays, and recommended analysis methods. A question and answer format is used with division into general introductory topics, queries about biologically focused "gene finding" studies, and questions about medically focused studies using genomics for predictive medicine.

Introductory Issues

What is the difference between genomic data and genetic data?

Genomic data provides information about the genome of a cell or group of cells. This includes both the genetic polymorphisms that are transmitted from parent to offspring as well as information about the somatic alterations resulting from mutational and epigenetic events.

Why are genomic data important?

Cancer is a disease caused by altered DNA. Some of these alterations may be inherited and some somatic. Genetic association studies attempt to identify the genetic polymorphisms that increase the risk of cancer. These contribute to understanding the molecular basis of the disease and permit identification of individuals for whom intensive surveillance or chemoprevention strategies may be appropriate. The genomics of tumors are studied in order to understand the molecular basis of the disease, to identify new therapeutic targets, and to develop means of selecting the right treatment for the right patient.

Is "the right treatment for the right patient" hype or substance?

Both. The phrase originated outside of oncology, where it was interpreted to mean personalizing therapy based on the genetic makeup of the patient. In oncology, personalization of therapy has mostly been based on the genomics of the tumor, not the genetics of the patient. The tumors originating in a given anatomical site are generally heterogeneous among patients; tumor genomics provides relevant information about that heterogeneity. In some areas of oncology targeted medicine is already a reality. For example, in breast cancer, treatment is often selected based on estrogen receptor status and *HER2* gene amplification.[1,2] Using genomics effectively for treatment selection depends critically on the predictive accuracy of the genomic test and the medical context. To withhold a potentially curative treatment from a patient

Biometric Research Branch, National Cancer Institute, Bethesda, MD.
Address correspondence to Richard Simon, DSc, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda MD 20892-7434. E-mail: rsimon@nih.gov

0037-1963/08/$-see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1053/j.seminhematol.2008.04.008

based on a test with less than perfect negative predictive value would be a serious mistake. A genomic test is only warranted if its predictive accuracy adds substantially to that of existing practice guidelines.[3] Extensive clinical studies are needed to demonstrate that a genomic test is ready and appropriate for clinical use.[4]

What kinds of genomic data are available?

Starting around 1996, DNA expression microarrays began to provide estimates of the abundance of mRNA transcripts genome wide. Today arrays are available to measure transcript abundance information for each exon of each gene in the genome. Within the past several years comparative genomic hybridization arrays and single-nucleotide polymorphism (SNP) arrays have been used to identify copy number variations and loss of heterozygosity on a genome-wide basis. Genome-wide genotyping is in use for identifying SNPs and in the next few years it will be economically feasible to completely resequence the genomes in individual tumors.

Is the challenge how to manage all of this data?

Managing the data is not the main challenge. The amount of data is well within the capability of modern information technology. For example, the BRB-ArrayTools software package developed by the author (available at http://linus.nci.nih.gov) can easily handle 1,000 expression profiles of 50,000 transcripts to develop predictive classifiers, fully cross-validated, on a personal computer within minutes.[5] The much greater challenge is the proper design, analysis, interpretation, and reporting of studies to utilize the technology in a way that provides meaningful biological information and diagnostic tests that have real medical utility.[6] A recent review by Dupuy and Simon indicated that half of published papers relating expression profiling to cancer outcome contained at least one error sufficiently serious as to raise questions about the conclusions of the study.[7] Because of the number of variables measured with genome-wide assays, there is great opportunity for discovery but also great risk of reaching misleading conclusions. The statistical analysis of such data is very challenging and it is critical that authors make their data, both the genomic data and the clinical data, publicly available for others to independently verify their claims and to utilize their data in meta-analyses. The restrictions on data sharing that have been practiced for clinical trials data are not desirable for whole-genome assay studies. Some journals require sharing, but it should be an absolute requirement for all cancer periodicals.

Isn't cluster analysis the way to analyze gene expression profiles?

The recent report by Dupuy and Simon[7] identified inappropriate use of cluster analysis as one of the most common flaws in published studies relating microarray gene expression to cancer outcome. The overuse of cluster analysis is indicative of a more fundamental problem that limits the effective use of genomic technology, the lack of adequate interdisciplinary collaboration. Analysis of genome-wide data is complex, and few biologists or clinical investigators have the appropriate training for this task. Many of the design and analysis problems presented by genomic data are also new for statisticians, and application of standard statistical approaches to high-dimensional genomic data often gives unsatisfactory results. Statisticians who invested substantial time learning about medicine made crucial contributions to cancer clinical trials. Making such contributions to biology and genomic medicine will require the same type of effort. Unfortunately, the organizational structures of many of our institutions are not well suited to effective interdisciplinary collaboration. Organizations sometimes overemphasize software engineering and database building and underemphasize high-level statistical genomics collaboration. Many cancer research groups have not made the resource commitments necessary to attract the best qualified individuals and to foster effective multidisciplinary collaboration.

Can biologists and clinical investigators analyze genome-wide data?

Multidisciplinary collaboration is most effective when there is substantial overlap of knowledge. One of the challenges in biomedicine today is training and retraining scientists in the effective use of whole-genome data. The challenge is not really in doing the assays, because assays quickly become commodities that can be ordered. Issues of how to design studies and analyze data involving genome-wide technology are important for biologists and clinical investigators, not just statisticians and computational scientists. One of the main objectives of BRB-ArrayTools[5] is to provide biomedical scientists a software tool for such training. It is also important that clinical scientists learn enough to be appropriately critical readers of the published literature; there are serious problems in some papers published in even the most prominent journals.[7] Many young biologists and clinical investigators are eager to develop their expertise in this area; the goal is important, but achieving it requires an investment of time.

What are the appropriate analysis methods?

The right methods and the right specimens depend on the objective of the study. Microarray expression profiling has allowed entirely new kinds of biological investigations. Traditionally, in studying biological mechanisms the focus was on a small number of proteins, development of assays to measure them, and then design of an experiment to test a hypothesis about how the concentrations of the proteins would vary under the experimental conditions. Today, one can measure the abundance of all transcripts in a single assay. Consequently, less focused kinds of experimentation are possible. Although microarray based studies do not require gene- or protein-specific hypotheses, a clear objective is still important in order to design an interpretable experiment with appropriate samples and an appropriate analysis. Many uses of microarrays can be categorized as (1) class discovery, (2) gene finding or class comparison, and (3) prediction.

What is class discovery?

Finding genes that are co-regulated or are in the same pathway can sometimes be accomplished by sorting genes into groups with similar expression profiles across a set of conditions. Many "cluster analysis" algorithms have been developed to accomplish this sorting. Cluster analysis algorithms are sometimes used to sort samples into groups based on similarity of their expression profiles over the set of genes. Clustering samples generally does not use any phenotype information about the samples. However, cluster analysis methods always result in clusters, and there is generally no appropriate way of "validating" a cluster analysis except by seeing whether the resulting clusters differ with regard to a known phenotype.

In seeking gene expression–based groupings of samples that correlate with a phenotype, it is generally much better to use "supervised" prediction methods, so-called because they use the phenotype class information explicitly. Often there may only be a small number of genes whose expression is correlated with the phenotype, and unsupervised cluster analysis will not group the samples in ways that correlate with the phenotype. A serious common mistake is to cluster the samples with regard to the genes found to be correlated with the phenotype. Showing that the samples can be thereby clustered into groups that differ with regard to the phenotype is erroneously used as evidence of the relevance of the selected genes. This practice violates the principle of separating the data used for developing a classifier from the data used for testing it. Since the same data are used for identifying the genes and for clustering the samples with regard to the selected set of genes, the process is invalid. As pointed out by Dupuy and Simon, this is one of the most frequent serious errors in studies relating gene expression to cancer outcome.[7]

Gene Finding

What methods are appropriate for gene finding problems?

Gene finding includes studies of mechanisms, like what genes are induced during wound-healing, or what genes are differentially expressed in normal mouse breast epithelium compared to a breast tumor in a genetically engineered mouse. Gene finding is sometimes called class comparison. For comparing gene expression between two classes of samples, familiar statistical measures such as significance tests of difference in mean expression between the classes can be employed. It is important, however, to take into account that differential expression is being compared for tens of thousands of genes. Hence the usual threshold of .05 for statistical significance is not appropriate. Using the .05 threshold there will be 500 false positive genes declared differentially expressed per 10,000 genes tested. This average false positive rate is independent of the correlation of expression among the genes. A threshold of statistical significance of .001 instead of .05 results in only 10 false positives per 10,000 genes tested on average.

For gene finding it has become standard to control the "false discovery rate." If *n* genes are reported in a publication to be differentially expressed between the classes and if *m* are false positives, then *m/n* is the false discovery rate. The simplest way to control the false discovery rate is by using the method of Benjamini and Hochberg.[8] Suppose that a publication reports *n* genes as differentially expressed and all have a *P* value less than p^*. Then an approximation to the false discovery rate is Np^*/n where *N* denotes the number of genes tested for differential expression. This is based on estimating the number of false positives as p^* times the number of genes tested, *N*. This calculation is generally somewhat conservative since some of the *N* genes are actually differentially expressed and other approximations are also used.[9,10] Other methods for finding genes that are differentially expressed, such as SAM[11] and the multivariate permutation test,[12] control the false discovery rate in a more sophisticated manner that takes into account the correlation among genes. The multivariate permutation test of Korn et al, SAM, the Benjamini Hochberg method, as well as more complex analysis of variance methods are available in BRB-ArrayTools.[5]

Class comparison methods are not limited to finding genes that are differentially expressed between two classes. There may be more than two classes or one may be interested in genes whose expression is correlated with a quantitative variable or a censored variable such as survival time. In time-course experiments, the interest may be in genes the expression of which changes with time after an experimental intervention.[13] Also of interest may be genes whose expression varies with time differently for two classes of samples. All can be viewed as gene-finding problems. Although the statistical measures of correlation of gene expression with phenotype depend on the nature of the problem, the control of the number or proportion of false positives is always important. Failure to provide adequate control of false positives was one of the three most common serious problems in expression profiling studies reported by Dupuy and Simon.[7]

How many samples are needed for gene finding with expression data?

For comparing classes, representative samples are needed from each class. In general, the biological variability in expression among samples of the same class is much greater than the variability among technical replicates (among replicate arrays of the same mRNA sample). The statistical power of gene finding studies primarily depends on the number of biological replicates, and it is often appropriate to not perform any technical replicates. These issues, particularly for dual-label arrays, are described by Dobbin et al.[14,15] The number of cases needed in each class depends on the fold difference in mean expression to be detected and the degree of biological variation in expression within each class. Often the studies are sized to detect a twofold mean difference in expression. The intra-class variation differs among genes and is greater for human tissues than for cell lines. Dobbin et al[15] provide simple formulas based on controlling the false discovery rate by using a stringent type one error level for sample size planning and these methods are available in BRB-ArrayTools.[5] Shih et al[16] have also shown that pooling of samples is rarely desirable unless necessary to obtain enough RNA for the assay.

How are lists of differentially expressed genes related to pathways?

Traditionally, first a list of differentially expressed genes is generated, and then software tools and genomic websites are used to annotate the genes appearing on the list. However, this approach has some serious limitations. To limit the false discovery rate, genes are usually included in the list only if their *P* value for differential expression is statistically significant at a stringent threshold, but many genes are differentially expressed but not to the extent required for inclusion in the gene lists. A popular alternative uses pathway information directly in the evaluation of differential expression, not post hoc to annotate the gene lists. Gene set enhancement analysis[17] is one method of this type. It focuses attention on a specified set of genes and computes a summary statistic of the extent to which the set is enriched in genes that rank high with regard to overexpression in the first class compared to the second class. However, in computing that enrichment score a binary categorization of genes as differentially expressed or not differentially expressed is not enforced. The method then computes the significance of the degree of summary enrichment relative to the expectation if no genes were differentially expressed among classes. Gene sets that are significantly enriched relative to that null distribution are identified. Tian et al[18] pointed out that there are various null hypotheses that could be tested, and that measuring enrichment or differential expression relative to the global null hypothesis that no genes are differentially expressed may not be useful in cases where there are many differentially expressed genes. Numerous alternative methods have been reported.[19-21] BRB-ArrayTools contains several methods for this purpose to evaluate the relationship of differential gene expression among classes to a variety of gene sets, including gene ontology categories, Biocarta signaling pathways, Kegg metabolic pathways, Broad Institute signatures, transcription factor targets, microRNA targets, and genes whose protein products contain a PFAM protein domain.[22]

Prediction

How do prediction problems differ from gene finding problems?

Prediction problems arise in many medical applications, as for example to predict which tumors are likely to respond to a given drug. Prediction may be thought of as a two-class problem with one class consisting of samples from patients who have responded to treatment and the other class of samples from non-responders. Although one component of developing a predictive classifier is selection of the informative genes to include, predictive problems are actually quite different from class comparison problems. In class comparison problems it is important to control the false discovery rate. In prediction problems, however, the objective is accurate prediction for independent data, not limiting the false discovery rate to an arbitrarily specified value. Thus the appropriate criteria for gene selection in prediction problems are different from those for class comparison problems. For example, in prediction it is often much more serious to miss informative genes than to include some false discoveries.[23] Class comparison or gene finding problems often have as their object understanding biological mechanisms. In some cases, it is much easier to develop an accurate predictor than to understand the biological basis of why the predictor works. (Understanding biological mechanisms is quite difficult and many excellent biologists have spent a career trying to understand experimental systems that are much simpler than mammalian cells.)

What kinds of predictive classifiers are best?

A class predictor, or classifier based on gene expression data, is a function that predicts a class from an expression profile. Specification of a class predictor requires (1) detailed description of the genes whose expression levels are to be utilized in the prediction; (2) selection of the mathematical form used to combine the expression levels of the component genes; and (3) specification of the parameters such as weights placed on expression levels of individual genes and threshold values used in the prediction. A predictive classifier is more than a set of genes. The development of a predictor has some similarities to logistic regression analysis. Statistical regression models have in the past usually been built using data in which the number of cases (n) is large relative to the number of candidate variables (p). However, in the development of class predictors using gene expression data, the number of candidate predictors is generally orders of magnitude greater than the number of cases. This has two important implications. First, only simple class prediction functions should be considered. Second, the data used for evaluating the class predictor must be distinct from the data used for its development. It is almost always possible to develop a class predictor even on completely random data that will fit that same data almost perfectly but be completely useless for prediction with independent data!

The most commonly used approach to selecting genes to include in the predictive classifier is to use the genes that by themselves most correlate with the outcome or the phenotype class. Traditionally, procedures like stepwise regression methods are used to select variables that make independent contributions to prediction. In traditional regression modeling, there is careful consideration of whether variables should be transformed and whether interactions among the effects of combinations of variables should be included in the model. This type of modeling requires large sample sizes. An accepted rule for traditional regression modeling is to have at least 10 times the number of cases as variables. With whole-genome assays, we have tens of thousands of variables, and the expression of each gene represented on a microarray is a variable. Consequently, the 10 to 1 ratio would require hundreds of thousands of cases for analysis, clearly an impossibility. As a result, the kind of regression modeling that statisticians have employed for problems with many cases and few variables does not work well for genomic problems. Accurate prediction is possible in high dimensional ($p>>n$) problems, but different methods of predictive modeling must be utilized to avoid over-fitting the data.

Numerous algorithms have been used effectively with DNA microarray data for class prediction. Many of the widely used classifiers combine the expression levels of the genes selected as informative for discrimination using a weighted linear function

$$l(\underline{x}) = \sum_{i \in G} w_i x_i \qquad (1)$$

where x_i denotes the log-expression for the i'th gene, w_i is the weight given to that gene, and the summation is over the set G of genes selected for inclusion in the classifier. For a two-class problem, there is also a threshold value d; a sample with expression profile defined by a vector $\underline{x}$ of values is predicted to be in class 1 or class 2 depending on whether (1) is less than the threshold d or greater than d, respectively. Many of the widely used classifiers are of the form shown in equation (1); they differ with regard to how the weights are determined.

Dudoit et al,[24,25] in comparing many classification algorithms, found that the simplest methods, diagonal linear discriminant analysis and nearest neighbor classification, usually performed as well or better than did more complex methods. Nearest neighbor methods are not of the linear form shown in equation (1); they are based on computing similarity of a sample available for classification to samples in a training set. Often Euclidean distance is used as the similarity measure, but it is calculated with regard to the set of genes selected during training as being informative for distinguishing the classes. The PAM method of Tusher et al is a popular form of nearest neighbor classification.[11] Ben-Dor et al[26] also found that nearest neighbor classification generally performed as well or better than did more complicated approaches. Similar results were found by Wessels et al.[27]

There is a substantial literature on complex methods for selecting small subsets of genes that work well together to provide accurate predictions. Such methods would be useful because a predictor based on a small number of genes may be more biologically interpretable than one based on hundreds of genes. It would also be easier to convert such a predictor to a reverse transcriptase–polymerase chain reaction (RT-PCR) platform so that it could be used with formalin-fixed paraffin-preserved tissue. Unfortunately, attempts to independently verify the performance of some of these methods have been disappointing.[27,28]

How is a predictive classifier validated?

A cardinal principle for evaluating a predictive classifier is that the data used for developing the classifier should not be used in any way in testing the classifier. The simple *split-sample* method achieves this aim by partitioning the study samples into two parts. The separation is often done randomly, with half to two thirds of the cases used for developing the classifier and the remainder of the cases used for testing. The cases in the test set should not be used for determining which variables to include in the classifier, and they should not be used to compare different classifiers built in the training set. The cases in the test set should not be used in any way until a single completely specified model has been developed using the training data. At that time, the classifier is applied to the cases in the test set. For example, with an expression profile classifier, the classifier is applied to the expression profiles of the cases in the test set and each of them is classified as a responder or non-responder to the therapy. The patients in the test set have received the treatment in question and therefore which predictive classifications were correct and how many were incorrect can be enumerated. In using the split-sample method properly, a single classifier should be defined on the training data; it is not valid to develop multiple classifiers and then use their performance on the test data to select among the classifiers.[29]

There are more complex forms of dividing the data into training and testing portions. These *cross-validation* or *resampling* methods use the data more efficiently than in the simple division described above.[30] Cross-validation generally partitions the data into a large training set and a small test set. A classifier is developed on the training set and then applied to the cases in the test set to estimate the error rate; the procedure is repeated for numerous training-test partitions and the prediction error estimates are averaged. Molinaro et al showed that for small data sets (<100 cases), leave-one-out cross-validation or 10-fold cross-validation provided much more accurate estimates of prediction accuracy than did either the split-sample approach or averaging results over random replicated split-sample partitions. Michiels et al[31] suggested that multiple training-test partitions be used. However, the split-sample approach is mostly useful when a completely defined algorithm for developing the classifier is not available. When there is a single training set–test set partition, numerous analyses on the training set can be performed in order to develop a classifier, and biological considerations of which genes to include can be introduced before deciding on the single classifier to be evaluated on the test set. However, with multiple training-test partitions, that type of flexible approach to model development is not feasible. If there is a completely defined algorithm for classifier development, it is generally better to use one of the cross-validation approaches to estimate the error rate because the replicated split sample approach does not provide as efficient use of the available data.

In order to adhere to the key principle of not employing the same data to both develop and evaluate a classifier, it is essential that for each training-test partition the data in the test set is not used in any way.[32-34] Hence a model should be developed from scratch in each training set; in other words, multiple classifiers are developed in the process of performing cross-validation and these classifiers will in general involve different sets of genes. It is completely invalid to select the genes beforehand using all the data and then to just cross-validate the model building process for that restricted set of genes. Radmacher et al[33] and Ambroise and McLachlan[32] demonstrated that such pre-selection results in severely biased estimates of prediction accuracy. In spite of this severe bias, this error is made in many developmental classifier studies.[7] The estimate of prediction accuracy resulting from complete cross-validation is an internally valid and unbiased estimate of the prediction accuracy for the model developed using the full set of data. A wide variety of classification models, variable selection algorithms, and complete cross-validation methods are available in BRB-ArrayTools.[5]

How can you determine whether a predictive classifier is statistically significant?

For predictive classifiers, "statistically significant" should mean "predicts more accurately than chance." If a separate test set of cases is available, then it is easy to compute whether the prediction accuracy in the test set fulfills this criterion. However, the prevalence of the classes needs to be taken into account. For example, if 20% of cases are responders, then one can be correct 80% of the time by always predicting non-response. If cross-validation is used, the statistical significance of the cross-validated estimate of prediction error can be determined by repeating the cross-validation for permuted data as described by Radmacher et al[33]; this approach is preferable to the approach proposed by Michiels et al.[31]

How can you determine whether a predictive classifier adds predictive value to standard prognostic factors?

Statistical significance of a predictive classifier should not be evaluated by using cross-validated class predictions in a multivariate regression model. Many studies utilize this approach to establish that the genomic prediction model provides "independent prediction value" over established covariates. The approach is not valid, because the cross-validated predictions are not independent[35] and because it mistakes statistical significance of association measures with predictive value.[36] It is much more meaningful to evaluate the cross-validated predictions of a genomic classifier within the levels of an established staging system.

Can predictive classifiers be used with survival data?

Such classifiers are best developed without attempting to convert survivals to binary categories. Several methods have been developed for categorizing patients into risk groups based on gene expression data.[37,38] BRB-ArrayTools[5] builds a Cox proportional hazards model within each cross-validated training set using the top principal components of the genes that are most correlated with survival in that training set. This model is used to classify the test-set cases as high or low risk. After the cross-validation loops are complete, Kaplan-Meier curves are constructed of the survivals of the cases classified as high risk versus those classified as low risk. The statistical significance of the difference between the cross-validated Kaplan-Meier curves is determined by repeating the entire procedure many times with the gene expression profiles permuted. Permutation is necessary because the standard log-rank test is invalid for cross-validated Kaplan-Meier curves, as the data sets are not independent. This approach is also used to determine whether gene expression classifiers predict survival risk better than do standard covariates, as well as to build models using genes whose expression adds to those of the covariates.

What is the difference between a developmental study and a validation study?

Predictive classifiers are constructed in developmental studies; validation studies test pre-specified classifiers. Developmental studies should provide some internal estimate of predictive accuracy for the classifier developed, usually based on splitting the data into a training set and a test set or using cross-validation. They are, however, both types of internal validation. Taking one set of data collected and assayed under carefully controlled research conditions and splitting it into a training and testing set is not equivalent to evaluating the predictive accuracy of a classifier on a new set of patients from different centers with tissue collection and assay performance more representative of "real world" conditions.[36]

Developmental studies are often too limited in size, structure, and the nature of the cases to establish the medical utility of a predictive classifier. Even in the pre-genomic era, prognostic factor studies were often conducted using a convenience sample of available specimens from a heterogeneous group of patients who had received a variety of treatments. Classifiers that are prognostic for such a mixed group often have uncertain therapeutic relevance.[39] The Oncotype DX classifier is one example of a prognostic classifier that does have therapeutic value[40,41] because it was developed and validated using cases appropriate for a therapeutic decision context. Predictive classifiers that identify which patients respond to specific treatments are also often more valuable than the more commonly reported prognostic studies of heterogeneous patients. Currently there is considerable interest in using predictive classifiers to increase the efficiency and informativeness of new drug development.[42-45]

In planning a study to develop a predictive classifier, considerable care should be given to selecting cases so that the result has potential therapeutic relevance. Very often this objective can be enhanced by selecting cases who participated in an appropriate clinical trial. Whereas developmental studies often provide some measure of predictive accuracy for the classifiers, such estimates may not establish real medical utility.[4] Medical utility often requires establishing that the predictive classifier is more effective than are standard practice guidelines for enabling treatment selection that results in better patient outcome (or a similar outcome with less adverse events). Establishing medical utility depends on available treatment options and current standards of care. A key step in developing a useful predictive classifier is identification of a key therapeutic decision setting that can potentially be improved based on genomic data.

How can medical utility of a genomic test relative to standard practice guidelines be evaluated?

The gold standard evidence might be a controllzed clinical trial in which patients are randomized to two groups. In one group, treatment is determined using the genomic classifier; in the other, treatment is determined by standard practice guidelines. This type of clinical trial is generally very inefficient and requires so many patients as to be impractical. Inefficiency is the result of many if not most patients receiving the same treatment regardless of randomizatiuon assignment; consequently, a huge sample size is needed to detect the small difference in overall outcome resulting from a difference for the patients whose treatment assignment actually differs between the two groups. A more efficient design involves measuring the genomic test in all patients before randomization, and then only randomizing those whose treatment as based on the genomic test is different from that based on practice guidelines. This design is being currently used in the MINDACT trial to test the medical utility of the 70-gene signature developed by van't Veer et al.[46,47] Medical utility can be evaluated separately for subgroups defined by the ways practice guidelines differ from genomic classifier recommendation. For example the TAILORx trial evaluates whether practice guideline-based chemotherapy can be withheld from patients with node-negative, Her2/neu-negative, estrogen receptor-positive breast cancer with Oncotype DX recurrence score < 11.

If a sufficiently complete and adequately preserved set of archived specimens is available from an appropriate clinical trial, it may be possible to reliably evaluate medical utility by performing a prospectively specified analysis using retrospective data. Technical validation of the robustness of the assay for use with prospectively collected tissues could be established separately. The advantage of such a prospective-retrospective design is a strong motive for archiving tumor specimens in all major randomized clinical trials.

Why do predictive classifiers developed in different studies for the same types of patients use very different sets of genes for prediction?

Validating a predictive classifier means establishing that the classifier predicts accurately for independent data—not that the same genes would be selected in developing a classifier with independent data. This point is often misunderstood and is a source of inappropriate criticism of expression profiling studies. Expression levels among genes are highly correlated. For regression model building in such settings, there are many models that predict about equally well, even more so for genomic studies where the number of candidate variables is large relative to the number of cases.[48] It would take enormous numbers of cases to distinguish the small differences in predictive accuracy among such models,[49] but it is a very inappropriate criterion for sample size planning. Dobbin and Simon have shown that much smaller sample sizes are generally adequate to develop predictive classifiers with accuracy within 5 to 10 percentage points of the those that could be achieved with unlimited cases.[50,51] In the Dobbin and Simon method, the sample size for a training set is planned so as to develop the genomic classifier (the method is available at http://linus.nci.nih.gov/brb/samplesize/). A substantial number of additional cases will be needed for a test set that provides precise estimates of sensitivity and specificity, particularly to determine whether the classifier adds sufficiently to the predictive accuracy of standard prognostic factors.

Why are so many molecular predictors available in the literature but so few find a use in clinical practice?

Pasztai et al identified 939 articles on "prognostic factors" or "prognostic markers" in breast cancer over 20 years, but only three were widely used in practice.[52] Kyzas et al reviewed 340 articles on prognostic marker meta-analyses and 1,575 articles on cancer prognostic markers published in 2005: more than 90% of the articles reported statistically significant findings[53]! There are multiple factors that account for the discrepancy between the many positive reports in the literature and the lack of clinical utility of such markers.

One of the most important reasons for the discrepancy is that prognostic factors that do not help in therapeutic decision making are not generally used. Most of the literature reports are based on evaluating prognosis using "convenience samples" of specimens from heterogeneous patients without focus on specific therapeutic decisions. Prognostic markers have potential value for therapeutic decision-making only under very restricted circumstances. If one studies prognosis for a set of patients who are receiving limited local treatment only, then the prognostic marker may help identify patients who do not need systemic therapy. Unless the prognostic study is focused it is unlikely to be therapeutically relevant.

Studies of predictive markers are likely to be more useful. A predictive marker provides information on the likelihood of benefit from a specific treatment. To study a predictive marker using survival or disease-free survival as an end point, a substantial number of specimens from patients in a randomized clinical trial of the treatment of interest versus an appropriate control treatment is needed. If objective tumor response is the end point, then a randomized clinical trial is not a requirement, but the specimens must be from patients who received the treatment in question. Such studies are much less common than are unfocused reports of prognostic markers in mixed populations.

A second key reason for the discrepancy between the large number of reports of prognostic or predictive markers and low number used in practice is that for a test to be useful for therapeutic decision-making, there generally need to be two reasonable treatment options and this is often not the case. If there is one good treatment and the prognosis for untreated patients is poor, then few physicians will order a test to determine who to leave untreated. In the case of Oncotype DX, the prognosis for many node-negative, estrogen receptor–positive patients who received tamoxifen alone was good, so for that population there were two viable treatment options, tamoxifen alone or tamoxifen plus chemotherapy.[41] In some contexts, there may be two treatment options, but the test does not have sufficient positive and negative predictive value for clinical utility. Many developmental studies do not even recognize the importance of predictive value and overemphasize statistical significance.[53]

Finally, it is very difficult to sufficiently develop a test that can be reliably used in routine medical practice. A robust assay that can be used broadly must be developed and then technically validated to show reproducibility and robustness despite variations in tissue collection and reagents prospectively establishing medical utility.

Conclusion

As pointed out by Dupuy and Simon, microarray-based clinical investigations have generated both unrealistic hype and excessive skepticism. Genomic technologies are tremendously powerful and will play instrumental roles in elucidating the mechanisms of oncogenesis and in the coming era of predictive medicine in which treatments are tailored to individual tumors. Achieving these goals involves challenges in rethinking many paradigms for the conduct of basic and clinical cancer research and for the organization of interdisciplinary collaboration. Whole-genome technology provides power for both discovery and for generating erroneous claims. We need to provide appropriate training and interdisciplinary research settings to enable laboratory and clinical scientists to utilize genomic technology effectively in collaboration with statistical and computational scientists.

Acknowledgment

The author would like to acknowledge the valuable comments of Dr John Ioannidis on an earlier draft of the manuscript.

References

1. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, et al: Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med 353:1659-1672, 2005
2. Paik S, Taniyama Y, Geyer CE: Anthracyclines in the treatment of HER2-negative breast cancer. J Natl Cancer Inst 100:2-3, 2008
3. Kattan MW: Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 95:634-635, 2003
4. Simon R: When is a genomic classifier ready for prime time? Nat Clin Pract Oncol 1:2-3, 2004
5. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y: Analysis of gene expression data using BRB-ArrayTools. Cancer Informatics 2:11-17, 2007
6. Simon R, Korn E, McShane L, Radmacher M, Wright G, Zhao Y: Design and Analysis of DNA Microarray Investigations. New York, NY, Springer Verlag, 2003
7. Dupuy A, Simon R: Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99:147-157, 2007
8. Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 57:289-300, 1995
9. Storey JD: A direct approach to false discovery rates. J R Stat Soc B 64:479-498, 2002
10. Wu B, Guan Z, Zhao H: Parametric and nonparametric FDR estimation revisited. Biometrics 62:735-744, 2006
11. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98:5116-5121, 2001
12. Korn EL, Li MC, McShane LM, Simon R: An investigation of SAM and the multivariate permutation test for controlling the false discovery proportion. Stat Med 26:4428-4440, 2007
13. Storey JD, Xiao W, Leek JT, Tomkins RG, Davis RW: Significance analysis of time course microarray experiments. Proc Natl Acad Sci 102:12837-12842, 2005
14. Shih JH, Michalowska AM, Dobbin K, Ye Y, Qui TH, Green JE: Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. J Natl Cancer Inst 95:1362-1369, 2003
15. Dobbin K, Simon R: Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6:27-38, 2005
16. Shih JH, Michalowska AM, Dobbin K, et al: Effects of pooling mRNA in microarray class comparison. Bioinformatics 20:3318-3325, 2004
17. Subramanian A, Tamayo P, Mootha VK: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 102:15545-15550, 2005
18. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci 102:13544-13549, 2005
19. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E: Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. Neurochem Res 29:1213-1222, 2004
20. Kong SW, Pu WT, Park PJ: A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics 22:2373-2380, 2006
21. Goeman JJ, Buhlmann P: Analyzing gene expression data in terms of gene sets: Methodological issues. Bioinformatics 23:980-987, 2007
22. Xu X, Zhao Y, Simon R: Gene sets expression comparison in BRB-ArrayTools. Bioinformatics 24:137-139, 2008
23. Breiman L, Friedman JH, Olshen RA, Stone PJ: Classification and Regression Trees. Belmont, CA, Wadsworth International Group, 1984
24. Dudoit S, Fridlyand J: Classification in microarray experiments, in Speed T (ed): Statistical Analysis of Gene Expression Microarray Data. Boca Raton, FL, Chapman & Hall/CRC, 2003, pp 93-158
25. Dudoit S, Fridlyand J, Speed TP: Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97:77-87, 2002
26. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: Tissue classification with gene expression profiles. J Comput Biol 7:559-584, 2000
27. Wessels LFA, Reinders MJT, Hart AAM, Veenman CJ, Dai H, He T, et al: A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 21:3755-3762, 2005
28. Lai C, Reinders MJT, van't Veer LJ, Wessels LF: A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 7:235, 2006
29. Varma S, Simon R: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91, 2006
30. Molinaro AM, Simon R, Pfeiffer RM: Prediction error estimation: A comparison of resampling methods. Bioinformatics 21:3301-3307, 2005
31. Michiels S, Koscielny S, Hill C: Prediction of cancer outcome with microarrays: A multiple validation strategy. Lancet 365:488-492, 2005
32. Ambroise C, McLachlan GJ: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci 99:6562-6566, 2002
33. Radmacher MD, McShane LM, Simon R: A paradigm for class prediction using gene expression profiles. J Comput Biol 9:505-512, 2002
34. 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 95:14-18, 2003

35. Lusa L, McShane LM, Radmacher MD, Shih JH, Wright GW, Simon R: Appropriateness of inference procedures based on within-sample validation for assessing gene expression microarray-based prognostic classifier performance. Stat Med 26:1102-1113, 2007
36. Ioannidis JPA: Is molecular profiling ready for use in clinical decision making? Oncologist 12:301-311, 2007
37. Bair E, Tibshirani R: Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2:511-522, 2004
38. Gui J, Li H: Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21:3001-3008, 2005
39. Simon R: Evaluating prognostic factor studies, in Gospodarowicz MK (ed): Prognostic factors in cancer (ed 2). New York, NY, Wiley-Liss, 2002, pp 49-56
40. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817-2826, 2004
41. Paik S: Development and clinical utility of a 21-gene recurrence score prognostic assay in patients with early breast cancer treated with tamoxifen. Oncologist 12:631-635, 2007
42. Simon R: A roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 23:7332-7341, 2005
43. Simon R, Maitournam A: Evaluating the efficiency of targeted designs for randomized clinical trials: Supplement and correction. Clin Cancer Res 12:3229, 2006
44. Simon R, Maitournam A: Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10:6759-6763, 2005
45. Freidlin B, Simon R: Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 11:7872-7878, 2005
46. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mab M, et al: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536, 2002
47. Bogaerts J, Cardoso F, Buysa M, Braga S, Loi S, Harrison JA, et al: Gene signature evaluation as a prognostic tool: Challenges in the design of the MINDACT trial. Clin Pract Oncol 3:540-551, 2006
48. Fan C, Oh DS, Wessels L, et al: Concordance among gene-expression based predictors for breast cancer. N Engl J Med 355:560-569, 2006
49. Ein-Dor L, Zuk O, Domany E: Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci 103:5923-5928, 2006
50. Dobbin K, Simon R: Sample size planning for developing classifiers using high dimensional DNA expression data. Biostatistics 8:101-117, 2007
51. Dobbin KK, Zhao Y, Simon RM: How large a training set is needed to develop a classifier for microarray data? Clin Cancer Res 14:108-114, 2008
52. Pusztai L, Ayers M, Stec J, et al: Clinical application of cDNA microarrays in oncology. Oncologist 8:252-258, 2003
53. Kyzas PA, Denaxa-Kyza D, Ioannidis JP: Almost all articles on cancer prognostic markers quote statistically significant results. Eur J Cancer 43:2559-2579, 2007