Published in final edited form as:

Semin Hematol. 2008 July ; 45(3): 196–204. doi:10.1053/j.seminhematol.2008.04.008.

Interpretation of Genomic Data Questions and Answers

Richard Simon, D.Sc.

Biometric Research Branch, National Cancer Institute

Abstract

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 hyperbole and excessive skepticism. Genomic technologies are tremendously powerful and will play instrumental roles in elucidating the mechanisms of oncogenesis and in developing an era of predictive medicine in which treatments are tailored to individual tumors. Achieving these goals involves challenges in re-thinking many paradigms for the conduct of basic and clinical cancer research and for the organization of interdisciplinary collaboration.

We will address some key issues on the use of genomic technology in biomedicine. Our focus will be on cancer therapeutics, although many issues have broader relevance. We will address study design for both developmental and validation studies. We also address 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?

Why are genomic data important?

Is “the right treatment for the right patient” hype or substance?

What kinds of genomic data are available?

Is the challenge how to manage all of this data?

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

Can biologists and clinical investigators analyze genome-wide data?

Richard Simon, D.Sc., Chief, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda MD 20892-7434, 301.496-0975 (tel), 301.402-0560 (fax), rsimon@nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

What are the appropriate analysis methods?

What is class discovery?

Gene finding

What methods are appropriate for gene finding problems?

How many samples do you need for gene finding with expression data?

How do you relate lists of differentially expressed genes to pathways?

Prediction

How do prediction problems differ from gene finding problems?

What kinds of predictive classifiers are best?

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

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

Can predictive classifiers be used with survival data?

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

How can you evaluate whether a genomic test improves medical utility relative to standard practice guidelines?

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

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

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 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 guidelines³. Extensive clinical studies are needed to demonstrate that a genomic test is ready and appropriate for clinical use⁴.

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 can identify copy number variations and loss of heterozygosity on a genome wide basis. Genome wide genotyping is in use for identifying single nucleotide polymorphisms and in the next few years it will be economically feasible to completely re-sequence 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 that I developed (available at <http://linus.nci.nih.gov>) can easily handle 1000 expression profiles of 50,000 transcripts to develop predictive classifiers, fully cross-validated, on a personal computer within minutes⁵. 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 utility⁶. 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 study⁷. 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 kind of restrictions on data sharing that have been practiced for clinical trials data is 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 paper by Dupuy and Simon⁷ identified inappropriate use of cluster analysis as one of the most common flaws in published studies relating microarray gene expression to cancer outcome. The over-use 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 inter-disciplinary 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 multi-disciplinary 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 re-training scientists in the effective use of whole-genome data. The challenge isn't 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-ArrayTools⁵ 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 journals⁷. 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 apposite analysis. Many uses of microarrays can be categorized as (i) class discovery; (ii) gene finding or class comparison; (iii) 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. Cluster analysis methods always result in clusters, however, 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 examining for 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 the expression of which 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 is 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 outcome⁷

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, to take into account that differential expression is being compared for tens of thousands of genes.--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 Hochberg⁸. Let us 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, 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 SAM¹¹ 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-ArrayTools⁵.

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 intervention¹³; 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 Simon⁷.

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 two-fold 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-ArrayTools⁵. 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. This approach has some serious limitations, however. In order 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 analysis¹⁷, is one method of this type, 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 over-expression in the first class compared to the second class. In computing that enrichment score, however, 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 domain²².

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 discoveries²³. 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 which predicts a class from an expression profile. Specification of a class predictor requires (i) detailed description of the genes whose expression levels are to be utilized in the prediction; (ii) selection of the mathematical form used to combine the expression levels to the component genes; and (iii) characterization 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). In the development of class predictors using gene expression data, however, 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 which will fit that same data almost perfectly but be completely useless for prediction with independent data!

One 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. This is not the method by which prediction models have been developed. Traditionally, procedures like stepwise regression methods are used to select variables that make independent contributions to prediction and which work well together. 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. 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, 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 used for problems with many cases and few variables does not work well for genomic problems. Accurate prediction is possible in high dimensional ($p \gg n$) problems, but different methods of predictive modeling must be utilized.

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

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

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

$\hat{x}$ of values is predicted to be in class 1 or class 2 depending on whether $l(\hat{x})$ as computed from equation (1) is less than the threshold d or greater than d respectively. Many of the widely used classifiers are of the form shown in (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 (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 classification¹¹. 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 an 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 classifiers²⁹.

There are more complex forms of dividing the data into training and testing portions. These *cross-validation* or *re-sampling* methods utilize the data more efficiently than in the simple division described above³⁰. 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 (less than 100 cases), leave-one-out cross validation or 10-fold cross validation provided much more accurate estimates of prediction accuracy than 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. The split sample approach is mostly useful, however, when 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. With multiple training-test partitions however, 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 principal 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 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 McLachlan³² 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 studies⁷. 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-ArrayTools⁵.

How a predictive classifier determined to be 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. The prevalence of the classes needs to be taken into account, however. For example, if 20 percent of cases are responders, then one can be correct 80 percent 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 a predictive classifier be judged to add 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, despite that any 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 independent³⁵ and because it mistakes statistical significance of association measures with predictive value³⁶. 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-ArrayTools⁵ 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. Of note, these are 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 conditions³⁶.

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 relevance³⁹. 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 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 utility⁴. 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 guideline be evaluated?

\Standard evidence might be a controlled 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 randomization 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 al⁴⁶. The MINDACT trial uses the Adjuvant! Online website as standard practice guidelines. The superiority of the 70-gene classifier to Adjuvant! Online with regard to predictive accuracy was independently established by Buyse et al., but that accomplishment itself did not establish medical utility⁴⁷. An alternative approach to establishing medical utility is to perform the genomic test on patients for whom practice guidelines specify a particular treatment, and randomize those for whom the genomic test suggests a different therapy. This strategy underlies the TAILORx trial. Patients with node negative ER positive breast cancer with tumors greater than xx cm and Oncotype DX score between 11 and 25 are randomized to receive either standard practice chemotherapy or no chemotherapy. Both of these trials are large and will require long follow-up. If a sufficiently complete and adequately preserved set of archived specimens were available from an appropriate randomized clinical trial, it might be possible to perform a prospective analysis using retrospective data, which would certainly expedite the evaluation of medical utility. Technical validation of the robustness of the assay for use with prospectively collected tissues could be established separately. The advantages 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 cases⁴⁸. It would take enormous numbers of cases to distinguish the small differences in predictive accuracy among such models⁴⁹ 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 five to ten 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?

Puzstai identified 939 articles on “prognostic factors” or “prognostic markers” in breast cancer over 20 years, but only 3 were widely used in practice⁵². Kyzas et al. reviewed 340 articles on prognostic marker meta-analyses and 1575 articles on cancer prognostic markers published in 2005: over 90% of the articles reported statistically significant findings⁵³! 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 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 endpoint, 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 endpoint, 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--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 chemotherapy⁴¹. 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 over-emphasize statistical significance⁵³.

Finally, it is very difficult to sufficiently develop a gthat it 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 functionality despite variations in tissue collection and reagents. Commercial development requires experience and dedication and a viable business plan. Clinical validation that the test has medical utility for treatment selection compared to standard practice guidelines is an even more formidable. Such studies, to achieve the highest level of evidence, are prohibitively expensive except in the limited cases where government or charitable agencies provide funding, unless they can be conducted as a prospective analysis of archived specimens. If the development of genomic tests is linked to the development and approval of new molecularly targeted drugs, then some of these obstacles may be avoided or overcome

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 on the coming era of predictive medicine in which treatments are tailored to individual tumors. Achieving these goals involves challenges in re-thinking 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.

Acknowledgements

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

REFERENCES

1. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *New England Journal of Medicine* 2005;353:1659–1672. [PubMed: 16236737]
2. Paik S, Taniyama Y, Geyer CE. Anthracyclines in the treatment of HER2-negative breast cancer. *Journal of the National Cancer Institute* 2008;100:2–3. [PubMed: 18159066]
3. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *Journal of the National Cancer Institute* 2003;95:634–635. [PubMed: 12734304]
4. Simon R. When is a genomic classifier ready for prime time? *Nature Clinical Practice: Oncology* 2004;1:2–3.
5. Simon R, Lam A, Li MC, et al. Analysis of gene expression data using BRB-ArrayTools. *Cancer Informatics* 2007;2:11–17.
6. Simon, R.; Korn, EL.; McShane, LM., et al. *Design and Analysis of DNA Microarray Investigations*. New York: Springer Verlag; 2003.
7. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute* 2007;99:147–157. [PubMed: 17227998]
8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 1995;57:289–300.
9. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* 2002;64:479–498.
10. Wu B, Guan Z, Zhao H. Parametric and nonparametric FDR estimation revisited. *Biometrics* 2006;62:735–744. [PubMed: 16984315]
11. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Science* 2001;98:5116–5121.
12. Korn EL, Li MC, McShane LM, et al. An investigation of SAM and the multivariate permutation test for controlling the false discovery proportion. *Statistics in Medicine*. 2008(In press),
13. Storey JD, Xiao W, Leek JT, et al. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Science* 2005;102:12837–12842.
14. Dobbin K, Shih J, Simon R. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *Journal of the National Cancer Institute* 2003;95:1362–1369. [PubMed: 13130111]
15. Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* 2005;6:27–38. [PubMed: 15618525]
16. Shih JH, Michalowska AM, Dobbin K, et al. Effects of pooling mRNA in microarray class comparison. *Bioinformatics* 2004;20:3318–3325. [PubMed: 15247103]
17. Subramanian A, Tamayo P, Mootha VK. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Science* 2005;102:15545–15550.
18. Tian L, Greenberg SA, Kong SW, et al. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Science* 2005;102:13544–13549.
19. Pavlidis P, Qin J, Arango V, et al. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research* 2004;29:1213–1222. [PubMed: 15176478]
20. Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 2006;22:2373–2380. [PubMed: 16877751]
21. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007;23:980–987. [PubMed: 17303618]
22. Xu X, Zhao Y, Simon R. Gene sets expression comparison in BRB-ArrayTools. *Bioinformatics*. 2008 (In press),
23. Breiman, L.; Friedman, JH.; Olshen, RA., et al. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group; 1984.

24. Dudoit S, Fridlyand J. Classification in microarray experiments. In: Speed, T., editor. Statistical analysis of gene expression microarray data. Chapman & Hall/CRC; 2003.
25. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 2002;97:77–87.
26. Ben-Dor A, Bruhn L, Friedman N, et al. Tissue classification with gene expression profiles. *Journal of Computational Biology* 2000;7:559–584.
27. Wessels LFA, Reinders MJT, Hart AAM, et al. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* 2005;21:3755–3762. [PubMed: 15817694]
28. Lai C, Reinders MJT, Veer LJvt, et al. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics* 2006;7:235. [PubMed: 16670007]
29. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;7:91. [PubMed: 16504092]
30. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: A comparison of resampling methods. *Bioinformatics* 2005;21:3301–3307. [PubMed: 15905277]
31. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: A multiple validation strategy. *The Lancet* 2005;365:488–492.
32. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Science* 2002;99:6562–6566.
33. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. *Journal of Computational Biology* 2002;9:505–512.
34. Simon R, Radmacher MD, Dobbin K, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003;95:14–18. [PubMed: 12509396]
35. Lusa L, McShane LM, Radmacher MD, et al. Appropriateness of inference procedures based on within-sample validation for assessing gene expression microarray-based prognostic classifier performance. *Statistics in Medicine* 2007;26:1102–1113. [PubMed: 16755534]
36. Ioannidis JPA. Is molecular profiling ready for use in clinical decision making? *The Oncologist* 2007;12:301–311. [PubMed: 17405894]
37. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology* 2004;2:511–522.
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* 2005;21:3001–3008. [PubMed: 15814556]
39. Simon R. Evaluating prognostic factor studies. In: Gospodarowicz, MK., editor. *Prognostic factors in cancer*. ed 2nd. New York, NY: Wiley-Liss; 2002. p. 49–56.
40. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* 2004;351:2817–2826. [PubMed: 15591335]
41. Paik S. Development and clinical utility of a 21-gene recurrence score prognostic assay in patients with early breast cancer treated with Tamoxifen. *The Oncologist* 2007;12:631–635. [PubMed: 17602054]
42. Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology* 2005;23:7332–7341. [PubMed: 16145063]
43. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: Supplement and Correction. *Clinical Cancer Research* 2006;12:3229.
44. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 2005;10:6759–6763. [PubMed: 15501951]
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. *Clinical Cancer Research* 2005;11:7872–7878. [PubMed: 16278411]
46. van't-Veer LJ, Dai H, Vijver MJvd, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–536. [PubMed: 11823860]

47. Buyse M, Loi S, Veer Lvt, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute* 2006;98:1183–1192. [PubMed: 16954471]
48. Fan C, Oh DS, Wessels L, et al. Concordance among gene-expression based predictors for breast cancer. *New England Journal of Medicine* 2006;355:560–569. [PubMed: 16899776]
49. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Science* 2006;103:5923–5928.
50. Dobbin K, Simon R. Sample size planning for developing classifiers using high dimensional DNA expression data. *Biostatistics* 2007;8:101–117. [PubMed: 16613833]
51. Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clinical Cancer Research*. 2008(in press),
52. Pusztai L, Ayers M, Stec J, et al. Clinical application of cDNA microarrays in oncology. *The Oncologist* 2003;8:252–258. [PubMed: 12773747]
53. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers quote statistically significant results. *European Journal of Cancer* 2007;43:2559–2579. [PubMed: 17981458]