Published in final edited form as:

Semin Cancer Biol. 2008 October ; 18(5): 305–310. doi:10.1016/j.semcancer.2008.03.001.

Using microarrays to study the microenvironment in tumor biology: The crucial role of statistics

Stuart G. Baker^{a,*} and Barnett S. Kramer^b

^a Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA

^b Office of Disease Prevention, National Institutes of Health, Bethesda, MD, USA

Abstract

Microarrays represent a potentially powerful tool for better understanding the role of the microenvironment on tumor biology. To make the best use of microarray data and avoid incorrect or unsubstantiated conclusions, care must be taken in the statistical analysis. To illustrate the statistical issues involved we discuss three microarray studies related to the microenvironment and tumor biology involving: (i) prostatic stroma cells in cancer and non-cancer tissues; (ii) breast stroma and epithelial cells in breast cancer patients and non-cancer patients; and (iii) serum associated with wound response and stroma in cancer patients. Using these examples we critically discuss three types of analyses: differential gene expression, cluster analysis, and class prediction. We also discuss design issues.

Keywords

Bonferroni; Class prediction; Cluster analysis; Differential expression; False discovery rate; Sample size

1. Introduction

Gene expression microarrays are potentially powerful tools for studying the role of the microenvironment and tumor biology. However, to obtain the most benefit from microarray studies and avoid pitfalls in analysis, an appreciation of the statistical complexity is crucial. It is not sufficient to apply microarray analysis software and report results without an understanding of the underlying statistical issues. Such a procedure can lead to a false sense of complacency, many false leads, and resulting squandered resources. The reason is that with so many genes under consideration, subtle issues of noise and variability need to be appreciated even when applying off-the-shelf methods.

* Corresponding author at: National Cancer Institute, EPN 3131, 6130 Executive Boulevard MSC 7354, Bethesda, MD 20892-7354, USA. Tel.: +1 301 496 7708; fax: +1 301 402 0816. E-mail address: sb16i@nih.gov (S.G. Baker). URL: <http://www.cancer.gov/prevention/bb/baker.html> (S.G. Baker).

Publisher's Disclaimer: This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues. Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit: <http://www.elsevier.com/copyright>

We discuss these issues in the context of three microarray studies investigating the role of the microenvironment in tumor biology. Many of our recommendations are similar to those from other reviews of statistical methods for microarrays but some differences arise from a different perspective and from the focus on the impact of the microenvironment on tumor biology.

2. Microarray studies of the microenvironment and tumor biology

We selected three microarray studies of the role of the microenvironment in tumor biology. The salient features of each of these studies are described below.

2.1. Prostatic stroma cells: cancer versus no cancer

Zhao et al. [1] used microarrays to compare gene expression of prostatic stromal cells in diseased versus normal tissues. They profiled gene expression among 18 samples of prostatic stromal cell cultures, 2 from normal transition zone (TZ), 7 from normal peripheral zone (PZ), 4 from benign prostatic hyperplasia (BPH), and 5 from cancer (CA). Using hierarchical clustering, they noted a separation between TZ and BPH samples and also a separate cluster of CA samples. They identified 34 genes differentially expressed between BPH and CA samples with a false positive rate (FPR) of 18% and 101 genes differentially expressed between PZ and CA samples with a FPR of 12%. They concluded that prostatic stromal cells have unique transcriptional programs, depending on the underlying state of health or disease.

2.2. Breast stroma and epithelial cells in breast cancer patients and non-cancer patients

Finak et al. [2] used microarrays to investigate gene expression in four types of histologically normal cells in 66 samples from 44 patients, 25 samples from epithelial duct cells adjacent to breast cancer, 26 samples from stromal cells adjacent to breast cancer, 9 samples from epithelial duct cells in patients undergoing cosmetic breast reduction surgery, and 6 samples from stromal cells in patients undergoing cosmetic breast reduction surgery. A cluster analysis on all genes and on 883 genes differentially expressed between stroma and epithelial tissue indicated that epithelium versus stroma, and not type of patient, was the dominant explanation for the clustering. A class prediction algorithm showed that gene expression in subjects with cancer did not predict clinical characteristics of the tumor. They concluded that there was no significant difference in gene expression between morphologically normal tissue adjacent to cancer and in breast reduction tissue.

2.3. Serum associated with wound response and stroma in cancer patients

Chang et al. [3] compared gene expression in 50 fibroblast cultures exposed to either serum associated with wound response or minimal amounts of serum. They found 677 genes with differential expression, which was reduced to 512 after eliminating those genes with expression related to cell cycle. They called these 512 genes the fibroblast core serum response (CSR) and label their expression in the presence and absence of serum as “serum-induced” and “serum-repressed”, respectively. Based on a cluster analysis of CSR genes in microarray experiments involving human cancers and normal cells, they reported all normal cells exhibited the serum-repressed signature, and almost all prostate and hepatocellular tumors exhibited the serum-induced signature. Because only some breasts, lungs, and gastric carcinomas exhibited the serum-induced signature, they concluded that the “wound healing phenotype” was a variable feature in these cancers. Using published microarray data from early breast cancer patients, they identified two clusters for which survival differed greatly. They concluded that there was a possible link between cancer progression and wound healing and that CSR are powerful predictors of clinical course.

3. Identifying differentially expressed genes

All of the microarray studies in Section 2 discussed selection of genes differentially expressed between two classes. Sometimes this process is called class comparison or outcome-related gene finding [4]. The selection of these genes involves two steps. The first step is to rank the genes by some measures. A popular measure for ranking is the t -statistic which equals the difference in average levels of expression between two classes divided by its standard error. The general consensus, with which we agree, is not to select genes simply by the difference in average levels of expression, because it is important to consider the underlying variability. The t -statistic is essentially a measure of the signal above the background noise.

Zhao et al. [1] and Finak et al. [2] ranked genes using a measure related to the t -statistic that is computed as part of significance analysis of microarrays (SAMs) [5]. Chang et al. [3] reported “selection for genes that were concordantly induced or repressed” which may be accounting for variability because the probability of concordance depends on both signal and noise.

Once the genes are ranked the next step is to select a cutoff level for how many of the highest ranking genes should be reported as differentially expressed. The general consensus is that this selection should be based on the false discovery rate of the set of genes under consideration [4]. The false discovery rate is the expected fraction of genes incorrectly called differentially expressed. According to one review article, a false discovery rate of 10–20% is desirable [4]. Zhao et al. [1] reported false discovery rates of 12 and 18%.

An alternative method is to select all genes satisfying statistical significance under the Bonferroni approach. (If one gene satisfies the Bonferroni criterion the higher ranking genes will also satisfy it.) The Bonferroni approach says that a gene expression is statistically significant if the p -value for a statistic measuring expression levels (such as a t -statistic) is smaller than 0.05 divided by the total number of genes examined. Under the conservative assumption that gene expression is independent among all genes, the Bonferroni approach ensures the probability of at least one gene having a statistic as extreme or more extreme than the observed level simply by chance is not more than 0.05. There are also other, more complicated approaches, for limiting the probability of at least one chance observation.

Some studies have reported both differential selections based on Bonferroni and false discovery methods. In one study with a sample size of 12 with 22,000 genes, 32 genes were selected as differentially expressed when the false discovery rate was 10% while only 4 genes were selected when the overall type I error was 10% [6]. In an early and influential study on the effect of radiation on cell cultures (which described the SAM method) involving 6800 genes in 8 samples, 0 genes were selected as differentially expressed genes under the Bonferroni criterion (the overall type I error was not reported), but 42 were selected as differentially expressed when using the false discovery rate of 18% [5].

Results such as these have led many investigators to conclude that the Bonferroni approach is too stringent. A review article on statistical methods for microarray analysis called the Bonferroni adjustment “reflexive, draconian, and largely unresponsive to the goals of biologists” [7]. The underlying viewpoint is perhaps best summarized by Gustano et al. [8] who wrote “this approach [Bonferroni type adjustments] is too conservative, as it will severely restrict the number of differentially expressed genes identified in the microarray analysis and miss a significant number of truly differentially expressed genes... which will impact the ability of the study to provide a global picture of expression patterns across the genome, which is the purpose of the microarray study in the first place.”

In our view, the above argument that Bonferroni method is too conservative is driven by the premise that false positives have little cost in microarray research. In this regard, it is instructive

to consider the experience of investigators searching the genome for polymorphisms related to cancer. Because false positives have plagued genome studies, there is a strong support for Bonferroni adjustment in this field in order to best identify genetic polymorphism that is truly associated with cancer [9]. This practice in genome research is justified for the high cost of a false positive. If the microarray analysts also find too many fruitless “leads” from differential mRNA expression studies, Bonferroni adjustment (or related methods) may have a resurgence of interest.

4. Cluster analysis

Statistical clustering methods are very common in microarray studies and were used in all the aforementioned examples involving the role of the microenvironment on tumor development. Clustering methods are used to group either genes or specimens with similar expression patterns, and the most popular clustering method is hierarchical clustering [4]. The following is a brief description of the hierarchical clustering method of Eisen et al. [10] that is often applied to microarray studies. Eisen et al. [10] described their method in terms of clustering of genes but the most common application is clustering of specimens. In terms of clustering of specimens the method can be described as follows. A measure of similarity is computed for the gene expression levels for each possible pair of specimens using a statistic similar to a correlation coefficient. The pair of specimens with the highest level of similarity of gene expression is selected as a node. The paired specimens forming the node are replaced by the averages of the gene expression levels of the paired specimens. Treating this average of expression levels as corresponding to a single specimen for the node, the process is repeated. The results are typically displayed in tree diagram with the specimens ordered by a measure of similarity. The bottom branches of the tree are nodes connecting the original specimens, higher branches are nodes connecting previous nodes.

Cluster analysis for microarrays was proposed because differential expression of individual genes was thought not to “address the full potential of genome scale experiments to alter our understanding of cellular biology by providing, through an inclusive analysis of the entire repertoire of transcripts, a continuing comprehensive window into the state of the cell as it goes through a biological process” [10]. Cluster analysis was touted as “a holistic approach... allowing biologists to develop an integrated understanding of the process being studied” [10]. While this goal of cluster analysis is laudable, cluster analysis has several important limitations.

One common problem with cluster analysis is inappropriate application to a subset of genes that have been selected based on differential expression between known classes. With simulated data having identical distributions of gene expression in two classes (so there are no clusters), cluster analysis incorrectly identifies the two classes as two distinct clusters [4]. In other words, applying cluster analysis to genes selected based on differential expression between classes increases the likelihood that the clusters will correspond to the classes regardless of whether or not there is truly differential expression between the classes. Although Finak et al. [2] correctly applied cluster analysis to the original genes (and identified separate clusters for stroma and epithelial specimens), they also inappropriately applied cluster analysis to a subset of genes differentially expressed between stroma and epithelial. This application increased the likelihood that the clusters would be dominated by stroma and epithelium cells, even if there were no differential expression of genes between stroma and epithelial cells.

Another problem with standard cluster analysis is that it contains no information on how the variability in the data affects the clustering. If the samples were randomly split into two parts and a separate cluster analysis were performed on each part, very different clusters would likely be created. However, more sophisticated methods of cluster analysis that assign a probability of occurrence to each cluster can address this problem [11], as used by Finak et al. [2].

A third problem with cluster analysis is that it is sometimes used as informal substitute for the more rigorous class prediction approach described next. This use of cluster analysis in this role is undesirable because it does not make full use of the data on classes, it does not provide a measure of accuracy of prediction, and yields more subjective conclusions than with class prediction because of the reliance on visual inspection. Finak et al. [2] used cluster analysis to conclude that gene expression differed between stroma and epithelial but not type of patient (cancer or not), when class prediction could have been better used. Also Chang et al. [3] used cluster analysis to conclude that serum-repressed gene signatures appear in normal tissue while serum-induced gene signatures appear in cancer tissue, instead of better using class prediction. Perhaps the informal nature of this analysis contributed to their explanation that lack of fit was due to the variability in the “wound healing phenotype.”

5. Class prediction

Class prediction is the formulation of a rule among genes, often called a gene signature, that distinguishes classes (such as cancer versus normal). A simple class prediction rule would be a linear combination of levels of gene expression yielding a score. If the score for a specimen is above a certain threshold, the specimen would be classified as cancer; otherwise it would be classified as normal. Class prediction differs from differential expression because it explicitly considers an ensemble of genes and their interactions, while differential expression is limited to investigating genes one at a time. Class prediction differs from cluster analysis because it uses known classes to formulate a rule instead of formulating a rule based solely on variations in data (in this regard, class prediction is often called a supervised method because the analysis is “supervised” by information on class, while clustering analysis, which does not use data on class, is called an unsupervised method). Class prediction also differs from differential gene expression and cluster analysis, because performance is measured using a clear, rigorous approach involving a test sample that is independent of the training sample used to formulate the rule (to be discussed). Thus, class prediction is perhaps the most important tool for the analysis of microarray data. Unfortunately, it was underutilized in the examples in Section 2. There are many examples of correct use of class prediction in literatures [5,12,13], and, in terms of selecting the number of genes to include, we will focus on one particular study [14].

An important consideration in class prediction is the avoidance of overfitting, which means that the classification rule captures so many nuances of one particular set of data used for estimation that it performs poorly on new data. One method for avoiding overfitting is to formulate the classification rule in a training sample and evaluate its performance in a separate independent test sample. Another approach is cross-validation in which the data are systematically split into various training and test samples (such as 10 splits, each removing a different 1/10 of the specimens to form a test sample) with performance estimated in the test sample and averaged over all splits. Yet another approach is multiple random validation, in which the data are repeatedly randomly split into training and test samples, with performance measured in the test samples, to provide a measure of the variability of the performance estimates [12].

There are two general methods for formulating a classification rule in the training sample. One approach is to first select genes with the highest differential expression (in a process called filtering) and then formulate the remaining part of the classification rule based on these genes. Another approach, called a wrapper, applies cross-validation to *only* the training sample to create a classification rule that includes gene selection as part of an iterative approach of rule creation and *preliminary* performance evaluation.

Although some very complicated classification rules have been proposed, empirical studies have shown that simple classification rules may perform as well or better than complicated

classification rules [13]. Also simple classification rules are often easier to implement and understand. One simple rule is classification by nearest centroid after filtering genes. The centroid for a class is the mean of the expression level for a list of genes based on specimens in the class. The classification rule assigns a specimen to the class whose centroid is closest to the gene signature of the specimen, based on a distance measure. One simple distance measure is the length of a straight line in multidimensional space, which equals the square root of the sum of the squared differences between the gene levels in the specimen and in the centroid. If the distributions of expression levels in each class are multivariate normal, another distance measure is basically a straight line distance weighted by the inverse of the covariance matrix and the approach is called linear discriminant analysis. One variant of classification by nearest centroid (including a filtering algorithm) is the prediction around medioids (PAMs) method [15], which filters genes using a modified *t*-statistic and then shrinks centroids. Finak et al. [2] used PAM to classify gene signatures into classes defined by either clinical or pathological features. It should be noted that, in another microarray analysis, the PAM method performed worse than classification by nearest centroid with filtering based on unmodified *t*-statistic [16].

Once the classification rule is formulated in the training sample its performance is evaluated in a separate independent test sample. To avoid overfitting bias when measuring performance, it is important that the final classification rule from the training set be applied to the test sample (as opposed to using data from the test sample in the formulation of the classification rule), including filtering genes in the test sample before applying the remaining part of the classification rule from the training sample [17]. Finak et al. [2] followed this recommendation.

With two classes, the performance of class prediction is best summarized via the false positive rate (FPR) and the true positive rate (TPR). The FPR is the probability that a gene in class 0 is classified as being in class 1 and the TPR is the probability that a gene in class 1 is classified as being in class 1. (The FPR should not be confused with the false discovery rate discussed previously with regard to differential gene expression.) Sometimes performance is measured in terms of accuracy, which often means $\text{TPR} + (1 - \text{FPR})$. However, separate estimates for FPR and TPR are preferable because there may be different “costs” to incorrect classification of class 0 or class 1. Prediction is often based on a score that combines levels of expression of various genes into a single number, and various cut-points on the score give different pairs of FPR and TPR. Prediction is sometimes also measured using survival curves.

Many methods for class prediction select a moderate or large number of genes as part of the formulation of the classification rule in the training sample. In contrast, we recommend selecting only a few genes for creating the classification rule in the training sample. Selecting only a few genes allows a better focus on the science, particularly in light of the results of one microarray study that found that classification rules based on completely different choices of all 70 yielded similar performances [18]. Also selecting a few genes for the classification rule has the mathematical justification that after a few predictors are included in a classification rule, the inclusion of additional predictors typically yields only a slight improvement in performance [19]. This result has been confirmed empirically in microarray studies [14].

We argue for investigating 1, 5, 10, and 20 possible genes in the classification rule and comparing performance in the test sample to select the best classification rule. Although this approach, strictly speaking, violates the aforementioned recommendation against using data from the test sample to formulate the classification rule, overfitting bias is likely to be minor because only a few models are considered. Alternatively one could simply fit the five-gene model which, based on theoretical considerations [19] and empirical results [14], would likely have similar performance as classification rules with more genes.

Another important issue is stability, which means selecting the same highest ranking gene or genes in most randomly selected training sample as part of multiple random validation [12, 14]. If the performance in the test sample is good, genes that are selected with high frequency are most likely to have biological or clinical significance. The identification of genes with high frequency is best achieved with a classification rule involving only a few genes. With a classification rule involving many genes, a large fraction of genes will be likely be selected with high frequency so that “noise” will overwhelm any “signal” [14].

6. Designing microarray studies

When a microarray study is proposed it should have a clear statement of the goal. Examples from Section 2 could be (i) compare gene expression in prostate stroma versus cancer cells, (ii) compare gene expression in specimens from breast cancer patients versus specimens from non-cancer patients (separately for stromal and epithelial tissue), and (iii) compare gene expression of wound response genes in specimens from patients with and without cancer.

In the design of a microarray studies, it is important to avoid any systematic bias. This means that the handling, storage, time of analysis, should be the same for different classes (such as cancer versus no cancer). Also informative baseline variables (e.g. age, sex, smoking habits) should be similar [20].

Also, particularly for studies of the role of gene expression on the microenvironment, it is also very important to consider whether the design affects the relevance of results. For example Zhao et al. [1] compared gene expression of different types of stromal cells following in vitro culture and implicitly assumed that the results would apply in vivo. However, cell culturing can change phenotype and hence gene expression pattern, making the link from in vitro to in vivo results very tenuous.

Another important issue is sample size. Sample size calculations differ for differential expression, cluster analysis, and class prediction. We discuss only on the sample size calculation for class prediction because, as we have argued, class prediction is more useful than analyses of differential expression or cluster analysis. Because training and test samples have different goals, the sample size calculations differ for each, and we discuss each in turn.

For the size of the training sample, we propose a simple calculation that requires only specification of target and null FPRs and TPRs. The size of the training sample is specified so that an observed target performance is unlikely to have occurred by chance if the true performance was at minimal acceptable level, such as $FPR = 0.30$ and $TPR = 0.07$. Occurrence by chance is based on the premise that the classification rules are dominated by a single gene, so that number of classification rules under consideration is effectively the number of genes. This means that a chance occurrence would correspond to a p -value of 0.05 divided by the number of genes. Computations are based on the binomial model. See Appendix A for additional details. Dobbin and Simon [21] proposed another calculation for the size of the training sample based on the difference between the estimated class prediction function and a class prediction function if the sample size were infinite. Their method requires an estimate of the largest effect size of any gene. Although their method was very different from ours, their recommendation of 20–30 specimens per class (40–60 total) is comparable to a size between our proposed sample sizes for $FPR = 0.1$ and $TPR = 0.9$ and $FPR = 0.05$ and $TPR = 0.95$.

For the size of the test sample we follow similar calculations in Baker et al. [22] and Dobbin and Simon [21]. We chose a size to yield a given standard error for a target performance level, with computations based on a binomial distribution for false and true positives. See Appendix B for more details.

As a check, the sample size of the microarray study by Golub et al. [23], for which excellent classification performance has been achieved, involved 92 specimens which agrees with one of the most optimistic performance scenarios (for training and test samples combined) in Table 1.

7. Discussion

In summary, microarrays are a potentially powerful tool for understanding how the microenvironment affects tumor biology. Importantly, the usefulness of a microarray approach will be dramatically enhanced by carefully applying an appropriate statistical design and analysis with a sufficiently large size of samples.

One major recommendation is a larger sample size than often used. A total size of 100 in the training sample and 100 in the test sample (with equal numbers in each class) should be adequate in many studies. Although many investigators will be challenged to obtain such a large sample size, it is preferable to using too small a sample size and getting inconclusive results. In fact, sample sizes that are too small for the task may lead to wasted resources if an inconclusive study triggers attempts to replicate the results using precious biospecimens in subsequent studies.

A second major recommendation is to emphasize the class prediction with at most five genes in the classification rule (gene signature). Unlike methods of differential gene expression, class prediction accounts for interactions among genes. Unlike cluster analysis, class prediction provides a simple rigorous framework for validation. By using only five genes in the class prediction rule, investigators have a better chance of identifying reproducible genes that may be the most important for understanding etiology. These recommendations were part of a recently proposed microarray study for the role of the microenvironment in tumor biology associated with foreign-body carcinogenesis [24].

Appendix

Appendix A. Sample size for training sample

Let FPR_0 and TPR_0 denote the true level of performance that one would like to reject in favor of higher level of performance. Let FPR and TPR denote the false and TPRs for which one would like to reject FPR_0 and TPR_0 in favor of a higher level of performance. Informally, FPR and TPR are the levels of performance that, if observed in the training sample, would let one conclude that there is good classification not due to chance.

Let $2n$ denote the sample size (with size n for class 0 and size n for class 1) and let $f = n \text{ FPR}$ and $t = n \text{ TPR}$, rounded to the nearest integer. To call the performance better than chance under FPR_0 and TPR_0 , one would like to detect f or fewer false positives and t or greater true positives. Based on a binomial distribution, the probability of observing f or fewer subjects with false

positives when the true FPR is FPR_0 is $p_F(n) = \sum_{i=0}^f \text{BIN}(n, i) FPR_0^i (1 - FPR_0)^{n-i}$, where $\text{BIN}(n, i)$ is the binomial coefficient denoting n choose i . Similarly the probability of observing t or

greater subjects with true positives when the TPR is $p_T(n) = \sum_{i=t}^n \text{BIN}(n, i) TPR_0^i (1 - TPR_0)^{n-i}$.

The overall probability is $p(n) = p_F(n)p_T(n)$. Let m denotes the number of genes and we assume that classification rules are dominated by a gene so there are approximately m classification rules, leading to a Bonferroni adjusted significance level of $0.05/m$. Tabulating $p(n)$ for various n , we solve for n such that $p(n) = 0.05/m$.

Appendix B. Sample size for test sample

We recommend choosing a sample size to yield a desired level of precision for the target FPR and TPR. Let s_F and s_T denote the desired standard errors for FPR and TPR, respectively and let n_F denotes the sample size for class 0 and n_T denotes the sample size for class 1. Under a binomial distribution $FPR(1 - FPR)/n_F = s_F^2$ and $TPR(1 - TPR)/n_T = s_T^2$ and we solve for n_F and n_T (except when $FPR = 0$, $TPR = 1$, in which case this formula is not applicable because the variance is 0). The total sample size is $n_F + n_T$.

References

1. Zhao HJ, Ramos CF, Brooks JD, Peehl DM. Distinctive gene expression of prostatic stromal cells cultured from diseased versus normal tissues. *J Cell Physiol* 2007;210:111–21. [PubMed: 17044071]
2. Finak G, Sadekova S, Pepin F, Hallett M, Meterissian S, Halwani F, Khetani K, Souleimanova M, Zabolotny B, Omeroglu A, Park M. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res* 2006;8:R58. [PubMed: 17054791]
3. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K, Chi JT, van de Rijn M, Botstein D, Brown PO. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2004;2:206–14.
4. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* 2007;99:147–57. [PubMed: 17227998]
5. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;98:5116–21. [PubMed: 11309499]
6. Bretz F, Landgrebe J, Brunner E. Multiplicity issues in microarray experiments. *Methods Inf Med* 2005;44:431–7. [PubMed: 16113769]
7. Allison DB, Cui XQ, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006;1:55–65. [PubMed: 16369572]
8. Gustano A, Calza S, Pawitan Y. Identification of differentially expressed genes and false discovery rate in microarray studies. *Curr Opin Lipidol* 2007;18:187–93. [PubMed: 17353668]
9. Hunter DJ, Kraft P. Drinking from the fire hose—statistical issues in genomewide association studies. *N Engl J Med* 2007;357:436–9. [PubMed: 17634446]
10. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* 1998;95:1483–68.
11. Suzuki R, Shimodaira H. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006;22:1540–2. [PubMed: 16595560]
12. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005;365:488–92. [PubMed: 15705458]
13. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002;97:77–87.
14. Baker, SG.; Kramer, BS. Identifying genes that contribute most to good classification in microarrays; *BMC Bioinformatics*. 2006. p. 407<http://www.biomedcentral.com/1471-2105/7/407>
15. Tibsharini R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunk centroids of gene expression. *Proc Natl Acad Sci USA* 2002;99:6567–72. [PubMed: 12011421]
16. Dabney AR. Classification of microarrays to the nearest centroids. *Bioinformatics* 2005;21:4125–48. [PubMed: 16144805]
17. Simon R, Radmach MD, Dobbin K, McShane L. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;95:14–7. [PubMed: 12509396]
18. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 2005;21:171–8. [PubMed: 15308542]
19. Hand DJ. Classifier technology and the illusion of progress. *Stat Sci* 2006;21:1–14. [PubMed: 17906740]

20. Ransahoff DF. Bias as a threat to the validity of cancer molecular marker research. *Nat Rev Cancer* 2005;5:142–9. [PubMed: 15685197]
21. Dobbin KK, Simon RM. Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* 2007;8:101–17. [PubMed: 16613833]
22. Baker, SG.; Kramer, BS.; Srivastava, S. Markers for early detection of cancer: statistical issues for nested case-control studies; *BMC Med Res Methodol.* 2002. p. 4<http://www.biomedcentral.com/1471-2288/2/4/>
23. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531–7. [PubMed: 10521349]
24. Baker, SG.; Kramer, BS. Paradoxes in carcinogenesis: new opportunities for research directions; *BMC Cancer.* 2007. p. 151(06 Aug 2007). <http://www.biomedcentral.com/1471-2407/7/151>

Table 1
Sample sizes for training and test samples in microarrays

Target performance of class prediction rule	Training sample ^a , number of genes		Test sample ^b , target standard error		
	10,000	3000	0.03	0.05	0.07
FPR = 0.2, TPR = 0.8	336	296	356	128	66
FPR = 0.1, TPR = 0.9	82	76	200	72	37
FPR = 0.05, TPR = 0.95	48	44	106	38	20

^a Size needed to reject chance as explanation for target performance under indicated number genes if true performance is FPR = 0.3 and TPR = 0.7.

^b Size needed to obtain target standard error under target performance.