The challenges of gene expression microarrays for the study of human cancer

Anna V. Tinker,[1] Alex Boussioutas,[1,2] and David D.L. Bowtell,[1,3,*]

[1]Ian Potter Centre for Cancer Genomics and Predictive Medicine, Peter MacCallum Cancer Centre, St. Andrew's Place, East Melbourne, 3002, Victoria, Australia
[2]Department of Medicine, Royal Melbourne and Western Hospitals, University of Melbourne, Footscray, 3011, Victoria, Australia
[3]Department of Biochemistry, University of Melbourne, Parkville, 3052, Victoria, Australia
*Correspondence: d.bowtell@petermac.org

Large-scale genomic studies promise to advance our understanding of the biology of human cancers and to improve their diagnosis, prognostication, and treatment. The analysis and interpretation of genomics studies have faced challenges. The retrospective and observational design of many studies has rendered them susceptible to confounding and bias. Technological variations and advances have impacted on reproducibility. Statistical hurdles in relating a large number of variables to a small number of observations have added further constraints. This review considers the promise and challenge associated with the large-scale clinically oriented genomic analysis of human cancer and attempts to emphasize potential solutions.

Introduction

While microarray studies have achieved much, the immense potential of large-scale genomics research to change the management of human disease remains to be fully realized. Practical constraints are imposed by the cost of genomic studies and difficulties in obtaining sufficient, well-annotated, and representative samples, particularly for human studies. While genomic technology is continuously improving in reliability and information content, comparing and combining data from different genomic platforms remains problematic. Most genomic experiments involve thousands of variables (such as gene expression values) measured against tens or, at best, hundreds of cases. False positive results and data overfitting are significant problems under these circumstances. Despite these challenges, validated findings have been made, and the first of these have become commercially available in some countries (Paik et al., 2004). Here, we provide an overview of the many complexities that face large-scale clinically oriented cancer genomic studies, with the goal of assisting readers and researchers in understanding and anticipating obstacles. We begin by examining the role of confounding and bias in study design, discuss technology-related limitations and statistical and analytical obstacles, and finish with several clinical considerations. We also provide recommendations for circumventing problems that have beset previous studies (Table 1).

Study design

The main objectives of most large-scale cancer genomics studies are to search for new molecular subtypes of cancer (class discovery); identify differentially expressed genes between predefined cancer classes, such as short- versus long-term survivors (class comparison); or predict membership to predefined cancer classes (class prediction) (Golub et al., 1999; Simon et al., 2003). Class discovery genomic studies have succeeded in identifying several important and reproducible molecular cancer subtypes. For example, the work of Perou et al. (1999) has identified several molecular subtypes of breast cancer, confirming the long held notion that breast cancer is comprised of more than one biological entity. These subtypes, with distinct gene expression profiles and patterns of oncogene activation or tumor suppressor loss, have been validated in independent data sets and correlated to clinical outcome (Sorlie et al., 2001, 2003). Likewise, class comparison studies have generated insights into the molecular relationships between other cancer subtypes. One example is the comparison of gene expression profiles in ovarian cancers from women with inherited BRCA1 and BRCA2 mutations to those with sporadic cancer (Jazaeri et al., 2002). It appears that the BRCA-associated pathways are also involved in sporadic cases of ovarian cancer, leading to speculation about the role of genetic and epigenetic alterations in BRCA genes and downstream regulators. Our own study has been able to compare distinct histological subtypes of gastric cancer, highlighting transcriptional differences between the intestinal and diffuse histologies (Boussioutas et al., 2003). Many researchers have attempted to springboard from class comparison to class prediction studies in order to develop valid molecular profiles with potential clinical applications. One class comparison study of histological grade 1 and grade 3 breast cancers has led to the identification of a gene expression profile that can be used to further classify histological grade 2 tumors into high versus low risk of recurrence categories, although the results of this study require further validation (Sotiriou et al., 2006). Despite the successes, challenges have been identified that affect all three study designs. Practical constraints in obtaining human cancer tissue have led many genomics studies to use a limited number of retrospectively collected samples. Therefore, the cases may not have been collected in a standardized fashion, and the observations may have been made from uncontrolled systems. Under these circumstances, data confounding and bias are particularly relevant.

Confounding refers to a factor that distorts the true relationship between the study variables of interest (Potter, 2003). A confounder is related to the outcome of interest, yet remains extraneous to the study question and is unequally distributed among the comparator groups. Confounding is important in cancer molecular profiling, especially for class comparison and class prediction studies. For example, in a study designed to derive a molecular predictor of chemotherapy responders and nonresponders, confounding can occur if other therapeutic modalities (e.g., surgery and radiotherapy) are not equally distributed amongst the two groups. In this situation, attributing a difference in gene expression to the characteristics of the cancer may be

Table 1. Problems and possible solutions in the design of clinically oriented microarray studies

Category	Problem	Potential solutions
Study design issues	Bias	Prospective design
		Randomization (where possible)
		Blinding (where appropriate)
		Avoid inappropriate pooling of samples
	Confounding	Complete clinical/pathological annotation
		Stratification using known confounders
		Use of prospective study design with structured reporting of key information
Array issues	Reproducibility	Choose a single molecular platform
		Standardization of technical protocols
		Biological and technical repeats
		Make available probe sequences for future reannotation
	Cross-platform comparison	Sequence verification of probes
		Removal of misannotated probes from analysis
		Utilize up-to-date version of genome annotation
		Use rank statistics rather than absolute values of gene expression
Statistical issues	Overfitting	Internal validation (leave one out cross-validation or split-sample analysis)
		External/independent validation
	Unstable gene lists	Multiple permutation of training and test sets
	Study power	A priori calculation of sample size using available methods
		Post hoc analysis of microarray data may indicate adequacy of sample size
	Data interpretation	Ranked biological themes
		Gene set enrichment analysis (GSEA)
		In vivo modeling

inappropriate. To correct the problem, patients should be balanced for known confounders; however, if the study sample size is limited, this may be particularly difficult to do and is likely to reduce the number of cases suitable for the analysis even further. Still, common confounders such as age, gender, cancer stage, tumor histology, and the treatment delivered require correction whenever possible. Therefore, having comprehensive clinical annotation of the biological samples is highly desirable. Accurate annotation can be especially difficult to obtain for archival, or ad hoc, sample collections. Retrieval of case histories may help to complete the sample annotation; however, if the samples do not encompass the full spectrum of the disease under study, or were compiled over a period when standards in clinical management changed, then generalizability to the present may be limited (Ahmed and Brenton, 2005). Given the importance of adequate clinical annotation for interpreting genomic studies, it would seem useful to develop a set of guidelines for recording a minimum clinical data set for human tissue used in microarray experiments, similar to the Minimal Information About a Microarray Experiment (MIAME) (Brazma et al., 2001) and the Standards for Reporting Diagnostic Accuracy (Bossuyt et al., 2003a, 2003b; Novere et al., 2005). A recent publication has taken the first steps in this direction (McShane et al., 2005).

Bias refers to a systematic difference in the way that study cases are handled or analyzed. Given that bias is a function of study design, every study should be carefully considered for all possible sources of bias at the outset. Some sources of bias are already acknowledged in the literature or are relatively easily identified. For example, differences in the physical handling and processing of cases and controls can introduce bias and lead to erroneous conclusions (Coombes et al., 2005). Technical factors, such as the time required to conduct an assay, the batch of reagents used, and the skill levels of different technicians are all possible sources of bias. Pooling of tissues from multiple tissue banks to increase the sample size is a common practice; however, this may increase heterogeneity and introduce additional biases. Conversely, some sources of bias may remain concealed, or the magnitude and direction of their effect may be difficult to ascertain (Ransohoff, 2005). Avoiding heterogeneity, randomization of processing steps, development of strict inclusion and exclusion criteria, systemization of protocols, and blinding of technicians to the class assignment of the specimens being handled are all valid methods for reducing systematic study bias. An extensive review of bias associated with molecular prediction studies has recently been published by Ransohoff (2005).

The use of a prospective study design is one of the most effective methods for controlling confounding and reducing biases. Investigators can plan in advance the hypothesis to be tested and the necessary sample annotation to be collected. In addition, it allows advance consideration of the required sample size (see "False findings, power, and sample size" in the "Statistical challenges" section below). And finally, a prospective design can ensure that all samples are handled and processed in a standardized fashion to minimize experimental bias. This approach has been adopted by the European Organization for Research and Treatment of Cancer (EORTC) in designing the Microarray In Node-negative Disease may Avoid ChemoTherapy (MINDACT) study (see "Clinical utility" section). Inevitably, a prospective

observational study will require long periods of time to allow for patient enrolment and adequate clinical follow-up. The use of biologic material collected previously, such as during a prospectively designed clinical trial, can be an attractive alternative. While some variations of sample handling may have occurred in trials not specifically intended for gene expression studies, methods to demonstrate standardization and quality control can be applied (Simon, 2005). A successful example of this approach is the study of Paik et al. (2004). Fixed material from over 600 breast cancer cases collected as part of a large National Surgical Adjuvant Breast and Bowel Project randomized clinical trial was utilized to validate a 21 gene RT-PCR-based molecular profile designed to predict clinical outcome in early-stage estrogen receptor-positive breast cancer patients.

Molecular platform

At the foundation of large-scale genomics is the ability to synthesize high-density microarrays. While high-density arrays can be constructed for many different purposes (such as measuring gene copy number, or epigenetic modifications, etc.), gene expression studies represent the most common use of this technology. Many of the problems observed with gene expression profiling apply to other large-scale genomic modalities. A variety of microarray platforms are available for expression analysis (Hardiman, 2004), with differences in the platform design, synthesis (in house versus commercial), and probe annotation. As a result, cross-platform comparisons of gene expression studies have been difficult. This is exemplified by two recent studies in breast cancer (van de Vijver et al., 2002; Wang et al., 2005) in which less than half of the key genes found to be predictive of distant metastases in women with lymph node-negative breast cancer were present on both sets of microarrays.

Most direct comparisons of different genomics platforms have found considerable variation in gene expression results (Bammler et al., 2005; Jarvinen et al., 2004; Kuo et al., 2002; Tan et al., 2003), although some exceptions exist (Irizarry et al., 2005; Larkin et al., 2005). Several factors contribute to the variation, including differences in the methods of RNA labeling, the process of hybridization, data acquisition, data preprocessing and normalization (Bammler et al., 2005), errors in probe placement and probe annotation (Carter et al., 2005; Jarvinen et al., 2004; Mecham et al., 2004a, 2004b), and differences attributed directly to the platform design (such as the different hybridization properties of cDNA arrays, and long and short oligonucleotide arrays) (Jarvinen et al., 2004; Kuo et al., 2002; Tan et al., 2003). Also, significant interlaboratory variation has been observed even when identical microarray platforms and starting nuclear materials are used (Irizarry et al., 2005). Concordance is highest when protocols are standardized and the same molecular platform is used, with the best results obtained from commercial platforms (Bammler et al., 2005; Dobbin et al., 2005). Comparison of data generated on different platforms is facilitated by sequence testing of probes, removal of misannotated genes from comparative analyses (Carter et al., 2005; Mecham et al., 2004a), using relative rather than absolute gene expression measurements (Irizarry et al., 2005; Jarvinen et al., 2004; Larkin et al., 2005), and applying improved preprocessing algorithms (Gentleman et al., 2004; Yang et al., 2002). In addition, comparisons at the level of biological processes, rather than single genes, are more informative and reproducible (Bammler et al., 2005). It is anticipated that technical variation will decline (Larkin et al., 2005), especially as guidelines and standard operating procedures for the execution of genomic experiments are further developed. Many journals mandate the use of the MIAME standard (Brazma et al., 2001) in the reporting of microarray experiments. Additional standards to improve reproducibility, sensitivity, and robustness in gene expression analysis are being proposed (see http://www.cstl.nist.gov/biotech/workshops/ERCC2003/).

The annotation of the human genome is continuously evolving, and therefore earlier annotations of technical platforms are slowly becoming obsolete, adding an additional challenge to the cross-platform comparison (Dai et al., 2005). A method for updating the annotation of probe sets within a platform, based on sequence alignments and specified probe selection, has been proposed by Dai et al. (2005). The authors suggest that updating probe set annotation should provide more accurate interpretation of gene expression data.

Finally, the number of transcripts in the mammalian genome is at least one order of magnitude greater than previously estimated because of the presence of alternative splicing and transcription from both strands of DNA (Bertone et al., 2004; Cheng et al., 2005; The FANTOM Consortium et al., 2005). A comprehensive compilation of splice variants has not yet been created (Marshall, 2004), and alternative transcripts have not previously been considered in the design of gene expression platforms. It is possible, then, that previous attempts at cross-platform comparisons have been hampered by a failure to adequately account for alternative transcription. Recognizing this, concerted efforts to incorporate the entire transcriptome onto a single microarray platform are being made.

Statistical challenges

In a typical microarray experiment, hundreds, if not thousands, of genes are individually subjected to a statistical test of significance. The process of multiple hypothesis testing leads to false positive findings, a phenomenon that needs to be controlled without markedly compromising study power. Other statistical challenges, such as determining sample size, data overfitting, and unstable gene lists also require consideration.

False findings, power, and sample size

The false positive rate of a test is associated with the standard p value. Invented for testing individual hypotheses, the p value requires adjustment for multiple testing in order to avoid abundant false positive results. Several approaches, varying in the stringency of the correction for multiplicity, are available (reviewed in Pawitan et al., 2005). Adjusting the p value has the potential to be overly conservative, resulting in loss of power i.e., low sensitivity, or high false negative rate. Therefore, the need for a less conservative approach for controlling multiple testing and retaining study power led to the derivation of the false discovery rate (FDR) (Benjamini and Hochberg, 1995; reviewed in Reiner et al., 2003). The FDR is the expected proportion of false discoveries among a declared significant result, and several methods of FDR multiple testing correction exist (Li et al., 2005).

Four characteristics of microarray experiments determine the FDR: (1) the proportion of truly differentially expressed genes, (2) the distribution of the true differences, (3) variability, and (4) the study sample size (Pawitan et al., 2005). Only sample size is under the control of the researcher, and it is of key importance in the planning of a microarray experiment. Sample size also directly affects the power of a study, that is, the probability of correctly identifying an effect of the desired (or greater) magnitude.

Optimally, the number of samples required to meet the study goals and achieve the defined statistical requirements would be determined at the study outset, helping to minimize resource waste, and creating a level of certainly about the results.

Several approaches for calculating the sample size of a microarray experiment have been described and tailored to specific study designs (e.g., class comparison or class prediction, etc.) (Black and Doerge, 2002; Dobbin and Simon, 2005; Hwang et al., 2002; Pawitan et al., 2005; Wei et al., 2004; Yang et al., 2003). Sample size calculations include several components: (1) the variance of individual measurements (associated with biological and experimental variation), (2) the magnitude of the effect to be detected, and (3) the stringency of the FDR. In microarray experiments, the first two components are difficult to predict. Estimates of variance can be obtained from prior studies, but careful case selection is required to ensure generalizability. For small studies, the estimated variance may be particularly unreliable, and the usual assumption of normal distribution is threatened (Dobbin and Simon, 2005). Iterative procedures to overcome limited sample size or statistical methods for inferring variance across genes may help to improve the estimation of variance (Wright and Simon, 2003). Larger sample sizes are required if variance is large; therefore, conservative sample size predictions should be based on estimates of variance derived from the most highly variable genes (Dobbin and Simon, 2005; Yang and Speed, 2002). The magnitude of effect can also be a challenge. It may be unknown what minimum change in gene expression is required for significant biological effects. There is evidence from a class comparison study that subtle, coordinated changes in expression levels can be biologically important (Mootha et al., 2003). If a study size was determined a priori using sound statistical methods, failure to derive a differential gene list using standard analytical approaches can suggest that alternative methods, such as gene set enrichment analysis (GSEA), may be appropriate (see "Interpreting gene lists" section).

The FDR, and hence sample size, are particularly sensitive to the proportion of genes that are truly differentially expressed (Dobbin and Simon, 2005). Experimenters may already have an idea about the number of differentially expressed genes from previous studies. However, if this information is not available, a highly conservative assumption can be made, or a pilot study (also useful to determine other variance components) can be conducted (Yang et al., 2003).

While the methodological details of sample size calculations are beyond the scope of this review, several approaches have been published, and in some cases software has been made publicly available (Pawitan et al., 2005). Researchers are encouraged to use the existing methods for sample size determination in experimental design. By incorporating sample size calculation into microarray experiments, the existing methods can be refined and improved, increasing further our ability to reliably design and analyze large-scale genome studies.

Data overfitting

When thousands of gene transcripts are measured in the development of a molecular classifier, there is a substantial chance that random associations between genes and the predefined classes of interest can occur (Simon et al., 2003). Such data "overfitting" is by definition random and nonreproducible (Lahad et al., 2005; Ransohoff, 2004; Simon et al., 2003). A good example of data overfitting is provided by Simon et al. (2003), who invented imaginary cases, ten with and ten without cancer, and a data set containing 6000 genes. The researchers were able to use standard research methods to generate models, at a high frequency, that perfectly fit the training set. Given the entirely arbitrary nature of the data, this finding underscores the potential impact of data overfitting.

Data overfitting can be reduced if the training set is subjected to a rigorous internal cross-validation (Simon et al., 2003). Cross-validation tests the model building process by removing one (e.g., leave one out cross-validation) or more samples out of the training set and scoring the ability of the prediction model, derived at each iteration, to classify the left-out samples. After the entire training set has been examined, a combined model is generated and the cross-validated model is reported with an estimate of the prediction error. The error rate can be significantly underestimated if investigators fail to account for all specimens (such as those that remained unclassified) or if cross-validation is improperly conducted (Simon et al., 2003). It is imperative that all aspects of the predictive model be subjected to the cross-validation procedure, meaning that for each iteration, the selection of informative genes, the computation of the gene weights, and the creation of the prediction rule should be repeated.

Once proper internal validation has been completed, the predictive model can be subjected to external validation using an independent test set. This is key to determining whether data overfitting has occurred (Ransohoff, 2004). Independent validation can involve a split-sample methodology, where some samples are used to generate the model (training) and others for testing the model (validation) (Ransohoff, 2004), or where the entire initial data set is used for building a classifier, and then this is tested on an independent data set. The latter method requires that the validation set comprises cases similar to those for which the classifier was designed (Simon, 2005; Simon et al., 2003). For example, a two-gene prognostic classifier for the prediction of disease recurrence in women with breast cancer (Ma et al., 2004) could not be validated in an independent data set (Reid et al., 2005), likely due to mismatched patient characteristics (Simon, 2005). Whenever a test and training set design is used, it is essential that complete separation of the samples comprising the sets is achieved to avoid overestimation of the prediction accuracy (Ntzani and Ioannidis, 2003; Simon et al., 2003). An early genomic study for the prediction of breast cancer recurrence (van de Vijver et al., 2002) is notable for incompletely separating the test and training sets (Ransohoff, 2004). Finally, it is important to note that small test and validation sets are unlikely to represent the heterogeneity of the condition under study, and the performance of the classifier can be markedly overestimated (Ransohoff, 2004; Simon et al., 2003).

Unstable gene lists

In a classic gene expression study, where thousands of genes are being tested against a relatively small set of cancer classes, it is possible that multiple and interchangeable gene combinations may simultaneously correlate to the classes of interest. Several class prediction studies have derived minimally overlapping, equally predictive, and valid gene lists (van 't Veer et al., 2002; Wang et al., 2005). While factors such as discrepant genomic platforms, confounding, and bias may contribute to such differences, other factors associated with sample composition may also play a role. Gene lists derived from a single data set have been found to be highly dependent on the composition of the training and test sets, displaying "instability" when the training and test sets are iterated (Ein-Dor et al., 2005; Michiels et al., 2005). The coinci-

dental grouping of several samples with a skewed correlation of gene expression (i.e., not representative of the entire population from which they were drawn) into the training set may contribute to this phenomenon. While, for class prediction, several gene lists may be operationally interchangeable, unstable gene lists may pose some challenges for studies aimed at gaining insight into the underlying biology of different cancer classes. One method for reducing gene list instability involves multiple training/test set partitions to find the most stable gene list (Michiels et al., 2005). When an inference about biological processes is desired, it may be more appropriate to interrogate the list for biological themes present in the data, rather than for individual genes (see "Interpreting gene lists").

Interpreting gene lists

A major goal of any microarray study is to identify biologically or clinically significant changes in gene expression. However, some analytical approaches may ignore genes that do not pass the threshold for differential expression (typically defined arbitrarily). Also, some biological processes exhibit only subtle, coordinated changes in the expression of single genes or groups of genes, such that they would not be identified through common analytical methods.

Methods for identifying and ranking biological themes from large-scale genomic data have been developed (e.g., DAVID, CLENCH, eGOn, GOstat, Onto-Miner, Avadis, etc.) (listing available at http://www.geneontology.org/GO.tools.shtml). Such post hoc analyses help to identify statistically significant occurrences of biologically relevant themes or phrases. While accessible and user friendly, all such software contain biases related to the use of only the top ranking genes to identify pathway membership and from dependence on published abstracts, such that publication bias may skew the results.

Searching for significant differences in expression of an a priori defined gene set, generated by incorporating biological knowledge, may facilitate the detection of modest but coordinate changes in sets of functionally related genes. One example of this approach is the GSEA, which was initially developed for the comparative study of muscle from patients with normal glucose tolerance, impaired glucose tolerance, and type II diabetes mellitus in whom differentially expressed genes could not be obtained using classic approaches (Mootha et al., 2003; Subramanian et al., 2005). When 149 a priori defined gene sets were tested using the GSEA approach, a group of coordinately downregulated genes was identified. The results were felt to be biologically plausible and have since been independently validated (Patti et al., 2003; Petersen et al., 2004). While it is possible that GSEA may be biased toward gene sets of larger size (Damian and Gorfine, 2004), it represents an important advancement in identifying biological processes from microarray data.

Several groups have successfully examined biological themes by integrating multiple independent data sets (Ramaswamy et al., 2003; Rhodes et al., 2004; Segal et al., 2004). Rather than focusing on ranked gene lists, computational techniques have been applied to identify common themes (or unique molecular "modules") underlying human malignancies (Segal et al., 2004, 2005; Segal et al., 2003). In addition, controlled in vitro and in vivo model systems with clearly defined molecular derangements have been used to deconvolute and define the molecular signatures of several key activated oncogenes (Huang et al., 2003; Lamb et al., 2003; Sweet-Cordero et al., 2005). An important observation is that oncogene signatures can be detected within human cancers and associated with disease outcomes (Bild et al., 2005; Glinsky et al., 2004, 2005). It is thought that these types of approaches will increase the likelihood of understanding the signals in microarray data and will provide results that are more interpretable than traditional gene lists (reviewed in Segal et al., 2005).

Clinical utility

The true test of a predictive profile, and its ultimate acceptance as a prognostic tool, requires a demonstration of its utility in the clinical setting. For many cancers, prognostic models based on clinicopathological factors already exist, and a genomic classifier should demonstrate added benefit beyond the best prognostic models already in use today. An example of this approach is the work of Rosenwald et al. (2002) in patients with diffuse large-B cell lymphoma who were treated with standard chemotherapy and whose prognosis was determined using the International Prognosis Index (IPI). When their molecular profile was incorporated into the IPI, the survival curves for the patients were further split, suggesting that the molecular profile further improved predictive ability. While such evidence is compelling, the ideal approach would be to demonstrate the added utility of a molecular predictor in the setting of a prospectively conducted randomized, controlled trial. The MINDACT study in breast cancer patients is a randomized clinical trial that incorporates the 70-gene predictive profile of Van't Veer et al. (van de Vijver et al., 2002; van 't Veer et al., 2002) with standard clinical pathological criteria. The study will randomize women with discordant molecular and clinicopathological risk to either receive chemotherapy or not. This design will help to determine whether patients with a low-risk molecular prognosis but a high-risk clinical prognosis can be safely spared adjuvant chemotherapy, thereby avoiding overtreatment of a potentially low-risk group of patients.

Conclusion

The identified challenges of large-scale cancer gene expression microarrays can, for the most part, be anticipated and managed. A multidisciplinary approach to the design of large-scale genomic studies, incorporating the expertise of clinicians, scientists, bioinformaticians, clinical epidemiologists, and statisticians will help to address many of the study design, technical, statistical, and analysis hurdles outlined in this review. The immense power of genomic research is yet to be fully uncovered. By optimizing the design and use of gene expression microarrays, and related technologies, the potential to change the way in which we understand and manage human cancers, and many other human diseases, may be realized.

Acknowledgments

We would like to thank Drs. Terry Speed, Izahk Haviv, Wayne Phillips, David Thomas, and Richard Tothill for critical appraisal of the manuscript. A.V.T. has been supported by a Research Fellowship from the Canadian Institutes of Health Research, by Eli Lilly Canada, and in part by Pfizer Canada Inc. in partnership with the Canadian Association of Medical Oncologists.

References

Ahmed, A., and Brenton, J. (2005). Microarrays and breast cancer clinical studies: forgetting what we have not yet learnt. Breast Cancer Res. *7*, 96–99.

Bammler, T., Beyer, R.P., Bhattacharya, S., Boorman, G.A., Boyles, A.,

Bradford, B.U., Bumgarner, R.E., Bushel, P.R., Chaturvedi, K., Choi, D., et al. (2005). Standardizing global gene expression analysis between laboratories and across platforms. Nat. Methods *2*, 351–356.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. [Ser A] *57*, 289–300.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A., et al. (2005). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature *439*, 353–357.

Black, M.A., and Doerge, R.W. (2002). Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Bioinformatics *18*, 1609–1616.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., and de Vet, H.C. (2003a). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann. Intern. Med. *138*, 40–44.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Moher, D., Rennie, D., de Vet, H.C., and Lijmer, J.G. (2003b). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann. Intern. Med. *138*, W1–12.

Boussioutas, A., Li, H., Liu, J., Waring, P., Lade, S., Holloway, A.J., Taupin, D., Gorringe, K., Haviv, I., Desmond, P.V., and Bowtell, D.D. (2003). Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. Cancer Res. *63*, 2569–2577.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. *29*, 365–371.

Carter, S., Eklund, A., Mecham, B., Kohane, I., and Szallasi, Z. (2005). Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. BMC Bioinformatics *6*, 107.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149–1154.

Coombes, K.R., Morris, J.S., Hu, J., Edmonson, S.R., and Baggerly, K.A. (2005). Serum proteomics profiling—a young technology begins to mature. Nat. Biotechnol. *23*, 291–292.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. *33*, e175. 10.1093/nar/gni179.

Damian, D., and Gorfine, M. (2004). Statistical concerns about the GSEA procedure. Nat. Genet. *36*, 663.

Dobbin, K., and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics *6*, 27–38.

Dobbin, K.K., Beer, D.G., Meyerson, M., Yeatman, T.J., Gerald, W.L., Jacobson, J.W., Conley, B., Buetow, K.H., Heiskanen, M., Simon, R.M., et al. (2005). Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin. Cancer Res. *11*, 565–572.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? Bioinformatics *21*, 171–178.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Glinsky, G.V., Glinskii, A.B., Stephenson, A.J., Hoffman, R.M., and Gerald, W.L. (2004). Gene expression profiling predicts clinical outcome of prostate cancer. J. Clin. Invest. *113*, 913–923.

Glinsky, G.V., Berezovska, O., and Glinskii, A.B. (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J. Clin. Invest. *115*, 1503–1521.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–537.

Hardiman, G. (2004). Microarray platforms—comparisons and contrasts. Pharmacogenomics *5*, 487–502.

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., D'Amico, M., Pestell, R.G., West, M., and Nevins, J.R. (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat. Genet. *34*, 226–230.

Hwang, D., Schmitt, W.A., and Stephanopoulos, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. Bioinformatics *18*, 1184–1193.

Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G.N., Geoghegan, J., Germino, G., et al. (2005). Multiple-laboratory comparison of microarray platforms. Nat. Methods *2*, 345–350.

Jarvinen, A.K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O.P., and Monni, O. (2004). Are data from different gene expression microarray platforms comparable? Genomics *83*, 1164–1168.

Jazaeri, A.A., Yee, C.J., Sotiriou, C., Brantley, K.R., Boyd, J., and Liu, E.T. (2002). Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. J. Natl. Cancer Inst. *94*, 990–1000.

Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L., and Kohane, I.S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics *18*, 405–412.

Lahad, J.P., Mills, G.B., and Coombes, K.R. (2005). Stem cell-ness: a "magic marker" for cancer. J. Clin. Invest. *115*, 1463–1467.

Lamb, J., Ramaswamy, S., Ford, H.L., Contreras, B., Martinez, R.V., Kittrell, F.S., Zahnow, C.A., Patterson, N., Golub, T.R., and Ewen, M.E. (2003). A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell *114*, 323–334.

Larkin, J.E., Frank, B.C., Gavras, H., Sultana, R., and Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. Nat. Methods *2*, 337–344.

Li, S.S., Bigler, J., Lampe, J.W., Potter, J.D., and Feng, Z. (2005). FDR-controlling testing procedures and sample size determination for microarrays. Stat. Med. *24*, 2267–2280.

Ma, X.-J., Wang, Z., Ryan, P.D., Isakoff, S.J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., and Tuggle, J.T. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell *5*, 607–616.

Marshall, E. (2004). Getting the noise out of gene arrays. Science *306*, 630–631.

McShane, L.M., Altman, D.G., Sauerbrei, W., Taube, S.E., Gion, M., and Clark, G.M. (2005). REporting recommendations for tumour MARKer prognostic studies (REMARK). Br. J. Cancer *93*, 387–391.

Mecham, B.H., Klus, G.T., Strovel, J., Augustus, M., Byrne, D., Bozso, P., Wetmore, D.Z., Mariani, T.J., Kohane, I.S., and Szallasi, Z. (2004a). Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Res. *32*, e74.

Mecham, B.H., Wetmore, D.Z., Szallasi, Z., Sadovsky, Y., Kohane, I., and Mariani, T.J. (2004b). Increased measurement accuracy for sequence-verified microarray probes. Physiol. Genomics *18*, 308–315.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet *365*, 488–492.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are

coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Novere, N.L., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). Nat. Biotechnol. *23*, 1509–1515.

Ntzani, E.E., and Ioannidis, J.P. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet *362*, 1439–1444.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N. Engl. J. Med. *351*, 2817–2826.

Patti, M.E., Butte, A.J., Crunkhorn, S., Cusi, K., Berria, R., Kashyap, S., Miyazaki, Y., Kohane, I., Costello, M., Saccone, R., et al. (2003). Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. Proc. Natl. Acad. Sci. USA *100*, 8466–8471.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics *21*, 3017–3024.

Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA *96*, 9212–9217.

Petersen, K.F., Dufour, S., Befroy, D., Garcia, R., and Shulman, G.I. (2004). Impaired mitochondrial activity in the insulin-resistant offspring of patients with type 2 diabetes. N. Engl. J. Med. *350*, 664–671.

Potter, J.D. (2003). Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs. Trends Genet. *19*, 690–695.

Ramaswamy, S., Ross, K.N., Lander, E.S., and Golub, T.R. (2003). A molecular signature of metastasis in primary solid tumors. Nat. Genet. *33*, 49–54.

Ransohoff, D.F. (2004). Rules of evidence for cancer molecular-marker discovery and validation. Nat. Rev. Cancer *4*, 309–314.

Ransohoff, D.F. (2005). Bias as a threat to the validity of cancer molecular-marker research. Nat. Rev. Cancer *5*, 142–149.

Reid, J.F., Lusa, L., De Cecco, L., Coradini, D., Veneroni, S., Daidone, M.G., Gariboldi, M., and Pierotti, M.A. (2005). Limits of predictive models using microarray data for breast cancer clinical treatment outcome. J. Natl. Cancer Inst. *97*, 927–930.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics *19*, 368–375.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A.M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc. Natl. Acad. Sci. USA *101*, 9309–9314.

Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N. Engl. J. Med. *346*, 1937–1947.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nat. Genet. *36*, 1090–1098.

Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005). From signatures to models: understanding cancer using microarrays. Nat Genet. *37* (*Suppl*), S38–S45.

Simon, R. (2005). Development and validation of therapeutically relevant multigene biomarker classifiers. J. Natl. Cancer Inst. *97*, 866–867.

Simon, R., Radmacher, M.D., Dobbin, K., and McShane, L.M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J. Natl. Cancer Inst. *95*, 14–18.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*, 10869–10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. USA *100*, 8418–8423.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. *98*, 262–272.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J.J., Ladd-Acosta, C., Mesirov, J., Golub, T.R., and Jacks, T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat. Genet. *37*, 48–55.

Tan, P.K., Downey, T.J., Spitznagel, E.L., Jr., Xu, P., Fu, D., Dimitrov, D.S., Lempicki, R.A., Raaka, B.M., and Cam, M.C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. *31*, 5676–5684.

The FANTOM Consortium, Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. *347*, 1999–2009.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet *365*, 671–679.

Wei, C., Li, J., and Bumgarner, R. (2004). Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics *5*, 87.

Wright, G.W., and Simon, R.M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics *19*, 2448–2455.

Yang, Y.H., and Speed, T. (2002). Design issues for cDNA microarray experiments. Nat. Rev. Genet. *3*, 579–588.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. *30*, e15.

Yang, M.C.K., Yang, J.J., McIndoe, R.A., and She, J.X. (2003). Microarray experimental design: power and sample size considerations. Physiol. Genomics *16*, 24–28.

DOI: 10.1016/j.ccr.2006.05.001