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