1

Variation, Variability, Batches and Bias in Microarray Experiments: An Introduction

Andreas Scherer

Abstract

Microarray-based measurement of gene expression levels is a widely used technology in biological and medical research. The discussion around the impact of variability on the reproducibility of microarray data has captured the imagination of researchers ever since the invention of microarray technology in the mid 1990s. Variability has many sources of the most diverse kinds, and depending on the experimental performance it can manifest itself as a random factor or as a systematic factor, termed bias. Knowledge of the biological/medical as well as the practical background of a planned microarray experiment helps alleviate the impact of systematic sources of variability, but can hardly address random effects.

The invention of microarray technology in the mid 1990s allowed the simultaneous monitoring of the expression levels of thousands of genes (Brown and Botstein 1999; Lockhart *et al.* 1996; Schena *et al.* 1995). Microarray-based high density/high content gene expression technology is nowadays commonly used in fundamental biological and medical research to generate testable hypotheses on physiological processes and disease. It is designed to measure variation of expression due to biological, physiological, genetic and/or environmental conditions, and it allows us to study differences in gene expression induced by factors of interest, such as pharmacological and toxicological effects of compounds, environmental effects, growth and aging, and disease phenotypes. We note that the term 'variation' describes directly measurable differences among individuals or samples, while the term 'variability' refers to the potential to vary.

Batch Effects and Noise in Microarray Experiments: Sources and Solutions edited by A. Scherer

As we shall see in more detail in Chapter 2, relative quantification of gene expression involves many steps including sample handling, messenger RNA (mRNA) extraction, in-vitro reverse transcription, labeling of complementary RNA (cRNA) with fluorescent dyes, hybridization of the labeled cDNA (target) to oligonucleotides with complementary sequences (probes), which are immobilized on solid surfaces, and the measurement of the intensity of the fluorescent signal which is emitted by the labeled target. The measured signal intensity per target is a measure of relative abundance of the particular mRNA species in the original biological sample.

Unfortunately, microarray technology has its caveats, as it is susceptible to variability like any other measurement process. As we will discuss in Chapters 2 and 3, technical variation manifests itself in signal intensity variability. This effect is informally called 'noise': technical components which are not part of the system under investigation but which, if they enter the system, lead to variability in the experimental outcomes. Note that noise is only defined in the context of technology. Since the early years of microarrays, noise and its impact on the reliability of large-scale genomics data analysis have been a much discussed topic. The team of Kerr *et al.* (2000b) was among the first to recognize the problem and to propose ANOVA methods to estimate noise in microarray data sets. Tu *et al.* (2002) addressed the issue of how to measure the impact of different sources of noise. Using a set of replicate arrays with varying degrees of preparation differences, they were able to characterize quantitatively that the hybridization noise is very high compared to sample preparation or amplification. They also found that the level of noise is signal intensity dependent, and propose a method for significance testing based on noise characteristics.

The unresolved issue of measurement variability and measuring variability has hampered the great hopes researchers had with the advent of microarray technology and the human genome sequence project. Since consensus technological, analytical, and reporting processes were (and still are) largely missing, it appeared that not only were gene expression data irreproducible, but also the results were very much dependent on the choice of analytical methods. A lively discussion on the validity of microarray technology resulted in publications and comments like 'Microarrays and molecular research: noise discovery?' (Ioannidis 2005), 'An array of problems' (Frantz 2005), countered by 'Arrays of hope' (Strauss 2006), and 'In praise of arrays' (Ying and Sarwal 2008), and publications which raise questions about the reproducibility of microarray data (Marshall 2004; Ein-Dor *et al.* 2006) or showing increased reproducibility (Dobbin *et al.* 2005b; Irizarry *et al.* 2005; Larkin *et al.* 2005).

Shi *et al.* addressed this issue in a systematic manner and in 2006 published a comparative analysis of a large data set which had been generated by MicroArray Quality Control Consortium (MAQC) with 137 participants from 51 organizations (Shi *et al.* 2006). The data set consists of two commercially available RNA samples of high quality – Universal Human Reference RNA (UHRR) and Human Brain Reference RNA – which were mixed in four titration pools, and whose mRNA levels were measured on seven microarray platforms in addition to three alternative platforms. Each array platform was deployed at three test sites, and from each sample five replicates were assayed at each site. This information-rich data set is an excellent source for the investigation of technological noise, and some of its data will be used in a number of chapters in this book. The project showed that quantitative measures across all one-color array platforms had a median coefficient of variation (CV) of 5–15%, and a concordance rate for the qualitative calls ('present', 'absent') of 80–95% between sample replicates. Lists of differentially expressed genes overlapped by about 89% between the test sites using the same platform, dropping to 74% overlap across platforms. The important conclusion the authors made is that the performance of the microarray technology in their study speaks for its use in basic research and may lead to its use as clinical diagnostic tool as well. The authors further point out that standardization of data reporting, analysis tools and controls is important in this process.

As pointed out earlier, 'noise' is used to informally describe measurement variability due to technical factors. In the context of biological variability, the term 'noise' will be avoided in the course of this book. Here we suggest the use of the more neutral term 'expression heterogeneity'. The basis of expression heterogeneity lies within the inherent differences in the nature of the subjects or specimen which are studied. It is dependent on the subjects' physiological states, their gender, age, and genetic aspects (Brem *et al.* 2002; DeRisi *et al.* 1997; Rodwell *et al.* 2004). Variability due to biological factors cannot be avoided or minimized and may sometimes even be useful and important. To minimize technical and biological variability, animal toxico- and pharmacogenomic studies are performed under standardized conditions until the tissue harvest (and further): housing in standardized cages, gender- and age-matching, and technical processes which adhere to standard operating procedures and good laboratory practice. However, one or more animals may react differently to the treatment than the others, and their expression signature may indeed provide very valuable information for the investigators. In another example, measuring gene expression heterogeneity is important in gaining understanding pathogenesis in the concept of personalized medicine (Anguiano *et al.* 2008; Bottinger *et al.* 2002; Heidecker and Hare 2007; Lee and Macgregor 2004).

In contrast to the random nature of 'noise', the nature of 'batch effect' is systematic. The term 'batch effect' refers exclusively to systematic technical differences when samples are processed and measured in different batches. Lamb *et al.* were confronted with batch effects when they tested 164 small molecules in cell culture. Since not all cells could be grown at the same time due to the large amount of cells they needed, the cells had to be grown in batches. Hierarchical clustering showed that this batch effect masked the more subtle effects of small-molecule treatment (Lamb *et al.* 2006). As we shall see in other examples and sources of batch effects in the course of this book, batch effects can virtually be generated at each step of an experiment, from sample manipulation to data acquisition. They are unrelated to the biological, primary modeled factors. Batch effects introduce system variability which can be of confounding nature and mask the outcome. Proper evaluation of sources and potential magnitude of technical noise during the planning, execution and analytical phase helps in extracting relevant biological information.

A wider term describing not only technical but also other aspects of confounding the data is 'bias'. We speak of bias where one or more systematic factors affect one or more experimental groups but not all. Bias may be defined as unintentional, systematic erroneous association of some characteristic with a group in a way that distorts a comparison with another group (Ransohoff 2005). There are different types of bias, among them the following: selection bias, when, for instance, the control population has less potential for exposure than the cases; self-selection bias, when only a certain, not-representative subpopulation serves as voluntary study population; measurement bias, due to systematic differences in the measurement process; and cognitive bias, where the decision is based on educational history.

Manageable potential sources of batch effects and bias should be accounted for during the experimental design phase. They should be as consistent as possible throughout the experiment. Monitoring these sources and reporting of deviations from the standard is detrimental. In Chapters 3 (by P. Grass) and 4 (by N. Altman) it will be shown that thoughtful experimental design can alleviate the impact of batch effects. Randomization and blocking are two concepts which accomplish this. Randomization is a concept in which experimental units (e.g. samples) are assigned to groups on a strictly random basis. This means that every sample has the same chance of being selected, and that the sample is representative of its study group. Blocking is a strategy of grouping samples into experimental units which are then homogeneous for the factors which are studied. This is important when samples cannot be processed on a single day. As in the case of Lamb *et al.*, growing all cells destined to be control cells on one day and growing all treated cells on another would introduce a confounding time effect. Lamb *et al.* (2006) carefully chose a setting where treated cells were grown on the same plate as the corresponding control cells.

Chapters 5 through 15 deal with descriptive and analytical ways of exploring the nature, extent, and influence of batch effects, in addition to providing statistical means of adjusting confounding effects. As the MAQC project has stressed, standardization of data acquisition, analysis and reporting is an important factor in making gene expression studies transparent and comparable. This is further highlighted by Frueh (2006), who points out the necessity of a ‘best microarray practices’ strategy to ensure quality of starting material, data, analysis, and reporting, and interpretation. In this book Shahzad *et al.* (Chapter 17) will show the benefit of the application of standard operating procedures in the development of a commercial genomics biomarker panel. The book will close with an overview of the status of various initiatives which are currently developing standardized procedures for biomedical research (Chapter 18, Rustici *et al.*).

The purpose of the book is to raise the awareness of sources of variability in microarray data, especially of batch effects and bias. It should serve as guidance and starting point for further studies at the same time. Biologists and managers who plan microarray studies are invited to read the book, as well as laboratory personnel, statisticians, and clinicians who execute the study and analyse the data.