5

Aspects of Technical Bias

Martin Schumacher, Frank Staedtler, Wendell D Jones,
and Andreas Scherer

Abstract

Variation in microarray data can result from technical and biological sources. While the extent to which technical factors contribute to this variation has been largely investigated (Bakay *et al.* 2002; Boedigheimer *et al.* 2008; Eklund and Szallasi 2008; Fare *et al.* 2003; Han *et al.* 2004; Lusa *et al.* 2007; Novak *et al.* 2002; Zakharkin *et al.* 2005), the nature and extent of the signal intensity changes with which variation manifests itself in the data has not been a major focus of research. Using several real microarray data sets with known batch effects, we analyze and describe how technical variation is translated into gene expression changes.

5.1 Introduction

When working with data from microarray experiments in which it is not possible to process all the samples (RNA extraction, labeling, hybridization, scanning, etc.) together simultaneously under identical conditions (i.e. in one batch), so-called batch effects are very often observed. Batch effects are defined as systematic differences in the gene expression intensities in samples processed in different batches which are exclusively introduced by technical factors (Zhang *et al.* 2004). The size of batch effects can be larger than the effects of biological variables which are investigated in such an experiment. Consequently, real biological effects can be distorted or even go undetected under these circumstances. The risk of false negative and/or false positive findings is clearly enhanced when batch effects are present in the data. Our focus is on large-scale microarray experiments with many samples which cannot be processed in only one batch and/or samples which were taken at times with large intervals, as is often the case in clinical trials. In such a context subsets of the samples will often be processed at different points in time (chronologically) to address specific questions before the whole experiment is completed, thus introducing

Batch Effects and Noise in Microarray Experiments: Sources and Solutions Edited by A. Scherer
 ISBN: 978-0-470-74138-2

batch effects by definition. Another scenario which generates batch effects is the reprocessing of samples which need to be repeated for some reason. These repeated samples are often distinctly different from the bulk of other samples processed together under identical conditions. Unfortunately, none of the currently known normalization methods (Stafford 2008) efficiently eliminates these technical biases. Careful experimental design and the subsequent application of statistical methods (e.g. analysis of variance (ANOVA)) for the modeling or elimination of batch effects can effectively handle the problem (Kerr *et al.* 2000b; see also various chapters in the present volume). In this contribution we are not concerned with the detection or mathematical handling of batch effects. These aspects are covered in other chapters of this book. Our aim is to characterize the nature and the extent (size) of batch effects. We will attempt to assess, *inter alia*, whether batch effects depend on the nucleotide sequence of probes, expression intensity, their characteristics with regards to their direction, which proportion of probe sets is affected, and whether batch effects in experiments with multiple batches are constant or variable.

5.2 Observational Studies

5.2.1 Same Protocol, Different Times of Processing

Technical bias can occur at various sample processing steps. An illustration of the deviations that one can see when one slightly changes a protocol or processes RNA from the same source at different times, in this case the Universal Human Reference RNA (UHRR), is provided in Figure 5.1. Hybridizations of UHRR (Stratagene, Inc.) to a human

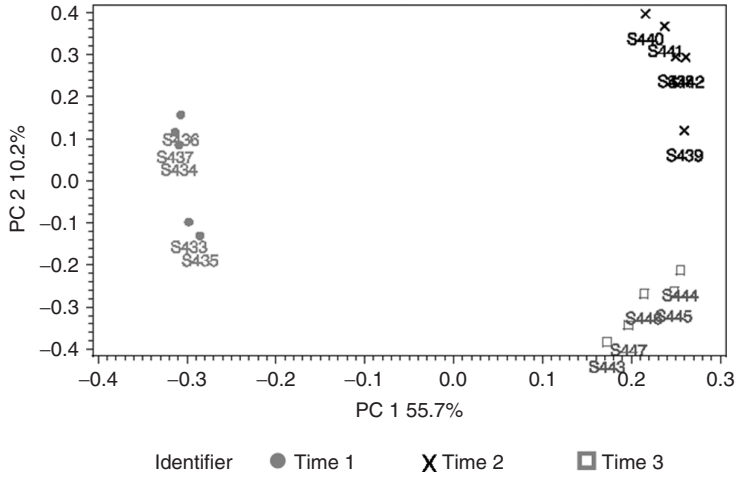

Figure 5.1 Influence of protocol changes and shift over time on gene expression data. Principal component analysis (PCA) of gene expression data from five replicate UHHR specimens processed in different batches. The same RNA samples were processed by operator 1 (time 1, gray dots) and operator 2, using the same protocol but different heating blocks. Operator 2 processed the same RNA at two different time points, separated by 1 month (time 2, crosses, and time 3, gray empty squares). Both choice of protocol and processing time point influence gene expression data, though not to the same degree.

whole-genome custom array (more than 50 000 probes) from time 1 differ from those at time 2 and time 3 in several important ways. First, two operators were involved: one who processed at time 1, another one who processed at times 2 and 3. While carrying out root-cause analysis of the differences, it was discovered that the two operators used heat blocks from two different manufacturers during their processing! This was the only significant deviation found in the protocol. Operator 2 processed at times 2 and 3, separated by one month. However, even running the same protocol with the same operator, though separated by one month, we see that time 2 and time 3 samples are still separable in the second principal component. Intensities within the time 2 and time 3 batches generally have very low coefficients of variation (median CV $\leq$5%, robust multi-array average (RMA) as summarization/normalization method). This allows for the detection of small biases. However, the biological effects in many experiments are usually larger than either variability due to repetition within time or between times.

5.2.2 Same Protocol, Different Sites (Study 1)

We designed an experiment in which we focus on the technical sources of variability following the RNA extraction. As depicted in Figure 5.2(a), five male rats (*Rattus norvegicus*) were treated with a PPARα antagonist orally once per day for two weeks, and five matched vehicle-treated male rats of the same age served as controls. After sacrifice, liver samples were snap-frozen and put into RNA-later (Ambion, Inc.) until further use. Animal housing and maintenance, treatment, liver dissection and RNA extraction were all performed in one facility, each step by a single, highly qualified technician. Total RNA was obtained by standard procedures as described elsewhere (Chomczynski and Sacchi 1987). An aliquot of 60 µg total RNA was sent to and processed by 12 technicians (operators). Explicitly, each operator followed the same experimental standard operating procedures. Labeling was done with a starting amount of 5 µg, using the Affymetrix labeling kit, following the manufacturer's instructions. RNA was hybridized to Affymetrix RAE230 arrays, and the resulting CEL files RMA-processed (with background correction, probe summarization using median-polish and quantile normalization). The data are available on the book companion website www.the-batch-effect-book.org/supplement. Figure 5.2(b) shows the boxplots (25−75% quantiles) of the signal intensities of the individual arrays after RMA normalization. As expected, the boxplots of all arrays/batches are similar. Lab5_AF has a slightly larger spread of signal intensities than the other batches, both for the untreated group and for the treated group. In Figure 5.2(c) it becomes more apparent that Lab5_AF (thick black line) is different from the other batches, as the distribution of the mean signal intensity values deviates from the intensity distribution of the other batches below $\log_2$ intensities of 8. As can be seen in the principal components analysis (PCA) score plot (Figure 5.2(d)), the main source of variation (along the first principal component) is the treatment factor, but there is an appreciable variance in the data coming from the different sites and the different technicians, along the second principal component. Multidimensional scaling (MDS; Grimm and Yarnold 2001) is another helpful statistical method for visualizing high-dimensional data. The nonparametric version of MDS used focuses on the representation of the (smaller) differences between individual batches. This explains the difference with respect to the PCA score plot and provides additional insight. In the MDS plot in Figure 5.1(e) the individual batches are more clearly separated compared to

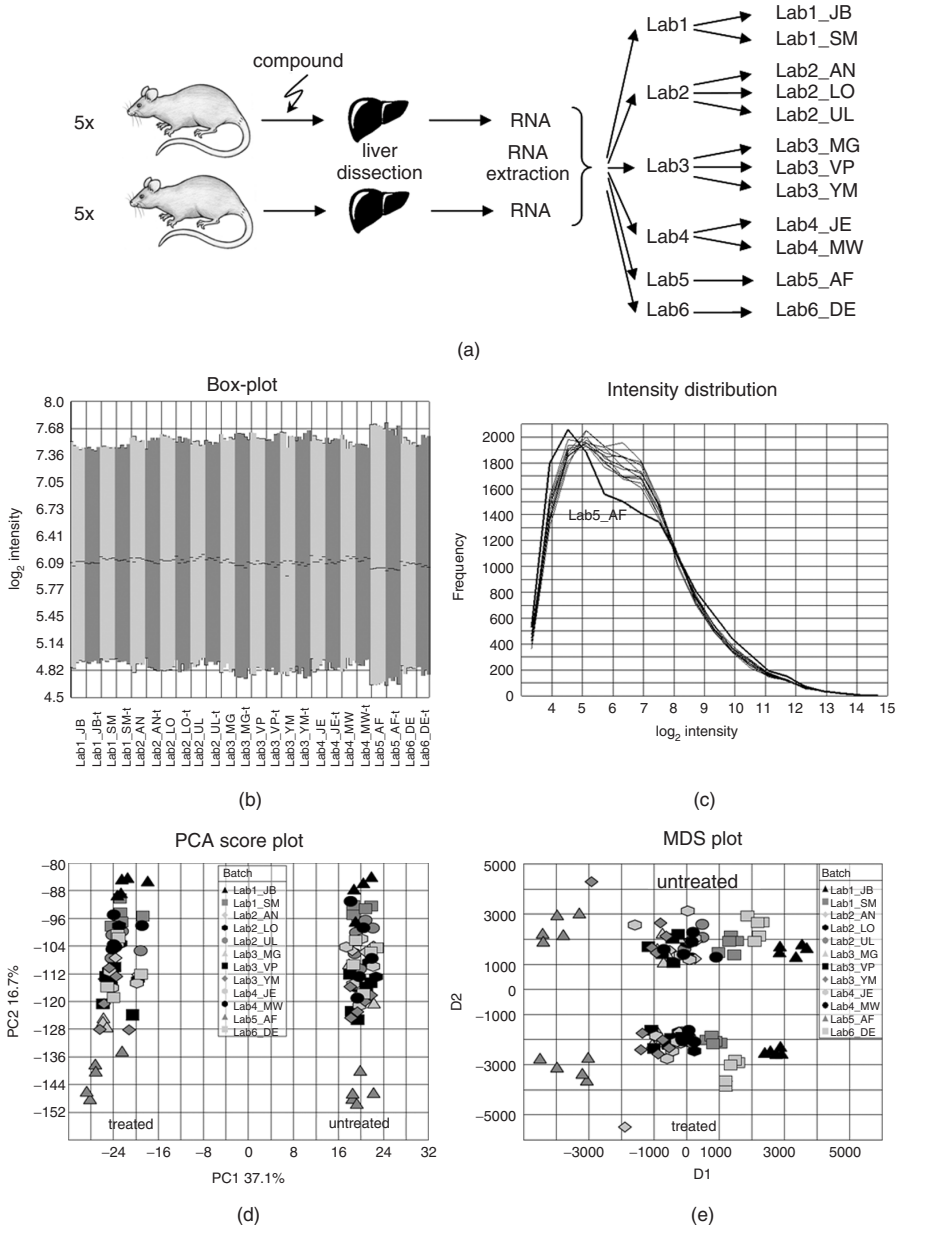

Figure 5.2 Inter-laboratory variation of gene expression caused by different handling of RNA- and microarray processing procedures. (a) Experimental design. (b) Box plots (25–75% quantiles) of signal intensities of individual arrays. (c) Intensity distributions of group means (by operator) of the untreated samples. (For simplicity the similar results for the treated samples are not shown). Lab5_AF is highlighted as the thick black line. (d) PCA score plot of 120 Affymetrix RAE230 microarrays from 12 operators in six laboratories. Along PC1 the treatment is the biggest source of variability. (e) MDS plot of the same samples as in (d). Along D1 the laboratories are the biggest source of variability.

the PCA score plot. This allows for a better qualitative estimation of the relative sizes of the overall batch effects.

As indicated by the hierarchical cluster analysis of the mean expression values (averaged over all samples per batch) of untreated samples shown in Figure 5.3(a), there is a trend that intra-laboratory variance (i.e. inter-operator variance in a single laboratory) is smaller than inter-laboratory variance (the result for the 'treated' group is very similar, and therefore not shown here). This makes sense as performance of an identical protocol within a laboratory, although handled by different technicians, should be more similar than the performance of different protocols.

What is the effect of the handling by different operators on gene expression levels? We initially address the issue of comparability of lists of differentially expressed genes per batch, the concordance between each batch and Lab1_JB as common reference. We compared lists of 1, 5, 10, 50, 100, 250, and 500 probe sets with largest fold changes between the means of treated and untreated animal groups, after application of a p-value filter ($p < 0.01$) to probe sets with a minimum mean expression of 6 (in $\log_2$ units) in the control group. The concordance of each batch with Lab1_JB as reference is between 67% and 80% for lists containing more than 10 probe sets, which shows that there is good agreement between the lists of differentially expressed genes (data not shown).

Next we analyzed how many probe sets within a range of signal intensities have a higher or lower mean signal intensity in one batch compared to another batch. The minimum mean ratio (i.e. fold change) should be larger than a threshold value which is usually thought of as a good starting point for gene discovery analyses, so we chose 1.2 or 1.5 (in either direction). We compared two very similar batches, Lab3_VP and Lab3_YM, treated and untreated samples separately (Figure 5.3(b)). We applied a fold change threshold of 1.2 which is fairly low. When using 1.5 as the lower cutoff there are hardly any probe sets changed (not shown). For a fold-change cutoff of 1.2, only a small percentage of probe sets per signal intensity range are affected at all (Figure 5.3(c), (d)). Next we turned to dissimilar batches and compared Lab5_AF and Lab1_JB, treated and untreated samples separately (Figure 5.3(e)). In each case the distribution of affected probe sets across signal intensity ranges is not uniform. The results for the dissimilar batches shows that only a few probe sets with low or very high signal intensities are affected, while up to 27% of all probe sets in the intensity range 8–9 ($\log_2$ scale) have a more than 1.5-fold difference between batches (Figure 5.3(f), (g)). In absolute numbers, looking at the untreated samples only (numbers for treated are very similar), there were 282 probe sets (i.e. 1.8% of all probe sets on the RAE230 array) affected by processing of the RNA in very similar batches, and 2940 probe sets (18.5% of all probe sets) had a fold change of at least 1.5 between two very dissimilar batches. Of those 2940 probe sets, 1168 have a p-value smaller than 0.05, and a signal intensity larger than 7. This means that 1168 probe sets would pass all standard criteria of being 'differentially expressed' between two data sets. Remember that the two data sets are based on aliquots from the same RNA!

5.2.3 Same Protocol, Different Sites (Study 2)

We wanted to see whether our findings can be also be observed in another data set with a similar setting. In 2006, an international initiative driven by the US Food

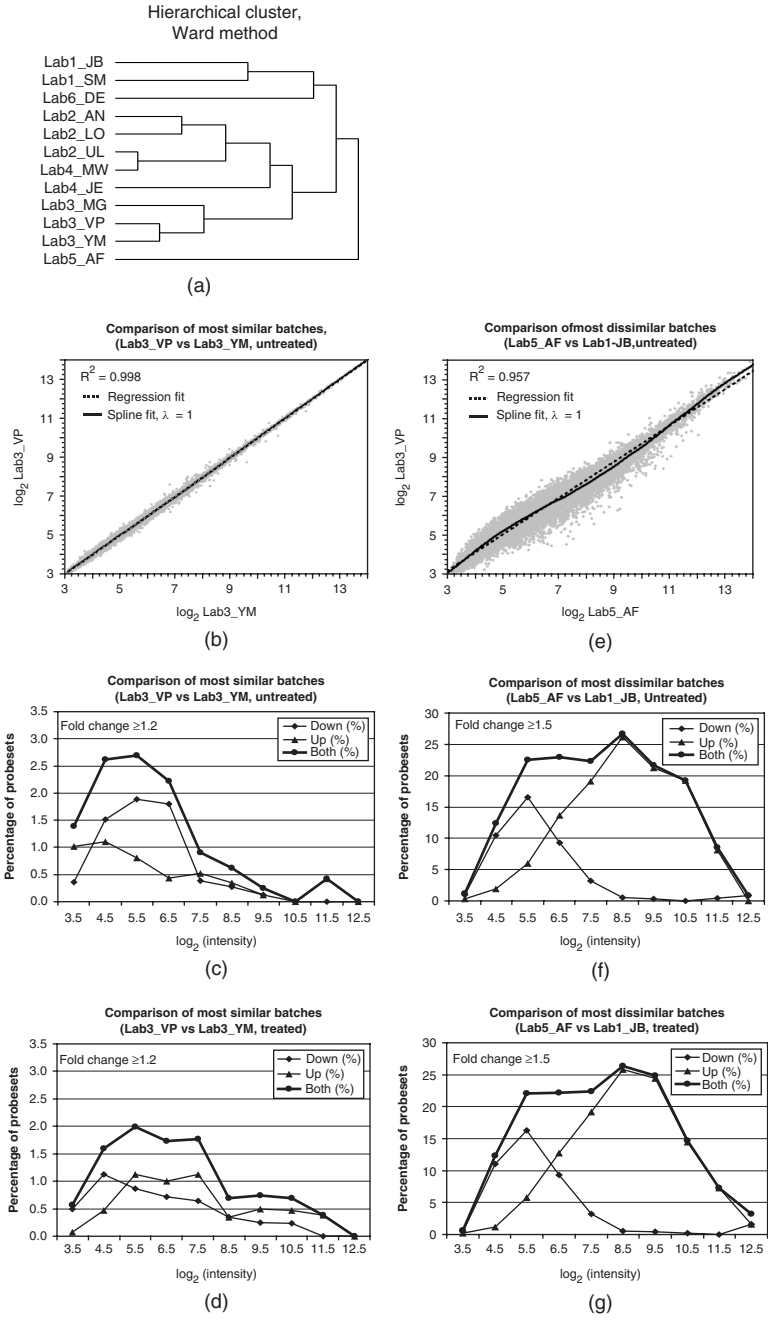

Figure 5.3 Hierarchical cluster analysis (Ward's method) of mean expression data of untreated rat liver samples (a). The percentage of probe sets with relevant fold changes between very similar batches ((b)–(d)) is smaller than between very dissimilar batches ((e)–(g)). The number of affected probe sets within a signal intensity range is given relative to the total number of probe sets within the signal intensity range.

and Drug Administration, the MicroArray Quality Control Consortium (MAQC), published data in an attempt to address the question whether microarray data sets (based on an identical set of samples) generated on different platforms and in different laboratories would yield reproducible results (Shi *et al.* 2006; http://www.fda.gov/science/centers/toxicoinformatics/maqc)). One of the MAQC data sets consisted of replicates of aliquots of two human reference RNAs and mixtures thereof: 'Universal Human Reference RNA, UHRR' (sample A, provided by Stratagene)) and 'Human Brain Reference RNA, HBRR' (sample B, provided by Ambion). The samples were processed at six sites on the Affymetrix platform, each site generating one HG-U133plus2 array per sample from five replicates (data are available from http://www.ncbi.nlm.nih.gov/geo; accession number GSE5350). Here we limit ourselves to the presentation of the UHRR data, as the results for the HBRR data are very similar. In a PCA score plot using all probe sets we can clearly see that site 6 is more different from sites 1–5 than those are from each other (Figure 5.4(a)). Unfortunately, the source of this outlying behavior of site 6 could not be identified (L. Shi, personal communication). We also noticed that the data generated at site 4 have a larger variance than the other sites, possibly due to technical issues, making it an outlier is this aspect. Probe sets which are affected most are in the intensity range of about 6–9 (on the $\log_2$ scale), which is a region whose expression data tend to be trusted as they are above array background noise and below saturation (Figure 5.4(b)).

When we plot between-batch fold changes versus signal intensity we obtain a picture similar to that obtained in the previous analysis. There are only very few affected probe sets when very similar batches are compared (Figure 5.4(c)), but a large number of probe sets in the comparison of dissimilar batches (Figure 5.4(d)). There appears to be a difference in the profiles of probe sets with higher/lower intensity in the reference batch in the UHRR data set compared to the rat data set, since in the UHRR the lines barely cross (Figure 5.4(d)) while they intersect in the rat data set (see Figure 5.4(a), (b)). This observation indicates that different batches may have different profiles.

5.2.4 Batch Effect Characteristics at the Probe Level

We wanted to investigate how a batch effect is characterized on the probe level, in contrast to the probe set level investigated so far. In an inter-laboratory comparison of gene expression data, run by Expression Analysis Inc., two laboratories processed identical samples. In this particular case, targets from the same mouse RNA source were prepared and labeled for the Affymetrix U74Av2 mouse chip. A batch effect at the probe level was detected that has been reproduced repeatedly in other contexts: A subset of probes (dark points in the scatterplot in Figure 5.5) was identified that consistently yielded higher intensities in one laboratory than another. All of these probes had a common attribute: each perfect match (PM) probe contains a set of four successive G nucleotides. Since our first discovery of this behavior in 2003, we have occasionally seen this bias occur in other Affymetrix arrays besides this illustration for the U74Av2 mouse chip. It has also been identified independently by others (Upton *et al.* 2008; Zhang, MD Anderson Cancer Center, private communication). Upton *et al.* believe that the tendency of GGGG-containing probes to be brighter than other probes in a probe set is essentially due to

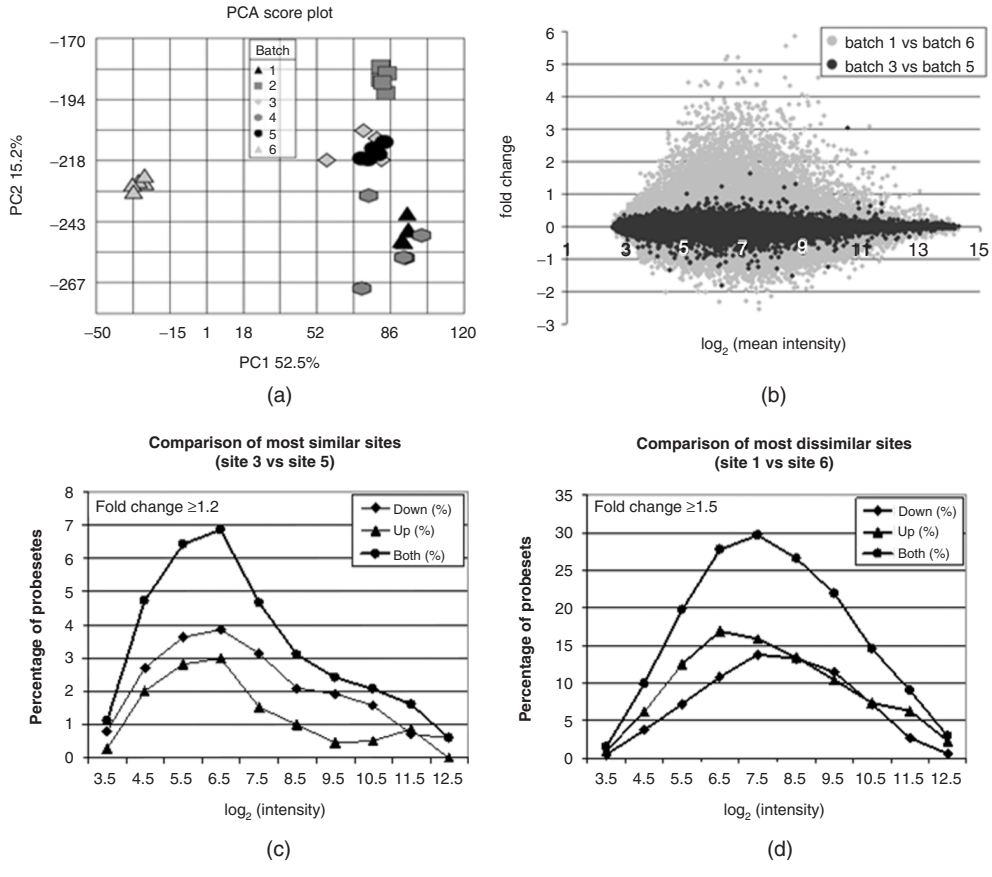

Figure 5.4 Inter-site variation (RNA processing and array hybridization) contributes to the signal intensity changes of identical RNA. (a) PCA score plot of Affymetrix HG-U133plus2 arrays hybridized with five UHRR samples in six laboratories (sites). Site 6 generated data which are markedly different from the other sites. (b) Scatterplot of mean signal intensities of site 1 and 6, and of site 3 and 5. (c), (d) Percentage of probe sets with relevant fold change from one site to another. Note the different *y*-axis scales.

probe–probe interaction on densely packed arrays. Thus the GGGG probes may form quadruplexes which show abnormal target binding qualities, as the effective association rate of probe and target is proportional to probe density. Many factors influence the stability of the quadruplexes, for instance the presence of monovalent cations or ethanol, which induces the formation of quadruplexes, storage conditions of arrays (low/high temperatures), or whether or not an array was heated prior to hybridization. In the example presented (Figure 5.5), the protocols may have been slightly different, but possibly other factors such as enzyme lot or array storage condition may have been slightly different as well. As the data were generated in 2003, a root-cause analysis is not possible anymore.

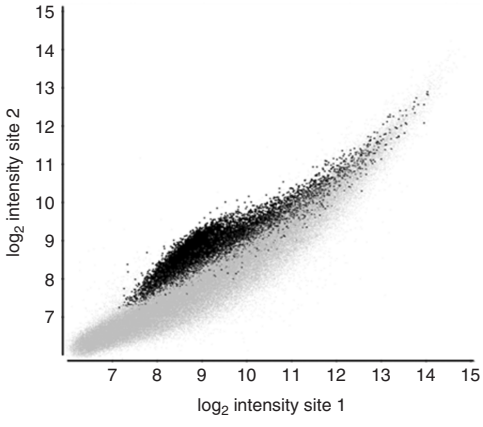

Figure 5.5 Approximately 4% (8500 of 197 000) of the perfect match probes on the Affymetrix U74Av2 array contain four consecutive G nucleotides. These probes are highlighted in black. The 'multiple-G' effect was detected in data sets hybridized in both lab 1 and lab 2. These results suggest that the hybridization changes are related to a protocol difference during target preparation.

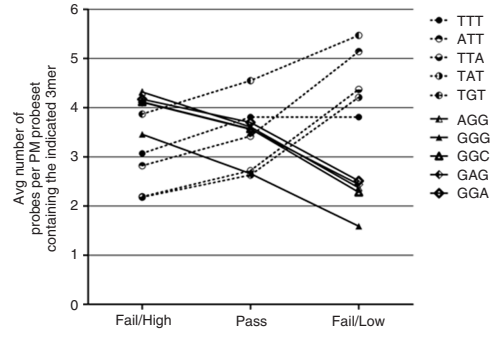

Figure 5.6 Nucleotide content of probes can be a factor in batch effects. In this example of an experiment with Affymetrix HG-U133A arrays, probe sets with higher signal intensities in lab 1 compared to lab 2 were enriched in PM probes with higher content of T nucleotides, while probe sets with lower signal intensities in lab 1 were enriched in PM probes with higher G nucleotides.

Depending on the brightness of other probes which make up the probe set, the GGGG-containing probes may receive an outlier call (other probes have low brightness) or may contribute stronger to the calculated signal intensity than other probes (gene is highly expressed and other probes have strong brightness). Thus, the probe sequence turns out to be a factor which contributes to the calculated signal intensity, which is dependent on protocol, array storage condition, and possibly other factors, which are not yet identified.

In yet another experiment, six identical replicates of RNA from the same source (UHRR, Stratagene Inc.) were prepared and hybridized to Affymetrix HG-U133A GeneChips at two different laboratories (Figure 5.6). One laboratory was used as a reference lab (lab 1) and the other laboratory was examined for consistency of absolute expression intensities (lab 2). Average signal intensity differences between the two laboratories were assessed for statistical significance using a t-test with adjustment for multiple testing. Each Affymetrix probe set was graded as 'pass' if the t-test did not indicate a significant difference. Fold change was not used as a separate criterion, although it plays an important part in the t-test itself. Due to the conservative p-value adjustment ($p < 0.000001$), very small fold changes (e.g. FC < 1.1) would rarely be flagged as significant. Roughly 5000 of the more than 22 000 HG-U133A probe sets did not pass the t-test. When root cause analysis was performed to understand this effect, a pattern was detected in those probe sets whose absolute intensity was consistently lower or higher in lab 2 than in lab 1. In particular, probe sets with probes enriched for the T nucleotide were found to have consistently

lower signals in lab 2 than in lab 1. Likewise, probe sets with probes enriched for the G nucleotide were found to have consistently higher signal in lab 2 than in lab 1. Figure 5.6 illustrates how nucleotide content/sequence can be an important factor in batch effects. The figure shows that, depending on the frequency of a particular sequence of three consecutive nucleotides within the PM probes included in a probe set, labs tend to show different signal values for that probe set. For example, when one examines those probe sets where lab 2 had greater (higher) signal than lab 1, the probes within the probe set had an average of 4 PM probes containing (for example) AGG. However, when the signals were equivalent, the probe sets had roughly 3.5 PM probes containing AGG and less than 2.5 PM probes with AGG when lab 2 probe sets had lower signal than lab 1. There were similar results for other 3-mers where there were at least two Gs. However, when we look at 3-mers containing multiple Ts, the results are reversed. Hence, again, probe sequence appears to play a big role in the observed inter-laboratory batch effects of many probe sets.

5.3 Conclusion

By using real microarray data from a toxicogenomics experiment and several experiments where RNA was reprocessed multiple times under identical or slightly different conditions, we see that batch effects can be substantial and greatly influence the data of microarray experiments and their interpretation.

The size and the direction of batch effects cannot be predicted. Our results show that batch effects are not unidirectional (i.e. all affected probe sets are biased in the same direction) or symmetrical (same number of probe sets biased positively and negatively). Additionally, the number of affected probe sets and the size of the batch effects can vary greatly. Another important finding is the observation that the proportion of affected probe sets depends on their expression intensity. The fact that we have found different signatures of batch effects in the experiments we have investigated points to a phenomenon which may not be extrapolated to other experiments and causes of bias. We suggest that each experiment should be investigated separately in this respect. Other findings indicate that the probes most affected by batch effects have certain characteristic nucleotide sequences.