BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 1. 22–31
Advance Access publication August 9, 2006
doi:10.1093/bib/bbl023

Statistically designing microarrays and microarray experiments to enhance sensitivity and specificity

Jason C. Hsu, Jane Chang, Tao Wang, Eiríkur Steingrímsson, Magnús Karl Magnússon and Kristin Bergsteinsdottir

Abstract

Gene expression signatures from microarray experiments promise to provide important prognostic tools for predicting disease outcome or response to treatment. A number of microarray studies in various cancers have reported such gene signatures. However, the overlap of gene signatures in the same disease has been limited so far, and some reported signatures have not been reproduced in other populations. Clearly, the methods used for verifying novel gene signatures need improvement. In this article, we describe an experiment in which microarrays and sample hybridization are designed according to the statistical principles of randomization, replication and blocking. Our results show that such designs provide unbiased estimation of differential expression levels as well as powerful tests for them.

Keywords: *microarray experiments; experimental design; familywise error rate; multiple comparisons; sensitivity and specificity*

THE POTENTIAL OF USING GENE EXPRESSION SIGNATURES AS DIAGNOSTIC TOOLS

The ability to analyze the expression of a large number of genes in a single experiment using microarrays has revolutionized biomedical research, particularly cancer research. Multiple microarray studies in various different cancers have resulted in gene signatures which have been proposed to be directly associated with the risk of disease recurrence, treatment or outcome. For example, by examining the expression of 25 000 genes in tumors from young (<55 years old at diagnosis), node-negative, sporadic breast cancer patients, van't Veer *et al.* [1] identified a 70 gene signature that was strongly predictive of short interval to distant metastasis. van de Vijver *et al.* [2] confirmed these findings in a larger group of 295 young patients (<53 years at diagnosis), with stage I or stage II disease. Also, Ma *et al.* [3] performed a microarray analysis on 22 000 genes in frozen breast tumors and micro-dissected tumor cells from 60 patients who had received Tamoxifen as only adjuvant therapy. From this analysis, two genes emerged, HOXB13 and IL17BR, which showed the strongest correlation with clinical outcome. Many similar studies have been performed in other diseases including renal cancer [4,5], non-small-cell lung cancer (reviewed in [6]) and in the various leukemias (reviewed in [7]).

However, several problems with microarray analysis have recently emerged. First, microarray analyses performed in different research laboratories have resulted in limited overlap in gene signatures in the same disease. In a study on gene-expression

Corresponding author. Jason C. Hsu. Department of Statistics, The Ohio State University, 1958 Neil Avenue Columbus, Ohio 43210, USA. Tel: 01 614 292 7663; Fax: 01 614 292 2096; E-mail: Hsu.1@osu.edu

Jason Hsu is a Professor in the Department of Statistics at the Ohio State University. He conducts research in multiple comparisons.
Jane Chang is an Assistant Professor of Applied Statistics and Operations Research at Bowling Green State University. Her research interests include optimal experimental design, data analysis and statistical design and analysis of microarray experiments.
Tao Wang is an Assistant Professor of Biostatistics at University of South Florida. His research interests include statistical design and analysis of microarray experiments, bioinformatics and statistical computing.
Eiríkur Steingrímsson is a Professor in the Department of Biochemistry and Molecular Biology in the Faculty of Medicine at the University of Iceland. His research focuses on the role of transcription factors in the development of melanocytes.
Magnús K. Magnússon, MD, is a consulting hematologist and principal investigator in the Departments of Laboratory Hematology and Molecular Medicine & Genetics at Landspitali-University Hospital in Reykjavik. His research interests include leukemogenesis, signaling transduction pathways and gene expression profiling in evaluating disease processes.
Kristin Bergsteinsdottir is a project leader in the Department of Genetics and Molecular Medicine at Landspitali-University Hospital, Reykjavik, Iceland. Her research focus is on gene expression analysis.

profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Wang *et al.* [8] recently found a 76 gene signature strongly predictive of metastatic disease. However, despite many similarities in study design, only three genes overlapped with the van't Veer *et al.* [1] and van de Vijver *et al.* [2] studies, namely cyclin E2, origin recognition complex and a tumor necrosis factor (TNF) superfamily protein [8]. Similarly, in diffuse large B-cell lymphoma, two groups have identified gene signatures to predict the outcome in this heterogenous cancer. Rosenwald *et al.* [9] reported a 17 gene signature and Shipp *et al.* [10] reported a 13 gene signature, both predicting survival after conventional chemotherapy. Interestingly, there was no overlap in the two gene signatures. Second, some of the proposed signatures have failed to reproduce in other populations. For example, using real-time quantitative polymerase chain reaction (PCR) and an independent cohort of patients, Reid *et al.* [1] failed to validate the predictive model of the two genes HOXB13 and IL17BR for the treatment outcome in breast cancer reported originally by Ma *et al.* [3].

It is clear that if microarrays are to be used as a diagnostic device, the results need to be reproducible and the sensitivity and specificity of the signatures need to be estimable. So far this has not been the case, and microarray studies have not paid serious enough attention to statistical *design* issues. Principles such as randomization, replication and blocking, which are considered essential to the integrity of clinical trials, generally have not been applied to microarray experiments. Different systematic (non-random) placement of probes on microarrays for different platforms may account for different biases among different platforms. Also, absence of true replications in gene probes on some platforms might have made estimations of variability unreliable. For example, the van't Veer *et al.* [1] study used Agilent microarrays with one 60-mer probe per gene per array/patient. Thus, there were no true replications to estimate the variability.

Despite the above problems and criticisms, gene expression signatures and the use of microarrays will clearly provide an important diagnostic tool for predicting disease outcome or response to treatment. Most of the gene expression signatures proposed to date usually consist of a few dozen or 100 genes, picked from 20–25 000 genes on a microarray. In order to confirm the validity of such gene expression signatures, and in order to use them in a diagnostic setting in the future, it is necessary to test them rigorously and determine their sensitivity and specificity. Before microarrays can be used as medical devices, statistical validation of proposed gene signatures in terms of sensitivity and specificity is required by the United States Food and Drug Administration (US FDA) [12,13] for pre-market approval. We believe that designing microarrays and microarray experiments statistically, using concepts that are routinely applied in clinical trials, makes such validation more feasible. Our reason for this is explained below. We will discuss statistical principles for placement of both gene probes and biological samples. For a general discussion of statistical design issues of microarray experiments, the reader is referred to Allison *et al.* [14] and Spruill *et al.* [15].

Blocking

Gene expression measurements from microarrays are potentially affected by extraneous effects such as array processing effects. In this context, a statistical 'block' is a condition under which measured gene expressions are likely to be equally affected by confounding factors. A block might, therefore, be an array or a batch of arrays processed together.

Some microarrays allow only one biological sample to be placed on each array. With one patient's sample per array, it is impossible to separate array to array variability from patient to patient variability. To normalize the expression levels from experiments using such arrays, it is unclear whether data observed under two or more conditions should be normalized together or separately. For example, suppose the purpose of the experiment is to compare expression levels of low-risk and high-risk patients. If expression levels from all the arrays are not normalized together, then the observed differences due to patients belonging to different risk groups are completely confounded with the potential differences due to array processing. On the other hand, if the arrays are normalized together using quantile normalization, then there is an implicit assumption made that switching a patient from a low-risk group to a high-risk group merely permutes the genes in terms of the ordering of their expression levels, while keeping the magnitudes of the expression levels of the genes same.

However, if the microarrays allow multiple samples to be placed on each array, then by keeping the proportion of samples from the groups to be

compared the same for each array, one can expect the collection of the magnitude of the expression levels of the genes to be approximately the same across arrays. It is then reasonable to normalize all the arrays together. Furthermore, basic statistical design principles suggest that group comparison is more efficient if equal numbers of samples from each group are placed in every block. Statistical analysis of a block design proceeds by first comparing different risk groups within blocks, and then combining such comparisons across the blocks. Such analysis increases sensitivity and specificity because it eliminates array to array variability in the comparisons. Some microarrays, including the NimbleGen 12-well arrays, can have 12 biological samples hybridized on the same slide (chip), so each array can conveniently form a block, as we shall demonstrate.

Randomization

If placement of the *biological samples* onto microarrays is not randomized, then observed differences in expression levels may be due to batch processing or position effects. To avoid such bias and confounding, we believe that the placement of biological samples onto the microarrays should be randomized. If placement of the *probes* on microarrays is not randomized, a prediction algorithm derived from one type of microarrays may contain bias and the prediction therefore not reproducible when expression levels are measured with another type of microarrays. To avoid such bias, we recommend that the placement of probes on the microarrays be randomized as well.

Replication

Measurements on gene expression levels inherently contain variability. To reliably estimate each patient's gene expression levels, we recommend each gene of a patient be probed with replicated probes or probe sets, if sample quantity and manufacturing technology allows it.

AN EXAMPLE TO DEMONSTRATE SENSITIVITY AND SPECIFICITY

To demonstrate the sensitivity and specificity of statistically designed microarrays, we conducted a proof-of-concept study using samples with known differences. We utilized four 12-well NimbleGen microarrrays, with the wells laid out in a three rows by four columns pattern on each microarray. Each of the 12 wells contained the same 200 probe sets, one probe set for each of 200 genes (167 purported breast cancer prognostic genes and 33 maintenance genes). The genes were selected from recent publications which reported gene expression signatures in breast cancer [1,16,17] and from those describing expression of maintenance genes [18–20]. The probe set for each gene consisted of sixteen 24-mer probe pairs. The placement of the probe pairs in each well was completely randomized, separately for each well.

Two different cell lines, HT-29 (colon cancer, denoted by T_c) and MCF-7 (breast cancer, denoted by T_b) were cultured. Total RNA was isolated and cDNA was synthesized for each cell line and then transcribed to generate biotin labeled cRNA.

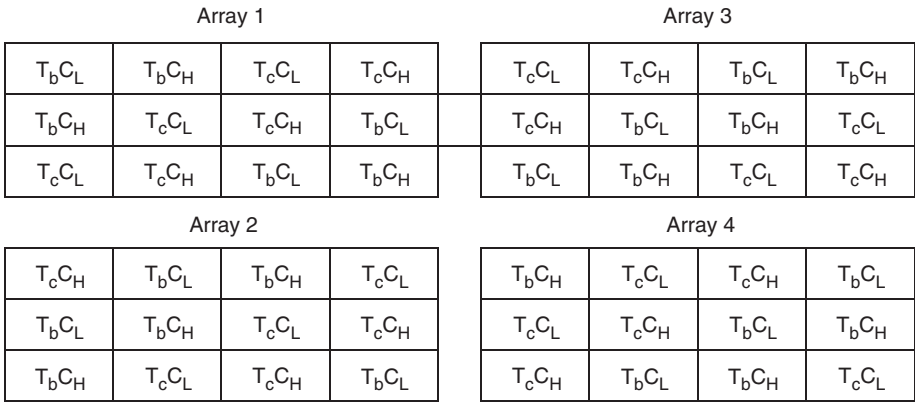
To facilitate sensitivity assessment, samples from the two cell lines were hybridized at two different concentrations: the routinely used concentration (0.5 μg/well, denoted by C_L), and at a three times higher concentration (1.5 μg/well, denoted by C_H). Samples of high concentration should show consistently higher intensities than samples of low concentration. The more significant the differences inferred, the more sensitive the microarray experiment. Specificity can be assessed by comparing the samples from the same cancer cell line, as no practical differences are expected among them.

Twelve sets of labeled cRNA samples of the four combinations T_bC_H, T_bC_L, T_cC_H and T_cC_L were hybridized to four 12-well micorrays, as three replications of 4×4 cyclic Latin Square designs (Figure 1). In each of the three Latin Squares, each combination appears exactly once in every row and column. As a result, any effect due to array, row or column will be automatically removed from the analysis by virtue of the design.

Finally, in order to avoid the systematic error (for example, T_cC_L is always between T_bC_H and T_cC_H) in this design, the actual order in which the treatment combinations were placed into the wells involved an additional randomization step, as follows. Starting with the original design, a random number generator was used to re-order the four columns and rows, separately for each of the four arrays (pp. 388–9 in [21]).

Background correction and array normalization

We performed background correction as described in Irizarry *et al.* [22]. With our design, every



Array 1			
T_bC_L	T_bC_H	T_cC_L	T_cC_H
T_bC_H	T_cC_L	T_cC_H	T_bC_L
T_cC_L	T_cC_H	T_bC_L	T_bC_H

Array 3			
T_cC_L	T_cC_H	T_bC_L	T_bC_H
T_cC_H	T_bC_L	T_bC_H	T_cC_L
T_bC_L	T_bC_H	T_cC_L	T_cC_H

Array 2			
T_cC_H	T_bC_L	T_bC_H	T_cC_L
T_bC_L	T_bC_H	T_cC_L	T_cC_H
T_bC_H	T_cC_L	T_cC_H	T_bC_L

Array 4			
T_bC_H	T_cC_L	T_cC_H	T_bC_L
T_cC_L	T_cC_H	T_bC_L	T_bC_H
T_cC_H	T_bC_L	T_bC_H	T_cC_L

Figure 1: Experimental plan—Latin Square design.

combination appears 12 times in the experiment, once in each row of every array, and three times in each column over the four arrays. This balances the potential well position effects. Also, with this design, every combination appears exactly three times in each array, so it is reasonable to expect the distribution of the expressions to be the same across the arrays. We therefore applied quantile normalization to the four arrays as described in Bolstad *et al.* [23]. Specifically, background-adjusted Perfect Matches (PM) intensities for each gene from each array were combined together as a single vector of measurement. Quantile normalization was then utilized to equalize the distributions of the four vectors from the four arrays.

Sensitivity analysis

Sensitivity, in the context of gene expression level analysis, means inferring genes that are truly differentially expressed to be differentially expressed. In our experiment, this means inferring samples of higher concentration to be differentially expressed from samples of lower concentration, within each cell line.

Let y_{gijk} denote the background-corrected and quantile-normalized logarithm of PM intensity for the gth probe set ($g = 1,\ldots,200$), ith treatment ($i = 1,\ldots,4$), jth well ($j = 1,\ldots,12$), and kth probe ($k = 1,\ldots,16$). The index $i = 1, 2, 3, 4$ corresponds to combinations T_bC_H, T_bC_L, T_cC_H and T_cC_L, respectively. We assume, y_{gijk} follows a linear additive model

$$y_{gijk} = \mu_g + \tau_{gi} + \beta_{gk} + \varepsilon_{gijk}, \quad i = 1,\ldots,4; j = 1,\ldots,12; k = 1,\ldots,16, \tag{1}$$

where μ_g is a representation of gene expression for the gth gene, τ_{gi} is the ith treatment effect on the gth gene and β_{gk} is the kth probe effect on the gth gene. The errors ε_{gijk} have mean zero, and we assume they have approximately the same variance σ_g^2 across the treatments, but not necessarily across genes. We further assume the correlation between ε_{gijk} and $\varepsilon_{gijk'}$ is ρ_g, for any $k \neq k'$.

Our model is similar to that of the robust multi-array average (RMA) in Irizarry *et al.* [22]. Theirs is a two-way linear model with independent identically distributed (*i.i.d.*) errors. We add a treatment effect term as appropriate for our design. Since our experiment has true replications (having eliminated the confounding effects by blocking), residuals are available for us to examine whether the errors are independent. If the errors were independent, then the heatmap in Figure 2 should show no particular pattern. However, there appear to be positive correlations among the probes within wells, as indicated by the vertical stripes in Figure 2. This may be due to the fact that if the sample going into a well is over-amplified, then the sequences for all the probes and genes in that well are over-amplified. We thus do not assume independent errors in our model. However, since different samples are placed in different wells, and the placement of the samples into the wells are randomized according to our statistical design, it is reasonable to assume the error vectors of array-normalized probe intensities to be *i.i.d.* for each cell line and concentration. To account for dependence among the error vectors, we re-sample with replacement *residual* vectors, each vector consisting of residuals from within each well.

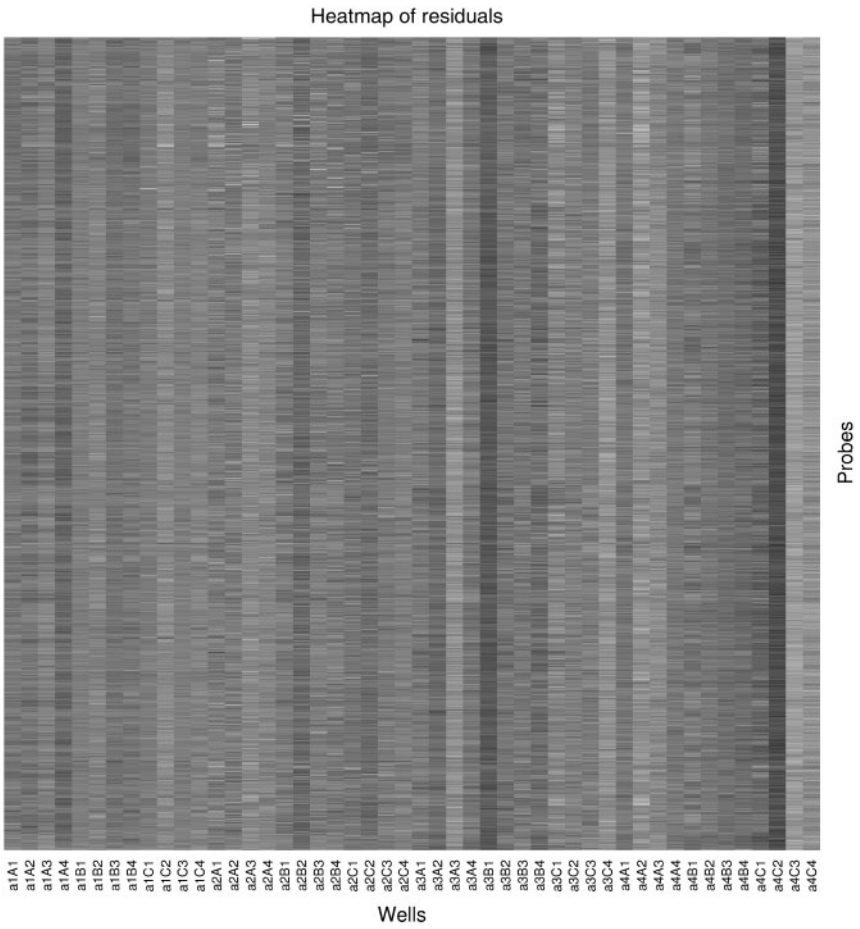


Figure 2: Heatmap of residuals, arranged by array, row and well (a1B3 is array 1, 2nd row, 3rd well from the left, for example), saturation of grey increases with residual increases (from negative to positive values).

In fitting the model, we use the constraint $\sum_{k=1}^{16} \beta_{gk} = 0$, in effect using the average of the probe intensities to represent each gene's expression level. This is in contrast to RMA, using essentially the median of the probe intensities.

Sensitivity can be assessed through comparisons of the expression levels for each gene within the same cell line between high and low concentrations, where there are known differences. The sensitivity parameters are $\tau_{g1} - \tau_{g2}$ and $\tau_{g3} - \tau_{g4}$, $g = 1, \ldots, 200$. Ordinary least squares (OLS) estimators, which are unbiased under model (1), are

$$\hat{\tau}_{g1} - \hat{\tau}_{g2} = \bar{y}_{g1\ldots} - \bar{y}_{g2\ldots},$$
$$\hat{\tau}_{g3} - \hat{\tau}_{g4} = \bar{y}_{g3\ldots} - \bar{y}_{g4\ldots}$$

For each gene, we assume the correlations among the probe intensities within wells to be the same (ρ_g) across the treatments. We allow for heterogeneity of error variances and correlations across genes. If the probes are chosen to have approximately the same affinity, then, since the placement of the probe sets and the probes within each probe set are completely randomized within each well, this assumption is not unreasonable. Under this assumption, OLS estimates turn out to be the same as generalized least squares (GLS) estimates, and they are the best linear unbiased estimates (BLUE). Their variances are

$$\mathrm{Var}\left(\hat{\tau}_{g1} - \hat{\tau}_{g2}\right) = \mathrm{Var}\left(\hat{\tau}_{g3} - \hat{\tau}_{g4}\right) = \frac{1 + 15\rho_g}{96}\sigma_g^2,$$

where ρ_g is the (unknown) correlation between the probe intensities within wells for gene g, and σ_g^2 can be estimated by

$$\hat{\sigma}_g^2 = \frac{\sum_i \sum_j \sum_k \left(y_{gijk} - \bar{y}_{gi\ldots} - \bar{y}_{g\ldots k} + \bar{y}_{g\ldots}\right)^2}{749}$$

Note that the variance of GLS estimates is a constant multiplier of σ_g^2. Thus, even though $\hat{\sigma}_g$ is not an appropriate standard error (SE) for the estimated

Table 1: Comparing the number of inferred genes differentially expressed between high and low concentrations by RMA and OLS methods

	RMA						OLS
Method	**Bonferroni**	**Holm**	**Hochberg**	**Sidak SS**	**Sidak SD**	**W&Y maxT**	**Vector re-sampling**
HT-29	0	0	0	0	0	16	80
MCF-7	47	76	198	47	76	199	200

differential expressions, it serves the purpose of appropriately scaling the estimates for the different genes in a re-sampling statistical analysis method.

All of the statistical methods we apply control the familywise error rate (FWER) at the 5% level. We performed step-down multiple testing as described in Section 5 of Hsu *et al.* [24]. To test for differential expressions of the gth gene, we used two-sample equal variance T-like statistics, $((\bar{y}_{g1\ldots} - \bar{y}_{g2\ldots})/\hat{\sigma}_g)$, based on differences of the estimates above scaled by $\hat{\sigma}_g$. Each null hypothesis H_{0I} that genes with indices in I, $I \subseteq \{1, 2, \ldots, 200\}$, are not differentially expressed (while the other genes are) is tested by the maximum T-like statistic. To calculate critical values, we re-sample with replacement the *residual* vectors of probe intensities vector at a time, resampling the residuals within each well as a unit. The residual vectors we re-sample from are pooled across the four treatments, following the results of Pollard and van der Laan [25] and Huang *et al.* [26] which state, when the sample sizes of the treatment are equal, re-sampling methods that re-sample the residuals pooled across the treatments remain approximately valid even if the variance–covariance matrices are not exactly the same. The critical value of the 5% test for the null hypothesis H_{0I} that genes with indices in the subset I are not differentially expressed is the 95th percentile of the bootstrap distribution of the maximum of the absolute values of the re-sampled T-like statistics with indices in I. Our multiple testing thus satisfy conditions S1, S2 and S3 in Section 5 of Hsu *et al.* [24] and can thus be executed in a step-down fashion. With a bootstrap replication size B of 50 000, the results of our step-down testing are reported in Table 1 under the heading OLS.

We first show that multiple tests based on OLS is competitive with those based on RMA median polish estimates. We applied the following multiplicity adjustment methods which do not take into account the joint distribution of the test statistics to RMA estimates: the single-step method based on the Bonferroni inequality, Holm's step-down method based on the Bonferroni inequality, Hochberg's step-up method based on Simes' inequality, Sidak's single-step (Sidak SS) method based on the product inequality and Sidak's step-down (Sidak SD) method based on the product inequality, to RMA estimates. We also applied the Westfall and Young [27] re-sampling method, which does take into account the joint distribution of test statistics, to RMA estimates. P-values were computed by complete enumeration of all permutations. The number of differentially expressed genes inferred by these methods is reported under the headings of Bonferroni, Holm, Hochberg, Sidak SS, Sidak SD and W&Y maxT in Table 1.

Given the same data and the same FWER control, product inequality methods will always do at least as well as Bonferroni inequality methods, but not by much (pp. 13–4 in [28]). Step-down and step-up methods will always do at least as well as their corresponding single-step methods. Whether a step-down or a step-up method does better, depends on data (Section 7 in [29]).

Our OLS analysis is a step-down maxT method which takes the correlations among the test statistics into account. So our probe level analysis result is directly comparable with the W&Y maxT RMA analysis result. Our OLS result is better than all RMA results. Clearly, the OLS method is competitive with the RMA methods.

We then show that statistically designing microarray experiments leads to better sensitivity. In order to assess the benefit of the proposed block design over unblocked ones, we generated 100 random designs, each randomly assigning the 48 treatment combinations (12 each for T_bC_H, T_bC_L, T_cC_H and T_cC_L) to the four arrays (12-wells per array). Background correction, array quantile normalization, OLS estimation and multiple testing were then performed as described earlier. For the HT-29

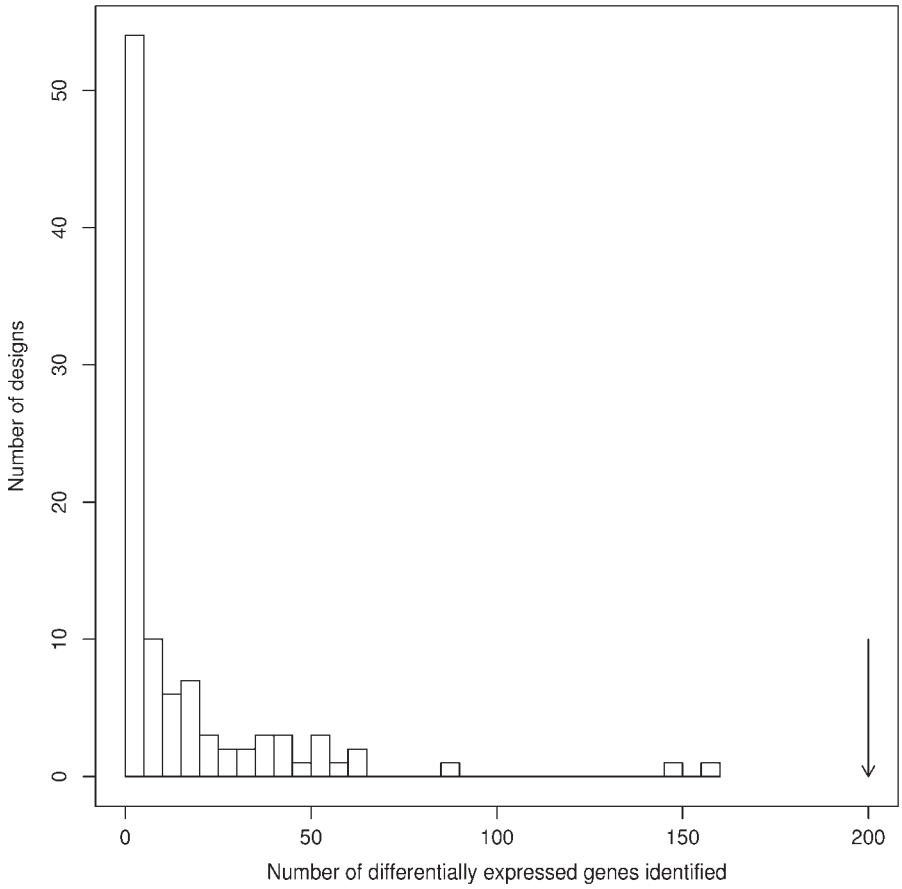


Figure 3: Comparing the number of inferred genes differentially expressed between high and low concentrations for MCF-7 (breast cancer) cell line. Random design *vs* statistical design

(colon cancer) cell line, 98 of the random designs did not find any gene to be significantly expressed, one random design found one gene, while another one found 17 genes to be differentially expressed. Figure 3 shows, for the MCF-7 (breast cancer) cell line, the histogram of the number of genes inferred to be differentially expressed between high and low concentrations by the random designs.

We found that, generally, the more balanced the allocation of treatments to wells is within each array, the higher the sensitivity. For example, 'Design III' in Table 2 is highly unbalanced, leading to no gene identified as differentially expressed with either cancer cell lines. Designs I and II are more balanced, leading to 17 and 160 genes identified as differentially expressed for the colon and breast cancer cell lines, respectively.

Specificity analysis

Specificity, in the context of gene expression level analysis, means inferring genes that are not differentially expressed to have expression levels close to each other. Specificity analysis should not be treated as tests for significant difference, as a lack of statistically significant difference can be due to small sample size or noisy data. Instead, specificity should be treated as an *equivalence* problem, in analogy to bioequivalence [30]. In our experiment, this means most genes from the same cell line and concentration from different arrays should have confidence intervals for differential expressions with limits not too far from zero.

For specificity analysis, we add *array* as a factor in modeling. For each of the 200 genes, for samples from the same cell line at the same concentration, we computed confidence intervals for pairwise differential expressions across the four arrays, resulting in $200 \times \binom{4}{2} = 1200$ confidence intervals. These confidence intervals are computed using a bootstrap technique similar to what we described for sensitivity analysis, except that they are not adjusted for multiplicity. (Equivalence confidence intervals do not necessarily have to be adjusted for multiplicity [30].)

Table 2: Comparing the number of inferred genes differentially expressed between high and low concentrations by OLS method

	Design I				Design II				Design III			
Cell line	**Colon**		**Breast**		**Colon**		**Breast**		**Colon**		**Breast**	
Concentration	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**
Array 1	3	1	4	4	4	2	3	3	1	4	6	1
Array 2	5	5	1	1	3	3	2	4	2	4	1	5
Array 3	3	4	3	2	2	4	4	2	3	0	4	5
Array 4	1	2	4	5	3	3	3	3	6	4	1	1
Differentially expressed genes identified	17		56		0		160		0		0	

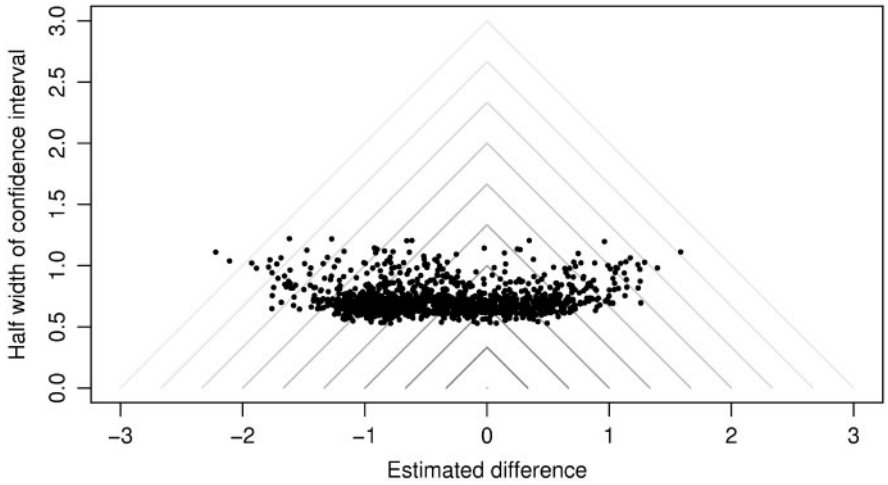


Figure 4: Equivalence confidence intervals for the HT-29 cell line.

Figures 4 and 5 are graphical representations of 1200 95% individual equivalence confidence intervals for low concentration for each of the cell lines. Confidence intervals for high-concentration comparisons are similar, and are not displayed here. Similar to volcano plots for *P*-values, these so-called Location-Scale displays map the intervals into points [31]. For example, a symmetric confidence interval is mapped to a point on the Location-Scale display with the horizontal coordinate being the center and the vertical coordinate being the half width of the confidence interval. It is a useful technique to display many confidence intervals simultaneously. The span of a confidence interval can be easily inferred from the location of its corresponding point on the Location-Scale display. In the equivalence setting, if the point representing a confidence interval is within the inverted-V contour corresponding to −3 and +3 say, then that confidence interval is contained in (−3, +3). The narrower these confidence intervals are around zero, the more specific the microarray experiment. Evidently from Figures 4 and 5, almost all the equivalence confidence intervals are contained in (−3, +3), which indicates that an agreement within ±3 is the extent specificity can be achieved for an experiment with sample sizes such as ours.

DISCUSSION

A good microarray experimental design randomly places probes in each well/array and randomly assigns biological samples to wells/arrays, to avoid systemic nuisance effects.

A good microarray experimental design also balances the allocation of treatments to avoid potential confounding effects. It is common to observe a batch processing effect in microarray experiments. Often, the batch effect is stronger than the treatment effect of interest.

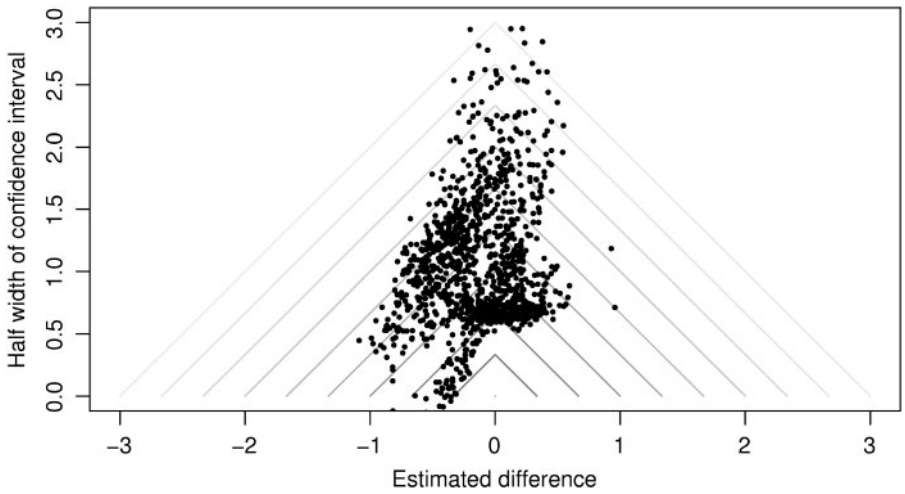


Figure 5: Equivalence confidence intervals for the MCF-7 cell line.

If the proportion of treatments within each batch is not the same between batches, then observed treatment differences may be confounded with batch effects, leading potentially to irreproducible results. Some microarrays such as those used in our experiment allow multiple samples to be placed on each array, so one can balance the number of samples from the groups to be compared on each array. One way to generalize our finding to microarrays limited to hybridizing one sample per array is to think of our experiment of four arrays with 12-wells each as processing 48 arrays in four batches of 12 arrays each. Our investigation shows that if one balances the proportion of treatments within each batch, then normalization across batches will not lead to confounding of batch and treatment effects.

Finally, a good microarray experimental design also balances the allocation of treatments with respect to array/batch to improve the sensitivity and specificity. Our investigation shows that, if one balances the proportion of treatments within each array/batch, then by eliminating array/batch variability in the comparisons, sensitivity of the statistical analysis is increased.

Key Point

- Statistically designing microarray experiments may improve the reproducibility of gene expression signatures for cancer prognoses. We describe an experiment in which microarrays and sample hybridization are designed according to the statistical principles of randomization, replication and blocking. Such designs avoid confounding effects, provide unbiased estimation of differential expression levels, as well as increase the sensitivity and specificity.

Acknowledgements

This study was approved by the National Bioethics Committee (ref. VSNb2005010026/03-15) and the Data Protection Authority (ref. 2005010047) of Iceland. This project was funded by the Icelandic Technology Development fund (Rannís – The Icelandic Centre for Research). The authors thank Sigridur Valgeirsdottir, Elsa Thorey Eysteinsdottir and Erla Hronn Geirsdottir (Nimblegen Iceland) for expertise assistance on sample processing, microarray hybridization and for their valuable discussions, Professor Yoonkyung Lee for her useful idea on sample size computation, and Nan Jiang for her comments. J.H.'s research was supported by NSF Grant No. DMS-0505519.

References

1. van't Veer L, Dai H, van de Vijver M, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.
2. van de Vijver M, He Y, van't Veer L, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**:1999–2009.
3. Ma XJ, Wang Z, Ryan PD, *et al*. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004;**5**:607–16.
4. Jones J, Out H, Spentzos D, *et al*. Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res* 2005;**11**:5730–39.
5. Kosari F, Parker AS, Kube DM, *et al*. Clear cell renal cell carcinoma: Gene expression analyses identify a potential signature for tumor aggressiveness. *Clin Cancer Res* 2005;**11**: 5128–39.
6. Petty RD, Nicolson MC, Kerr KM, *et al*. Gene expression profiling in non-small cell lung cancer: from molecular mechanisms to clinical application. *Clin Cancer Res* 2004;**10**: 3237–48.
7. Margalit O, Somech R, Amariglio N, *et al*. Microarray-based gene expression profiling of hematologic malignancies: basic concepts and clinical applications. *Blood Rev* 2005; **19**:223–34.

8. Wang Y, Klijn J, Zhang Y, *et al*. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;**365**:671–79.
9. Rosenwald A, Wright G, Chan WG, *et al*. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002;**346**: 1937–47.
10. Shipp MA, Ross KN, Tamayo P, *et al*. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002; **8**:68–74.
11. Reid JF, Lusa L, De Cecco L, *et al*. Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *J Natl Cancer Inst* 2005;**12**:927–30.
12. Food and Drug Administration. Pharmacogenomic Data Submissions: Guidance for Industry, 2005. http://www.fda.gov/cber/gdlns/pharmdtasub.htm.
13. Food and Drug Administration. Drug-diagnostics Co-development Concept Paper, 2005. http://www.fda.gov/cder/genomics/pharmacoconceptfn.pdf.
14. Allison D, Cu X, Page G, *et al*. Microarray data analysis: from disarray to consolidation and consensus. *Nature Rev Genet* 2006;**7**:55–65.
15. Spruill S, Lu J, Hardy S, *et al*. Assessing sources of variability in microarray gene expression data. *Bio techniques* 2002; **33**(4):916–23.
16. Yu K, Lee CH, Tan PH, *et al*. A molecular signature of the Nottingham prognostic index in breast cancer. *Cancer Res* 2004;**64**:2962–8.
17. Ahr A, Holtrich U, Solbach C, *et al*. Molecular classification of breast cancer patients by gene expression profiling. *J Pathol* 2001;**195**:312–20.
18. Hsiao LL, Dangond F, Yoshida T, *et al*. A compendium of gene expression in normal human tissues. *Physiol Genomics* 2001;**7**:97–104.
19. Warrington JA, Nair A, Mahadevappa M, *et al*. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* 2000;**2**:143–7.
20. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet* 2003;**19**:362–5.
21. Dean A, Voss D. *Design and analysis of experiments*. New York: Springer, 1999.
22. Irizarry RA, Hobbs B, Collin F, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;**4**:249–64.
23. Bolstad BM, Irizarry RA, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**: 185–93.
24. Hsu JC, Chang YC, Wang T. Multiple comparisons in screening for differential expressions from microarray data. In: Angela Dean and Sue Lewis, (eds), *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics* Chapter 6, New York: Springer-Verlag, 2006;139–55.
25. Pollard KS, van der Laan MJ. Resampling-based multiple testing: asymptotic control of type I error and applications to gene expression data. *J Stat Plan Infer* 2005;**125**:85–100.
26. Huang Y, Xu H, Calian V, *et al*. JC. To permute or not to permute. To appear in *Bioinformatics* 2006
27. Westfall P, Young S. *Resampling-based multiple testing*. New York: Wiley, 1993.
28. Hsu JC. *Multiple comparisons: theory and methods*. London: Chapman and Hall, 1996.
29. Huang Y, Hsu JC. Hochberg's step-up method: cutting corners off Holm's step-down method. Submitted for publication 2006.
30. Berge RL, Hsu JC. Bioequivalence trials, intersection–union tests, and equivalence confidence sets. *Stat Sci* 1996; **11**:283–319.
31. Peruggia M, Hsu JC, Huang Y. The location-spread displays of estimates and their associated measures of dispersion. Technical Report 730, Department of Statistics, The Ohio State University. *Am Stat* 2005.