5610–5618 *Nucleic Acids Research, 2009, Vol. 37, No. 17* *Published online 17 July 2009*
doi:10.1093/nar/gkp573

Importance of randomization in microarray experimental designs with Illumina platforms

Ricardo A. Verdugo[1,*], Christian F. Deschepper[2], Gloria Muñoz[3], Daniel Pomp[3] and Gary A. Churchill[1]

[1]The Jackson Laboratory, Bar Harbor, ME 04609, USA, [2]Institut de Recherches Cliniques, Montreal, QC, Canada and [3]Department of Nutrition, University of North Carolina, Chapel Hill, NC 27599, USA

Received March 4, 2009; Revised June 3, 2009; Accepted June 22, 2009

ABSTRACT

Measurements of gene expression from microarray experiments are highly dependent on experimental design. Systematic noise can be introduced into the data at numerous steps. On Illumina BeadChips, multiple samples are assayed in an ordered series of arrays. Two experiments were performed using the same samples but different hybridization designs. An experiment confounding genotype with BeadChip and treatment with array position was compared to another experiment in which these factors were randomized to BeadChip and array position. An ordinal effect of array position on intensity values was observed in both experiments. We demonstrate that there is increased rate of false-positive results in the confounded design and that attempts to correct for confounded effects by statistical modeling reduce power of detection for true differential expression. Simple analysis models without *post hoc* corrections provide the best results possible for a given experimental design. Normalization improved differential expression testing in both experiments but randomization was the most important factor for establishing accurate results. We conclude that lack of randomization cannot be corrected by normalization or by analytical methods. Proper randomization is essential for successful microarray experiments.

INTRODUCTION

Establishing causality is the ultimate goal of any experiment aiming to discover the mechanisms underlying natural phenomena. Among several approaches for establishing causality, one of the most widely used is randomization, as first described by Sir Ronald Fisher in 'The design of experiments' (1). The random assignment of experimental units to treatments controls the likelihood that any factor other than the treatment is the cause of the association (2,3). This recommendation is explicitly stated in most reference books devoted to the analysis of microarray data (4). Nonetheless, this basic principle, that has been widely accepted and applied in many fields of science, is often ignored at several levels in the design of microarray experiments. Consequently, many investigators and consulting analysts are faced with the challenging and sometimes impossible task of performing *post hoc* analyses for experiments that were not properly randomized. As a result, causality can no longer be established with confidence.

Experimental designs should be tailored to the microarray platform. Numerous studies can be found in the literature on how to design efficient experiments with two-color (5–8) and even three- or four-color arrays (9). Such designs try to minimize or balance the variability introduced by design factors such as dye bias, and array effects. For two-color arrays, dye-swap and blocking are commonly employed. Provided that appropriate designs are used, linear models can account for dye and array effects in an analysis of variance (10). The multiple alternatives for pairing samples in either references or different versions of loop designs have been studied both theoretically (11,12) as well as empirically (13). However, the role of sample position in one-color platforms that hold multiple samples in a single slide has received less attention. These platforms present features that may lead to sources of technical variation that had not been anticipated. For custom arrays (NimbleGen, 200 probes), samples are placed in a 3-row by 4-column arrangement of wells and confounding effects may arise from the array, row, or column of the samples. It has been shown that accuracy and reproducibility of differential expression testing in this platform is improved by experimental designs that employ blocking, randomization and replication (14).

We examined the effect of sample position effects using whole-genome Illumina BeadChips in which multiple samples are hybridized on a single BeadChip. Each BeadChip (chip hereafter) represents an experimental block and all

*To whom correspondence should be addressed. Tel: +1 207 288 6715; Fax: +1 207 288 6847; Email: ricardo.a.verdugo@gmail.com.

samples on a single chip can potentially share common effects due to processing. Furthermore, samples are hybridized to arrays in ordered positions, which are labeled in alphabetic order. For instance, positions A–H for MouseRef-8 and A–F for Mouse-6 Sentrix® mouse platforms.

We present empirical evidence for the presence of significant chip and position effects in an Illumina microarray study. We compared the results from two experiments that used the same set of RNA samples, but where samples from each experimental group were placed on the chips using either a confounded design layout or a randomized arrangement. In addition, we considered the impact of both normalization and the choice of statistical model on the results of the analyses. We discuss the implications for experiments performed on Illumina and other microarray platforms.

MATERIALS AND METHODS

Experimental design

Two experiments were performed using the same RNA samples obtained from cardiac muscle of 16 individual mice from one of two different genetic backgrounds (genotype): the C57BL/6J (B) inbred strain and the C57BL/6J-chrY$^{A/J}$/NaJ (BY) congenic strain in which the Y chromosome from the A/J strain has been introgressed onto the B background (15). Four mice from each strain were castrated (treatment C) and four mice were subject to sham operations but remained intact (treatment I). In the Confounded experiment (Figure 1), all samples from the B genotype were hybridized to the first chip and samples of the BY genotype were hybridized to the second chip. Samples from intact mice are in positions A–D and samples from castrated mice are in positions E–H. Thus, genotype was fully confounded with chip, and treatment was partially confounded with array position on the chip. In the Randomized experiment (Figure 1), block randomization was used. Samples were selected at random, subject to the constraint that two samples of each type appear on each chip. Since the same RNA samples were used in both experiments, any differences can be attributed to technical factors. Normalization should, in principle, eliminate these effects. In the following sections, we assess the impact of normalization and design factors (chip and position) on tests for genotype, treatment and interaction in these two experiments.

Subjects and samples manipulation

Experimental animals (corresponding to the offspring of breeding pairs obtained from the Jackson Laboratory) were euthanized at 12 weeks of age, between 9:00 and 10:00 AM, for tissue collection. Procedures were approved by the Institut de Recherches Cliniques de Montréal (IRCM) Institutional Animal Care Committee and conducted according to guidelines issued by the Canadian Council on Animal Care. RNA was isolated from samples of myocardium from left ventricles of 16 mice using the *RNeasy minikit* (Qiagen Canada, Mississauga, ON, Canada). Biotinylated probes were prepared from 50 ng of total RNA, using the Ambion Illumina TotalPrep RNA Amplification kit (Applied Biosystems, Streetsville, ON, Canada). The complete set of RNA was hybridized to MouseRef-8 BeadChips (25K, Illumina, San Diego CA, USA). The two experiments were performed on different days. Bead level intensity values were summarized using BeadStudio v3.1 without normalization. Local background correction was applied by default using BeadStudio. Raw probe intensity values were imported to the R 2.7.2 (UNIX) language/environment for normalization and analysis (R. Development Core Team, http://www.r-project.org).

Probe annotation

In the Illumina platforms, probes are bound to a set of ~30 beads. We will refer to the trim-average of intensity across each set as the probe level values. Probes were annotated using the ArrayGene software as described previously (16). Briefly, every sequence or gene id associated with a given probe in the gene list obtained from the Illumina website was cross-referenced to a local MySQL database of sequence and gene identifiers that is based on EntrezGene IDs (Genome build 37). Probes associated with more than one EntrezGene were not annotated and were not used for the functional analysis. More than 4700 out of 17 077 genes on the MouseRef-8 platform are targeted by more than one probe. Due to the potential variation of transcripts (e.g. alternative splicing), we did not merge probe level values into a single gene-level summary. When reporting effects at the probe level, we use the terms probe and transcript interchangeably.

Data preprocessing

All samples passed quality control inspection in both experiments (Supplementary Figures 1S and 2S). Quantile normalization was applied to data from each experiment separately (17). Variance was stabilized with a $\log_2$ transformation. Probes for unexpressed genes were removed based on Present/Absent calls as recommended in (18). In short, transcripts were called as present when probability of detection was ≥ 0.96 (as estimated with BeadStudio, using the intensity distribution of negative probes), and were retained when present in at least 50% of samples from any treatment/genotype group in either experiment. Out of 25 697 probes, 13 903 were retained for statistical analysis.

Assessment of array-level design effects

Linear models were fit to the median intensity across all probes from each array on each chip. The following model was fit to the data from each experiment separately before normalization:

$$m_{ij} = \mu + c_i + \beta_i p_j + e_{ij}, \quad \mathbf{1}$$

where m_{ij} is the $\log_2$ transformed median intensity from array j of chip i, μ is the mean, c_i is the effect of the chip i, β_i is the coefficient of regression on position p_j within chip i, and e_{ij} is the residual. Within-chip R^2 was estimated by fitting a reduced model with only the position term for

each chip with the *lmList* function in the *nlme* package for the R language/environment (nlme package; http://cran.r-project.org).

Assessment of probe level effects

Differential expression and technical effects were tested by fitting a set of linear models at the probe level to data from each experiment separately, with and without normalization (Table 1). All models in Table 1 are some version a cell means model where μ_k is the mean intensity of samples from four experimental groups: B.I, B.C, BY.I and BY.C. The effects of genotype, treatment and their interaction were estimated by appropriate contrasts between experimental groups: genotype = (B.C + B.I) – (BY.C + BY.I), treatment = (B.I + BY.I) – (B.C + BY.C) and interaction = (B.I – BY.I) – (B.C – BY.C). To assess association to within chip position effects, two additional contrasts were calculated: treatment_B = B.I – B.C and treatment_BY = BY.I – BY.C. The *UnAdj* model tests experimental groups without adjustment for chip and position. The *LinReg* model includes a linear regression adjustment for position. The approximation of position effects by a linear regression was based on empirical observations from this and other data sets (Appendix). The *Full* model relaxes the linearity assumption by including position as a random effect (Table 1). The effect of Chip was included in the *LinReg* and *Full* models only for the Randomized experiment as this factor was completely confounded with genotype in the Confounded experiment and could therefore not be estimated.

All model fitting and ANOVA analyses were performed in the R language/environment version 2.7.2 using the R/maanova software version 1.13 (19). *F*-values were calculated using shrinkage estimates of error variance (20). *P*-values were derived from expected *F* distributions and corrected for multiple comparisons with the *q*-value method described in (21). Probes were selected by FDR < 0.1. For the genotype or treatment effects, only probes without a significant interaction were selected. The proportion of differentially expressed genes π_1 was estimated from the distribution of *P*-values as described in (21).

The rank order of *P*-values was used as a selection index and $1 - r_s$ was used as distance measure for hierarchical clustering, where r_s is the Spearman correlation between *P*-values from each pair of models. Negative correlations result in distances >1. Use of Spearman correlation allows us to compare relative order of significance without specifying significance thresholds. The *hclust* function of the *stat* package for R was used to generate an agglomerative average hierarchical clustering of ranked *P*-values from each model. Venn diagrams were produced with the *limma* package for R (22).

Functional testing for lists of differentially expressed genes

Only probes that could be associated with a unique EntrezGene as in (16) were used for functional testing. In cases where genes were targeted by multiple probes, genes were selected if at least one probe was significantly differentially expressed (DE) (FDR < 0.1). Over-representation on Gene Ontology (GO) terms was assessed by a Fisher's exact test comparing the odds ratio for membership to a given GO term between DE and non-DE genes (23).

RESULTS

We compared data from two experiments: one with a Confounded design and another with a Randomized design, using the same samples, on Illumina BeadChips (Figure 1). Statistical modeling was used to evaluate the effects of both design and experimental factors in each experiment separately. By design factors, we refer to technical factors that are associated with the processing of samples, hybridization protocols, and/or microarray platform. Experimental factors refer to biological factors of interest, such as genotype, treatment and their interaction.

Table 1. Statistical models applied to probe level data

Experiment	Models	df_1	df_2	Abbreviation
Confounded	$y_{kl} = \mu + \mu_k + e_{kl}$	3	12	c.unadj
	$y_{ijkl} = \mu + \beta_i p_j + \mu_k + e_{ijkl}$	5	10	c.linreg
	$y_{jkl} = \mu + P_j + \mu_k + e_{jkl}$	9	6	c.full
Randomized	$y_{kl} = \mu + \mu_k + e_{kl}$	3	12	r.unadj
	$y_{ijkl} = \mu + c_i + \beta_i p_j + \mu_k + e_{jkl}$	6	9	r.linreg
	$y_{ijkl} = \mu + C_i + P_j + \mu_k + e_{kl}$	11	4	r.full

df_1, model degrees of freedom; df_2, error degrees of freedom; y, $\log_2$ probe level intensity; μ, overall mean; μ_k, mean for experimental group k; β_i, coefficients of regression within chip; p_j, position covariate (values 1–8); P_j, random position effect (levels A–H); C_i, random chip effect; Lower case indicates fixed and upper case random effects; c, confounded; r, randomized; unadj, unadjusted; linreg, adjustment by linear regression; full, adjustment by full mixed model.
The prefix 'raw.' or 'norm.' is applied to the model abbreviation in the text and figures to indicate if the model was fit to raw or normalized data, respectively (Figure 4).

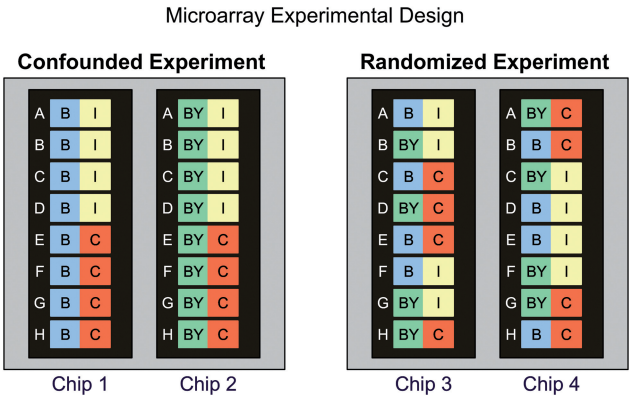

Figure 1. Experimental design. Layout of samples for the Confounded and Randomized experiments. Black rectangles represent BeadChips. Sentrix Position for individual arrays are displayed along the left side of BeadChips (A–H). Each experiment used two BeadChips. Colors represent genotype [blue = C57BL/6J (B); green = C57BL/6J-chrYA/J/NaJ (BY)] and castration treatment [yellow = intact (I); red = castrated (C)].

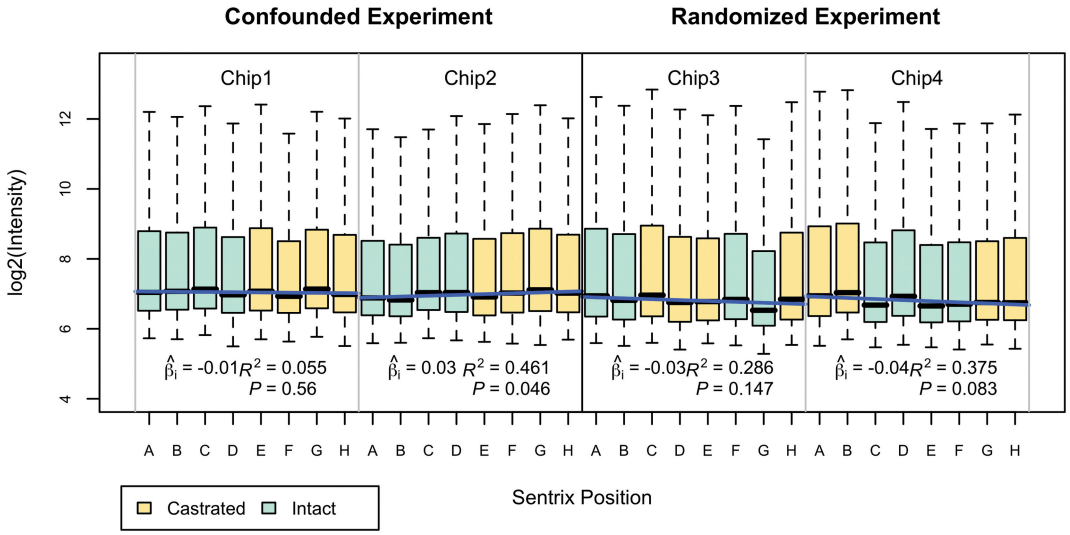

Figure 2. Boxplot for raw data from the both experiments. Outliers are not shown for clarity. Boxplots of raw intensity values for negative probes in the Confounded and Randomized experiments are shown by position in the four Chips. Color differentiates castration treatment (yellow = castrated; green = intact). Blue lines are best linear fit on the medians by position.

We assessed the effects of design factors on bias and power of statistical tests for differential expression. By differential expression we refer to variation in the abundance of a gene's messenger RNA (mRNA) across samples in an experiment. The amount of mRNA in a sample is measured indirectly by fluorescence intensity of a given probe in the microarray. Design factors can affect intensity but only experimental factors can produce differential expression. With this distinction made between biological and technical effects on intensity, we note that statistically significant probes do not imply differential expression. Throughout the text, we use this distinction and try to differentiate true differential expression from increased number of selected probes that result from confounded design effects.

Assessment of array-level design effects

We initially investigated the overall intensity distribution from each array to assess systematic effects due to chip and position by fitting model 1 (see 'Materials and Methods' section) to the median intensity before normalization (Figure 2). We tested for differences in intensity between chips within each experiment and for a linear trend of intensity across positions within a chip. Chip had a suggestive effect in the Confounded experiment (P-value = 0.137) but not in the Randomized experiment (P-value = 0.955). In addition, we observed position effects: (i) in the Confounded experiment, a positive trend was observed in Chip 2 ($R^2 \sim 0.25$, P-value = 0.046); (ii) in the Randomized experiment, a decreasing trend across positions was observed in both chips ($R^2 \sim 0.17$, P-value = 0.09). This pattern was also observed in negative control probes, i.e. probes that have no target in the mouse genome and samples are therefore expected to reflect background signal (Supplementary Figure 3S). For subsequent analyses, we adjusted the data using quantile normalization (17) to equalize the median intensities across all chips and arrays within each experiment.

Effect of design on number of differentially expressed genes

Differential expression associated with experimental factors was assessed using two-way ANOVA. This is the simplest and most obvious method to study the effects of genotype, treatment, and their interaction. If an interaction is present, it means that the effect of a factor on the response variable depends on the levels of the other factor. In the present study, a significant interaction can be interpreted as treatment effects that depend on the genotype of the animals or equivalently, a genotype effect that depends on the treatment. This was done by an ANOVA of the *UnAdj* model with raw and normalized data from the Confounded and Randomized experiments (Table 1).

We found that more probes were selected for genotype, treatment and interaction in the Confounded experiment. However, we caution that this does not necessarily indicate greater power to detect biological effects in this experiment. To assess the biological information content of the differentially expressed gene lists in each experiment, we examined the lists for enrichment of GO biological processes (23). Although this is not a perfect test, it is reasonable to expect greater enrichment in a biologically coherent gene list. The DE gene list for treatment and genotype in the Randomized experiment, although shorter, identified more biological process terms than the corresponding list from the Confounded experiment (Figure 3). For the interaction gene list, the Confounded experiment identified more terms. GO biological processes are highly interrelated and multiple GO terms may effectively represent the same groups of DE genes (Supplementary Table 3S). Nevertheless, our results suggest that the longer gene lists for main effects selected in the Confounded experiment may be due to detection of

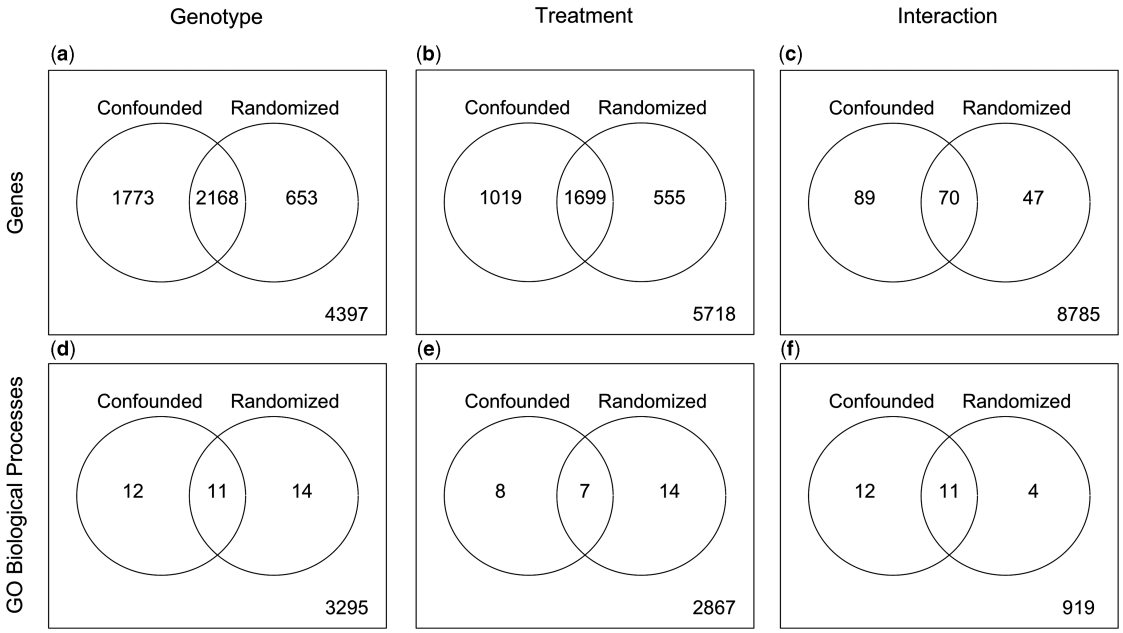

Figure 3. Confounded experiment is enriched with false-positive results. Genes selected for differential expression in two replicated experiments. Genes were selected by the *UnAdj* model on normalized data. Venn diagrams group number of unique genes selected (**a–c**) and GO Biological Processes associated to those genes (**d–f**).

technical effects on intensities. We explore the importance of such design effects below.

Accounting for design effects by statistical models

The differential expression analysis presented above ignores any effects that chip or position within chip may have on intensity measurements. This is a common procedure in microarray analysis, since it is expected that normalization would correct for technical factors before gene level model testing. However, for the purpose of assessing the fraction of probes that are affected by such design effects, we tested differential expression with two additional models: *LinReg*, and *Full* models (Table 1). We fit the *UnAdj, LinReg* and *Full* models to data from each experiment separately, with and without normalization. The *Full* model is a mixed model that accounts for chip and position when testing for experimental effects. The *Full* model was fit using REML method (24). The *LinReg* model accounts for Chip effects as in the *Full* model, but position effects are assumed to follow a linear trend. *LinReg* was fit as a fixed effects model.

We performed an unsupervised hierarchical clustering of results from the *UnAdj*, *LinReg* and *Full* models by the ranking of *P*-values (Figure 4). The branch length in the tree is inversely proportional to the correlation of the rank order of genes (r_s) between a given pair of models (see 'Materials and Methods' section for details). Test results for the Genotype, Treatment and Interaction terms showed a hierarchical pattern in which normalization was the most important factor, followed by experimental design and lastly by the analysis model. The exception to this pattern was the fitting of the *LinReg* model to the Confounded experiment. This model produced very different results, with zero or negative rank correlation compared to all other cases. Furthermore, *LinReg* model selected no probes for genotype effects in normalized data. Inspection of the effects estimated from each model in the Confounded experiment revealed that genotype effects from the *LinReg* model are affected by the slope of position effects in a given chip. In this model, the genotype effect is effectively the difference between the intercepts of the regression lines fit to position effects (Supplementary Figure 6S). The *LinReg* model does not seem to be appropriate in the Randomized experiment either. It produced unexpected *P*-value distributions for Chip and Position effects (Supplementary Figure 8S). Although the intent was to achieve *post hoc* correction for position effects, the *LinReg* model is clearly problematic and should be avoided in practice.

The *UnAdj* and *Full* models in the Confounded experiment produced the closest results to those from the Randomized experiment ($r_s \sim 0.4$–0.5). Both models performed similarly in terms of ranking of probes for all three experimental factors ($r_s > 0.97$). This was true for both the Confounded and the Randomized experiments. However, the absolute difference in their *P*-values changed with normalization. In the Randomized experiment, the *UnAdj* model using raw data mainly gave higher *P*-values than the *Full* model and the *Full* model therefore selected more probes. However, these *P*-values had little correlation to the results from either model on normalized data. Therefore, although the largest number of probes for treatment and genotype is selected with the *Full* model using raw data (Figure 4), the selected set of probes is very different than when normalized data is used. In contrast, once data is normalized, the *Full* model selects fewer probes than the *UnAdj* model and both models produce *P*-values with high rank order agreement.

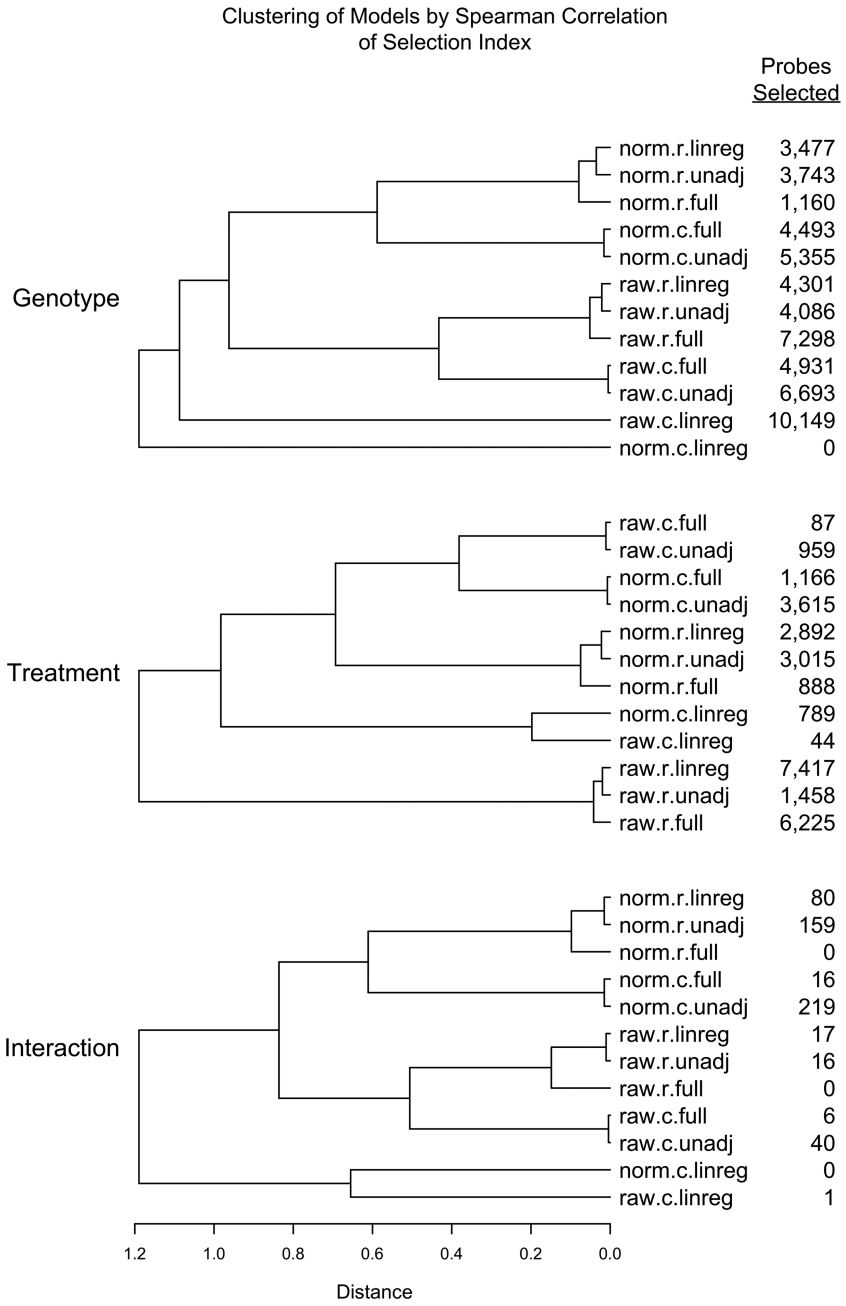

Figure 4. Hierarchical clustering of fitted models for adjustment of chip and position effects (Table 1). Models distance was measured as 1-Spearman correlation between *P*-values. Negative correlations produce distances higher than 1. Branches are labeled to indicate normalization method (norm, raw), experiment (c,r) and analysis model (unadj, linreg, full). Number of probes selected at FDR < 0.1 are shown at right.

This indicates that although the *Full* model is effective at reducing the residual variance in raw data, it is not a replacement for the use of normalization to remove systematic effects. Once data are normalized, the cost paid in error degrees of freedom (Table 1) outweighs the benefit from the reduction in residual variance, decreasing the model's power and consequently the *UnAdj* model performs better.

Overall, no combination of normalization and modeling examined here could provide results for the Confounded experiment with a correlation to the Randomized experiment that was >0.5. The *LinReg* model was the worst choice for the confounded experiment. The *UnAdj* model proved to be the best option for normalized data, regardless of the design layout.

Effect of confounding 'Chip' and 'Genotype'

The *LinReg* and *Full* models account for design effects at the probe level by estimating probe specific coefficients for chip and position. In the Confounded experiment, adjusting for the effects of chip would also remove any genotype effects due to the complete confounding of these two factors. Therefore, we cannot adjust for chip effects in this

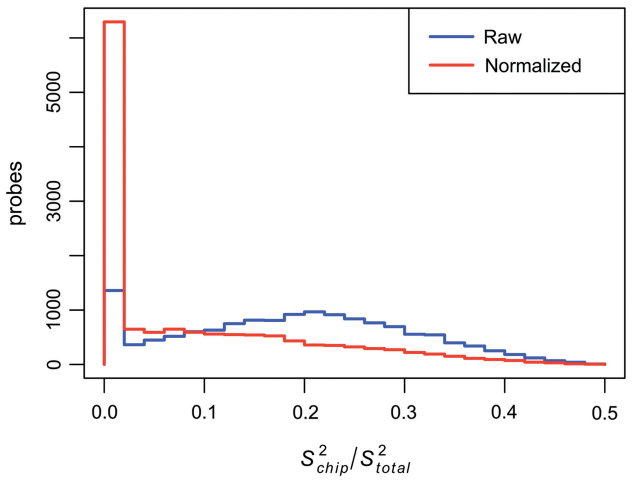

Figure 5. Variance due to chip effects. The variance component associated to chip effects (S^2_{chip}) was estimated by REML from the *Full* model in raw and normalized data from the Randomized experiment. The histogram shows the distribution of $S^2_{\text{chip}}/S^2_{total}$ across probes.

model and as a consequence, differential gene expression between genotypes cannot be distinguished from technical effects due to chip (for the mathematical arguments on this issue see pp. 181–183 of ref. (3)). We inspected the proportion of total variance from the *Full* model due to chip effects across probes in the Randomized experiment (Figure 5). Although no significant difference between chips was detected on the median intensity from raw data in the Randomized experiment (Figure 2), we found a large proportion of probes where chip effects accounted for a significant fraction of the total variance (>0.1 = 10 614 probes; >0.3 = 2511). These chip effects were greatly reduced by normalization but were not eliminated from all probes (>0.1 = 5130 probes; >0.3 = 928). Similar random chip effects must also be present in the Confounded experiment, resulting in apparent genotype effects in hundreds of probes even after normalization.

Effect of confounding 'Position' and 'Treatment'

In order to assess the association between position and treatment effects, we calculated treatment effects for each genotype separately (see 'Materials and Methods' section). In the Confounded experiment, the treatment comparison for the BY genotype was performed in a single chip, i.e. Chip 2. Therefore, we assessed presence of association between position effects for Chip 2 and treatment effects in the BY genotype (Treatment_BY) in this experiment using the *LinReg* model (Figure 6). Despite the bad performance of this model shown above, we use it here because it provides insights into position effects and the properties of this model. Position and Treatment_BY effects were highly correlated both before and after normalization. This correlation was not observed in either chip from the Randomized experiment (see Randomized panels in Figure 6 and Supplementary Figure 9S). Under the null hypothesis, the distribution of effects should center around the expected value of zero. However, the distribution of position effects in raw data from Chip 2 and 4 were shifted to the positive and negative side, respectively. This is evidence of bias in estimated position effects was produced by the systematic position effects observed both in target (Figure 2) and control probes (Supplementary Figure 3S). Treatment_BY effects also showed deviation form the zero expectation, which could be due to the confounding with Position. Normalization moved the distributions of intensities, centering gene-specific position and treatment effects in both experiments around zero. However, random gene-specific position effects were still present for many probes (Figure 6). Thus, normalization was effective at correcting overall systematic effects but it does not break correlations between the partially confounded factors nor does it eliminate many probe-specific position effects.

To explore the effects of these corrections on power for detecting differential expression, we selected lists of genes with the largest adjusted treatment effects (FDR < 0.1; Table 2). The *P*-value distributions for treatment effects (Supplementary Figures 7S and 8S) were used to estimate the proportion of DE genes π_1 as explained in methods. Estimates of π_1 from all three models in raw data from the Confounded experiment were lower (0.06–0.25) than in the Randomized experiment (0.52–0.68, Table 2). Although, normalized data showed much more similar estimates, $\hat{\pi}_1$ was smaller for the *LinReg* and *Full* models in the Confounded experiment, whereas $\hat{\pi}_1$ for *UnAdj* model was larger (Table 2). These results indicate that correcting for chip and position effects in raw data causes large reduction in power for the Confounded experiment and almost no probes are selected (44 in *LinReg* and 87 in *Full*). Once systematic effects are removed by normalization, adjustment by the *LinReg* and *Full* models still reduced power, although to a lesser degree, when compared to the Randomized experiment. Not adjusting for position effects (*UnAdj* model), selects more probes in the Confounded experiment but at the expense of increased false positive results. Therefore, neither normalization nor post hoc adjustment could match the precision and power of the Randomized experiment.

DISCUSSION

Inspection of raw data from an Illumina microarray experiment revealed significant chip and position effects at the array and probe levels, with an approximately linear trend of decreasing intensity values from positions A to H (Figure 2 and Supplementary Figure 3S). These trends were also observed in a large and independent Illumina dataset (Appendix). In that experiment, position effects showed significant linear trends in 10 out of 24 chips, all of negative sign. Some of those trends were twice as big as those observed here. This and previous observations from other experiments lead us to conclude that position effects are prevalent in Illumina experiments and can be of large magnitude. Results in the Appendix also show that position trends can extend beyond single chips and even across batches, pointing at lability of the fluorescent dye as a possible cause for these position effects. If this is the case, microarray experiments with

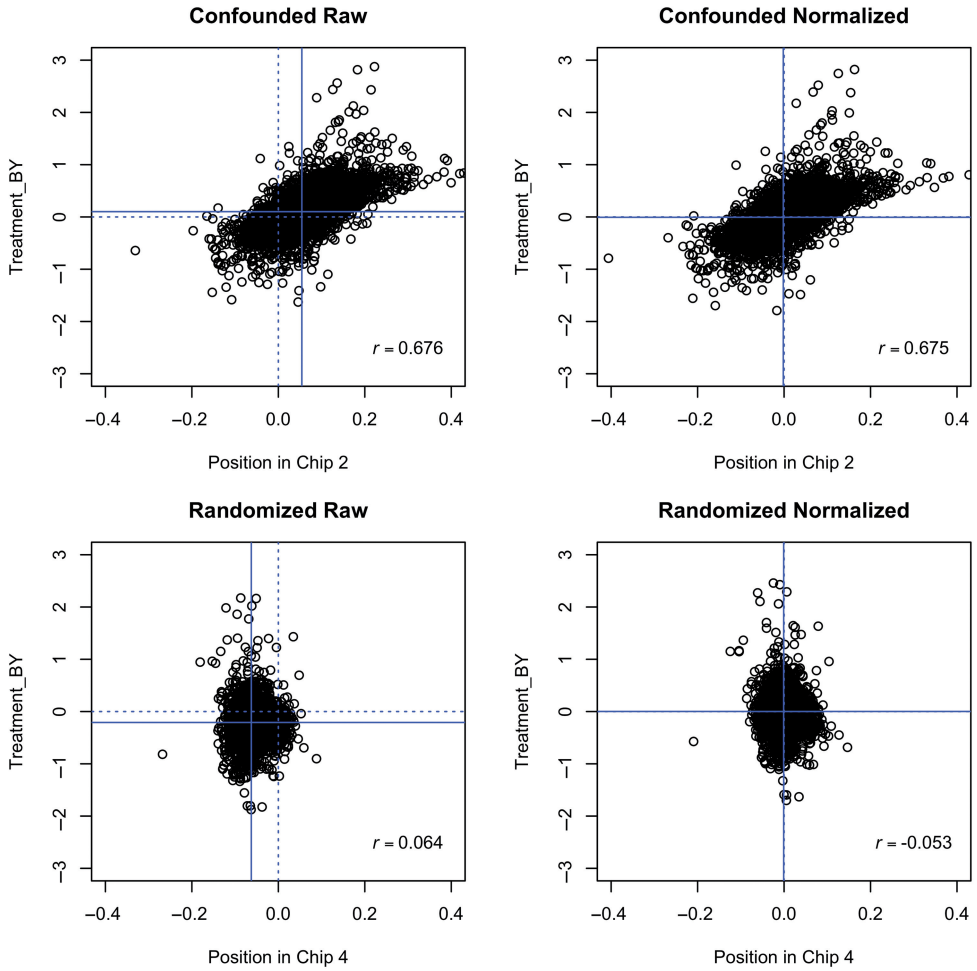

Figure 6. Association between treatment and position effects introduced by a confounded design. Scatter plots show position in Chip 2 (Confounded) and Chip 4 (Randomized) versus treatment effects in the BY genotype (Treatment_BY) from each experiment. Dotted blue lines cross the *y*- and *x*-axis at 0. Solid blue lines denote the median position (vertical) and Treatment_BY effects (horizontal).

Table 2. Differentially expressed genes tested for Treatment effects before and after normalization

Model	Normalization	Confounded experiment		Randomized experiment	
		$\hat{\pi}_1$	Probes elected	$\hat{\pi}_1$	Probes selected
UndAdj	Raw	0.25	959	0.52	1472
	Normalized	0.43	3615	0.39	3123
LinReg	Raw	0.06	44	0.68	7417
	Normalized	0.34	789	0.41	2892
Full	Raw	0.11	87	0.67	6225
	Normalized	0.33	1166	0.41	888

Estimated π_1 and number of probes selected by FDR < 0.1 are shown by experiment, normalization procedure and model used.

any platform may suffer systematic effects due to order of sample processing. Regardless of the source or sign of position effects, the confounding of factors of interest with order of hybridization would have similar consequences on power and accuracy as those reported here. This could explain the higher power observed from balanced over unbalanced designs on a custom NimbleGen chip (14) and highlights the relevance of our findings for any microarray platform.

Performing a standard test for differential expression due to genotype, treatment, and interaction effects with the *UnAdj* model demonstrates that the Confounded experiment selects more probes than the Randomized experiment. This could be due to higher power or to increased false discoveries. The confounding of design factors with experimental factors does not allow us to distinguish these explanations in the Confounded experiment (3). The reduced enrichment for biological annotations compared to the Randomized experiment favors the hypothesis of increased false discoveries. Analysis of the Randomized experiment demonstrated that hundreds of probes can show significantly large chip and position effects even after normalization. It is likely that the same effects occur in the Confounded experiment, which is consistent with the larger number of selected probes.

We further explored the possibility of controlling for design effects by statistical modeling. This resulted in reduced power to detect genotype, treatment, and interaction effects in both designs (Figure 4). Adjusting for

position effects by a linear regression was a particularly bad choice, producing the least consistent ranking of probes compared to all other models. Presumably, this was a consequence of controlling for a factor that, even after normalization, was correlated with treatment (Figure 6). The *UnAdj* and *Full* models produced highly similar results ($r_s > 0.97$), but both had a correlation of only ~0.4–0.5 to results from the randomized experiment. Therefore, when an experiment has confounded design factors, one cannot improve on the *UnAdj* model and normalized data. This combination also presented the highest power in the Randomized experiment (Supplementary Figure 5S). Based on our results, we discourage the use of *post hoc* corrections for design effects, whether the experiment was confounded or randomized. Furthermore, we conclude that randomization of samples to position in the Illumina chip is essential for reliable inferences of differential expression.

When randomization is not explicitly spelled out in protocols, the natural tendency is to perform hybridizations following a logical order according to the identifiers of the samples, which often include information on the factors of interest in the experiment. We demonstrated that this leads to a confounded design that can significantly impact the outcome of the analysis by increasing the false positives rate. Therefore, statistical designs that randomize the arrangement of samples on chips or any other systematic features of a protocol are advised. Randomized relabeling of samples before processing provides a simple strategy to avoid confounding. Consideration of potential blocking factors as well as balance and replication can suggest more sophisticated designs for accurate and unbiased experiments (14). All of these principles are accepted requirements for clinical trials testing new drugs (25) and should be required as minimal standards for publication of microarray results. These measures can improve reproducibility of microarray experiments by avoiding augmented rates of false discoveries and providing confidence that differential intensity reflects differential expression.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the 'Genome Quebec Innovation Centre' for performing the Illumina BeadChip hybridization experiments, and thank in particular Isabelle Guillet for excellence of the service provided. We also acknowledge Imogen Hurley for careful proofreading of the manuscript.

FUNDING

The NIGMS National Center for Systems Biology: Center for Genome Dynamics (grant number GM076468 to G.A.C.); the Canadian Institutes for Health Research (CIHR; grant number MOP-64391 to C.F.D.). Funding for open access charge: GM076468.

Conflict of interest statement. None declared.

REFERENCES

1. Fisher,R.A. (1951) The design of experiments 6th ed. Olivier and Boyd, London, UK.
2. Rubin,D.B. (2008) For objective causal inference, design trumps analysis. *Ann. Appl. Stat.*, **2**, 808–840.
3. Lehmann,E. and Romano,J.P. (2008) *Testing Statistical Hypotheses* 3rd ed. Springer, New York, USA.
4. Draghici,S. (2003) *Data Analysis Tolls for DNA Microarrays*. Chapman and Hall/CRC, Boca Raton, FL, p. 196.
5. Churchill,G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.*, **32(Suppl)**, 490–495.
6. Kerr,M.K. (2003) Design considerations for efficient and effective microarray studies. *Biometrics*, **59**, 822–828.
7. Yang,Y.H. and Speed,T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.*, **3**, 579–588.
8. Kerr,M.K. and Churchill,G.A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, **77**, 123–128.
9. Woo,Y., Krueger,W., Kaur,A. and Churchill,G. (2005) Experimental design for three-color and four-color gene expression microarrays. *Bioinformatics*, **21(Suppl. 1)**, i459–i467.
10. Churchill,G.A. (2004) Using ANOVA to analyze microarray data. *Biotechniques*, **37**, 173–175, 177.
11. Altman,N.S. and Hua,J. (2006) Extending the loop design for two-channel microarray experiments. *Genet. Res.*, **88**, 153–163.
12. Kerr,M.K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.
13. Vinciotti,V., Khanin,R., D'Alimonte,D., Liu,X., Cattini,N., Hotchkiss,G., Bucca,G., de Jesus,O., Rasaiyaah,J., Smith,C.P. *et al.* (2005) An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics*, **21**, 492–501.
14. Hsu,J.C., Chang,J., Wang,T., Steingrímsson,E., Magnússon,M.K. and Bergsteinsdottir,K. (2007) Statistically designing microarrays and microarray experiments to enhance sensitivity and specificity. *Brief. Bioinform.*, **8**, 22–31.
15. Nadeau,J.H., Singer,J.B., Matin,A. and Lander,E.S. (2000) Analysing complex genetic traits with chromosome substitution strains. *Nat. Genet.*, **24**, 221–225.
16. Verdugo,R.A. and Medrano,J.F. (2006) Comparison of gene coverage of mouse oligonucleotide microarray platforms. *BMC Genomics*, **7**, 58–58.
17. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
18. McClintick,J.N. and Edenberg,H.J. (2006) Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinform.*, **7**, 49.
19. Wu,H., Kerr,M., Cui,X. and Churchill,G. (2002) MAANOVA: a software package for the analysis of spotted cDNA Microarray experiments. In Parmigiani,G., Garett,E.S., Irizarry,R.A. and Zeger,S.L. (eds), *The analysis of gene expression data: methods and software, Statistics for Biology and Health*. Springer, New York, USA, pp. 313–341.
20. Cui,X., Hwang,J.T., Qiu,J., Blades,N.J. and Churchill,G.A. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.
21. Storey,J.D. (2002) A direct approach to false discovery rates. *J. Roy. Stat. Soc. B*, **64**, 479–498.
22. Smyth, G. (2005), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York, pp. 397–420.
23. Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
24. McCulloch,C.E. and Searle,S.R. (2001) *Generalized, Linear, and Mixed Models*. Wiley-Interscience, New York.
25. ICH. (1998) E9: Statistical Principles for Clinical Trials. Available at: http://www.ich.org/cache/compo/475-272-1.html#E9 (28 June 2009, date last accessed).