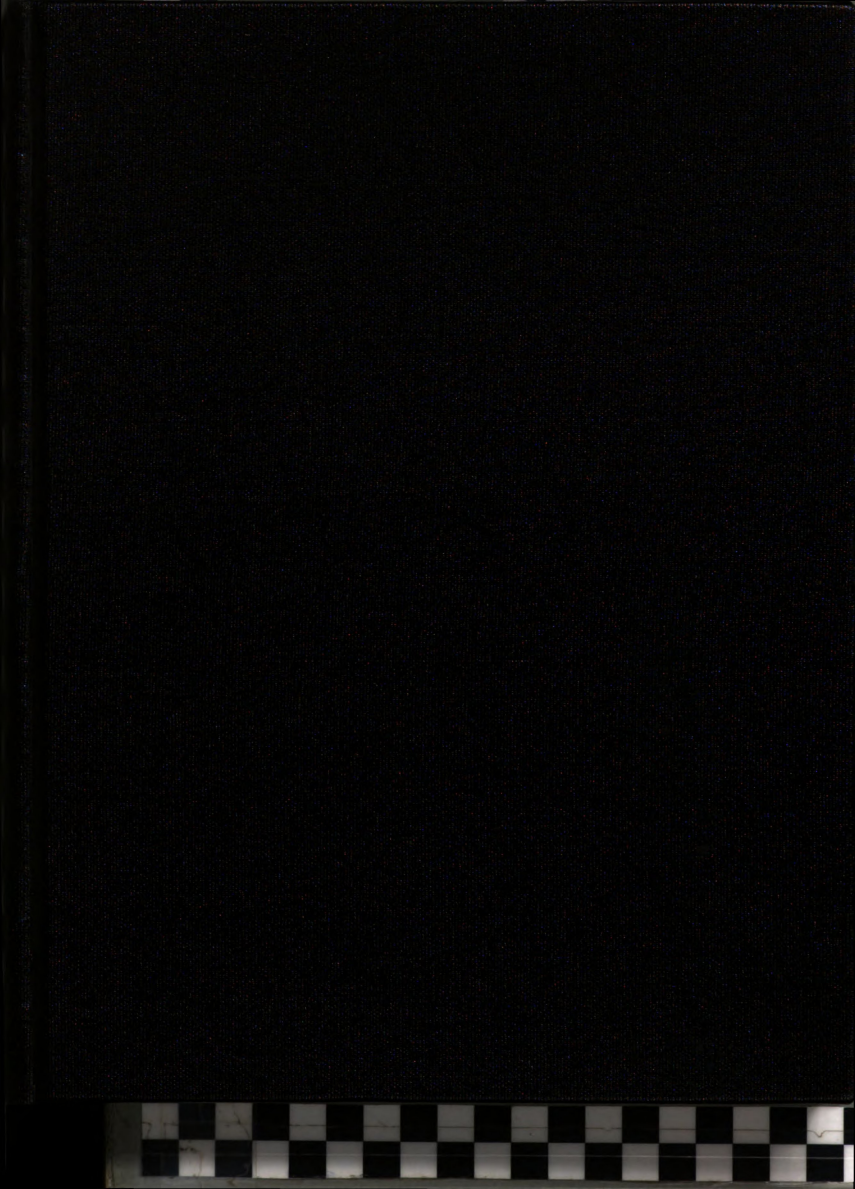


This is to certify that the
dissertation entitled

IMPROVING EXPERIMENTAL DESIGN AND STATISTICAL INFERENCE FOR TRANSCRIPTIONAL PROFILING EXPERIMENTS

presented by

Juan Pedro Steibel

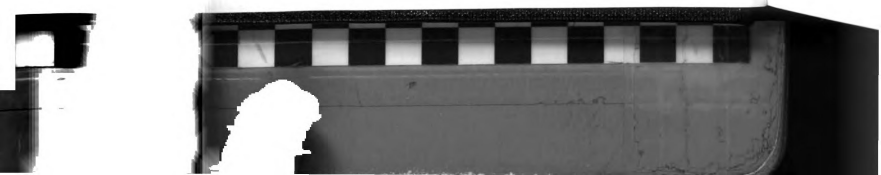
has been accepted towards fulfillment
of the requirements for the

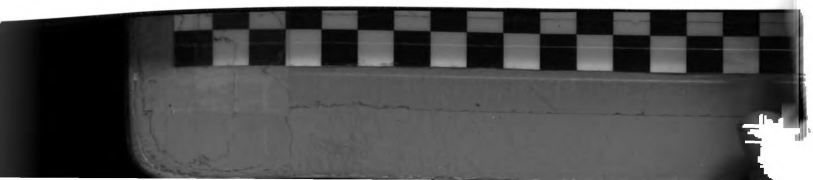
PhD degree in Animal Science

Robert J. Tempelman
Major Professor's Signature

March 7, 2007
Date

MSU is an affirmative-action, equal-opportunity employer





IMPROVING EXPERIMENTAL DESIGN AND STATISTICAL INFERENCE FOR TRANSCRIPTIONAL PROFILING EXPERIMENTS.

By

Juan Pedro Steibel

A DISSERTATION

Submitted to
Michigan State University
In partial fulfillment of the requirements
For the degree of

DOCTOR OF PHILOSOPHY

Department of Animal Science

2007

ABSTRACT

IMPROVING EXPERIMENTAL DESIGN AND STATISTICAL INFERENCE FOR TRANSCRIPTION PROFILING EXPERIMENTS.

By

Juan Pedro Steibel

Functional genomic studies resort to transcriptional profiling experiments in order to gain insights into the function of genes and their patterns of expression and regulation. The most commonly used techniques for gene expression studies in animal science are two-color microarrays and quantitative reverse transcription polymerase chain reaction (qRT-PCR). The overall objective of this dissertation was to increase the efficiency and statistical power of these studies by further optimizing experimental designs and statistical analysis methodology for microarray and validation (qRT-PCR) studies.

The first study addressed the comparison of alternative reference designs, including a potentially more efficient variant called the blocked reference design (BRD). The relative efficiency of the various designs was shown to depend on the number of treatments and the relative magnitude of biological and residual variability. All designs were deemed practically equally efficient when the magnitude of the biological variances is negligible relative to the residual variance; however, with a large number of treatment comparisons and a larger biological to residual variance ratio, the BRD can be 1.5 to 2 times more efficient than more traditional alternatives.

The second study compared models based on using either log-intensities or log-ratios as the response variables in linear model analyses, demonstrating their inferential equivalences and differences. The results indicated that the main difference between the two models primarily depended on the ability of the corresponding design to recover

inter-slide information using a log-intensity model compared to log-ratio model. The amount of this recovery also depended on the relative amount of variability between slide, biological and residual variances.

The third study presented a novel method for the analysis of relative quantification RT-PCR data. Using linear mixed models, we accounted for hierarchical replication due to more than one gene assayed in each sample and multiple repeated measures in each sample-gene combination. Extensive simulations showed that the proposed model controlled the Type I error rate at the nominal level and yielded better power than traditional analysis methods.

The last chapter focused on jointly optimizing microarray and qRT-PCR validation experiments in order to maximize sensitivity while controlling the false discovery rate (FDR) within a two-stage testing framework. This optimization was based on partitioning a set of biological samples into two groups to be separately utilized for the microarray and validation steps. The results indicated that most of the samples (typically more than 60%) should be assigned to the microarray experiment. Even though the original optimization assumed independent genes, known variances, and uniform effect sizes, we showed that the results are valid for moderate departures from those assumptions. The problem of designing an adequate validation experiment, conditional on an existing microarray experiment was also studied. We found that if independent samples are used for the validation, a limited sample size and a liberal significance level (>0.15) could be used to properly control FDR after the second stage test.

To my daughter, Maria Eva

AKNOWLEDGEMENTS

In the first place, I would like to acknowledge my family, especially my parents, Pedro and Silvia, and my daughter, Maria, for their constant support and understanding. Despite the distance, Dad and Mom have always been close to me. Without their support and encouragement, it would have been hard to keep up. Maria is probably the most wonderful child that I could have dreamed of and her patience and creativity has contributed a lot to my life and provided inspiration to finish this program.

To the members of my committee: Drs Rob Tempelman, Guilherme Rosa, Paul Coussens, R.V. Ramamoorthi and Cathy Ernst. For the encouragement, constructive criticism and guidance that they provided. I would like to say thanks especially to Dr. Coussens for allowing me to work with part of his group and for providing data, advice, ideas and help whenever I needed them.

Guilherme and Rob definitely deserve a dedicated paragraph. From the beginning, Guilherme trusted me, and gave me space and freedom to propose and develop ideas, while at the same time he offered wise advise when he considered that I needed some help. And he was right every single time. But over all, I am very grateful to him for offering me his friendship and support in every single issue of my academic and personal life. Rob Tempelman "adopted me" towards the end of my program but treated me as if I were his student from the beginning. He was very considerate of the work done and accepted to keep the course of the research that I had started. His generosity and support have made the difference to me. Both, Rob and Guilherme, at this point, are not only academic role models for me, but they are also models as friends and parents. The things

I learned from them are so numerous, valuable and diverse that I will not attempt to list them here. Enormous thanks go to both of them.

The experience that I gained by collaborating with other fellow students and researchers is as valuable as what I learned though my own research. Among the people I'd like to thank especially for this are Rosangela Poletto, Mahmoud Romehi, Osman Patel, Nora Bello and Patricia Almeida. It has been a pleasure to work with them, learn from them and transfer to them part of my knowledge.

The fellow students and researchers in statistical genetics, Fernando Cardoso, Shirley (Lan Xiao) and Ashok Ragavendran were always ready to help with computational resources, share a good idea or get involved in a fruitful discussion. Their contribution has made a difference in more than one occasion.

Last, but not least, thanks to my friends. Especially to those who have been close to me during these last four years. Federico Aime, Pablo Ross, Nora Bello and Patricia de Almeida are among the most emblematic of the friends that I ever made. And I met them here, during my PhD program. These folks accepted me (and my circumstances) as a whole package. Their unconditional generosity and support have helped me during some very rough times of my life and I will always have a special place for them in my memories and in my heart.

"If the pages of this book contain some successful verse, the reader must excuse me the discourtesy of having usurped it first. Our nothingness differs little; it is a trivial and chance circumstance that you should be the reader of these exercises and I their author."

Jorge Luis Borges.

TABLE OF CONTENTS

LIST OF TABLES x
LIST OF FIGURES xii

INTRODUCTION 1
1. Experimental design and analysis of two-color microarray experiments. 2
2. Analysis of relative quantification qRT-PCR data. 5
3. Design of screening and validation experiments. 7
4. General hypothesis 8
5. Specific Aims 9

CHAPTER ONE 11
I. On reference designs for microarray experiments 12
1. Introduction 14
2. The traditional reference designs 15
2.1. General layouts 15
2.2. Required resources: arrays and mRNA samples 16
2.3. Linear models and contrast variance 17
3. An improved reference design 19
3.1. Layout 19
3.2. Required resources 20
3.3. Statistical efficiency 20
4. Results and discussion 22
4.1. Biological and technical replication 22
4.2. Relative efficiency of traditional reference designs 22
4.3. An alternative reference design 23
4.4. Technical and logistic issues 23
4.5. Concluding remarks 25

CHAPTER TWO 33
II. Comparison of log-intensity and log-ratio linear models for two color microarrays. 33
1. Introduction 34
2. Linearly transforming a log intensity model to a log ratio model. 35
3. Ratio models for specific designs 36
4. Recovery of inter-array information. 45
5. Discussion 49

CHAPTER THREE 60
III. Linear mixed models for the analysis of relative quantification RT-PCR data 60
1. Background 61
2. Results 64

2.1. Motivating example 64
2.2. Testing and estimating differential expression 66
2.3. Validation through simulation 67
2.4. Model checking and result comparison in experimental datasets 70
2.5. Comparison of results with $2^{-\Delta\Delta CT}$ methods 72
3. Discussion 73
4. Conclusions 76
5. Methods 76
5.1. Materials and RT-PCR reactions 76
5.2. Model derivation 77
5.3. Hypothesis testing and estimation 78
5.4. Data and response variable 79
5.5. Programs for analysis 79
5.6. Simulation study 79
5.7. Model selection 80

CHAPTER FOUR 88
IV. Design and analysis of two-stage experiments for transcription profiling 88
1. Introduction 89
2. Methods 90
2.1. Type I error rate in a two-stage single test. 91
2.2. FDR and sensitivity in two-stage tests 92
3. Results 94
3.1. Sample size and sample allocation 94
3.1.1. Known variance and constant effect sizes across independent genes 94
3.1.2. Distributed effect sizes 97
3.1.3. Unknown variances 98
3.1.4. Implementing two stage tests 98
3.1.5. Correlated genes 100
3.2. Design of a validation experiment 101
3.2.1. Sample size calculation of a (independent) validation experiment 101
3.2.2. Non-independent sample sets in screening and validation experiments 103
3.3. Estimation of effect sizes and correlation for microarray and qRT-PCR experiments 105
4. Discussion 107

CHAPTER FIVE 129
General discussion 129
1. Objectives revisited and their impact in animal functional genomics. 129
2. Future research directions 135

APPENDIX ONE 132

REFERENCES 147

LIST OF TABLES

Table 1.1. Number of biological replicates in non-control treatments (r), for a given number of arrays (n) and treatments (K), for each alternative reference design.....16

Table 1.2. Total number and specific variances of direct and indirect contrasts between treatments groups, and overall average variance of contrasts for replicated (RRD), classical (ClRD) and common (CRD) reference designs. 17

Table 1.3. Resources required in each of the four variants of reference design for comparing K treatments, with r biological replicates for the non-control groups. q: necessary aliquot of mRNA for each hybridization.....20

Table 2.1. Expected mean squares of relevant terms in models [4] and [5]. In both cases, the model [4] is assumed to be the data generation model. σ_e^2 Is the residual variance in the log-intensity scale, σ_B^2 is the individual variance in the log-intensity scale.....52

Table 2.2. Relative efficiency of intra-inter-slide (random array) analysis compare to the intra-slide (fixed array) analysis in the loop design (Figure 4). σ_e^2 Is the residual variance in the log-intensity scale, σ_B^2 is the individual variance in the log-intensity scale.....53

Table 3.1 Properties of the point estimates of the fold change.....81

Table 3.2. Properties of the interval estimates of the fold change.....82

Table 3.3. Properties of the hypothesis tests.....83

Table 3.4. Experimental data for model checking and result comparison.....84

Table 4.1 Outcomes of $m=m_0+m_1$ two-stage tests. A represents accepted hypotheses, R represents rejected hypotheses, the sub index 0 indicates true null hypotheses and 1 indicates true alternative hypotheses. The super index indicates the stage of the test.....112

TABLE 4.2 Error measures in two stage tests based on the quantities presented in Table 4.1.....113

Table 4.3. Sensitivity and FDR for two stage tests based on correlated test statistics. $\pi_0=0.99$, $\delta^{1)}=0.6$, $\delta^{1)}=0.72$, $n_1=n_2=25$.....114

LIST OF FIGURES

Figure 1.1: Alternative reference designs with three treatments (A, B and C) and 12 arrays: a) Common reference design (CRD) with four replicates (subindexes); reference sample (R) is the same in all arrays; b) Classical reference design (ClRD) with four replicates (subindexes) in 12 arrays; reference sample (R) is replicated; and c) Replicated reference design (RRD), in which six replications of treatments B and C are hybridized together with independent replicates of a control treatment (A). 15

Figure 1.2. Relative efficiency of the replicated reference design (RRD) as compared to the common reference design (CRD), considering: a) the same number of replicates (r), and b) the same number of arrays (n). X-axis is the log2 of the variance ratio; curves correspond to different numbers of treatments (K).....18

Figure 1.3. Efficiency of the: a) common reference design (CRD), and the b) replicated reference design (RRD), relative to the classical reference design (ClRD). ●: fixed number of biological replicates, ■: fixed number of arrays.....19

Figure 1.4. Blocked reference design (BRD) with 3 treatments (A, B and C) and 6 replicates (subindexes) in 12 arrays. Each test sample is hybridized together with a replicate from the control treatment (A); the same set of replicates of the control group is used with the other treatments. 19

Figure 1.5. Relative efficiency of the blocked reference design (BRD) compared to: a) the replicated reference design (RRD), and to b) the common reference design (CRD). A fixed number of replicates (r) is considered. X-axis is the log2 of the variance ratio. Each curve corresponds to a different number of treatments (K). 21

Figure 2.1. Common reference design. B_{kl} is the l^{th} biological replicate within the k^{th} group. 54

Figure 2.2. Complete block design with dye swap . B_{kl} is the l^{th} biological replicate within the k^{th} group. 55

Figure 2.3. Dye swap design with technical replication. B_{ij} is the j^{th} biological replicate within the i^{th} group.....56

Figure 2.4. Connected loop design. B_{kl} is the l^{th} biological replicate within the k^{th} group 57

Figure 2.5. Split-plot design. B_{kls} is the l^{th} biological replicate within the k^{th} group of factor T receiving the s^{th} level of factor S. The same experimental unit ($\boldsymbol{B}_{ij}$) is assigned to two levels of the factor S (subplot factor) in the same array. Independent experimental units are used for the two levels of the factor T (whole plot factor).....58

Figure 2.6. Relative contrast variance under different assumptions for the array effect in reference designs.....59

Figure 3.1.Fold change estimates.....85

Figure 3.2.Histograms of p-values under null hypothesis. 86

Figure 3.3.Result comparison between ΔCT (vertical axis) and linear mixed model (horizontal axis). 87

Figure 4.1. a) Sample allocation and b) sensitivity of optimal designs as a function of the relative effect size between the two stages. The lines represent different proportions (0.005, 0.01, 0.015, 0.02 and 0 =1 stage) of selected hypotheses to be carried to the second stage. The results were calculated for: π_0=0.99, f=0.05, $\delta_j^{(1)}$=0.4, and n=170 (see text). 115

Figure 4.2. Distribution of the effect size for stage 1. a) constant, b) symmetric. c) decreasing. The three distributions were scaled to have mean 1.0 or 0.4.....116

Figure 4.3. Sample size allocation and sensitivity as function of the distribution of the first and second stage effect size, average $\delta^{(1)}$=1, n=28. π_0=0.99, f=0.05. The single stage sensitivity corresponds to a decaying distribution.....117

Figure 4.4. Sample size allocation and sensitivity as function of the distribution of the first stage effect size, average $\delta_j^{(1)}$=0.4, n=170 (c,d). π_0=0.99, f=0.05. The single stage sensitivity corresponds to a decaying distribution. 118

Figure 4.5. Distribution of the ratio of the effect sizes. a) constant, b) symmetric. c) decreasing. 119

Figure 4.6. Sample size allocation (a) and sensitivity (b) as function of the distribution of the first stage effect size for different effect size rations (lines). π_0=0.99, f=0.05..... 120

Figure 4.7. Sample allocation (a) and sensitivity (b) of optimal designs as a function of the relative effect size. π_0=0.99, f=0.05, $\delta_j^{(1)}$=1, n=30..... 121

Figure 4.7. Distribution of the sensitivity and false discovery rate obtained from simulations based on different procedures to select the significance cutoff ($\alpha^{(1)}$). The points indicate the mean whereas the line is the expected value obtained from the optimization algorithm..... 122

Figure 4.8. Distribution of a) sensitivity and b) false discovery rate obtained from simulations for different correlation structure in the data. The x-axis numbers indicate the correlation coefficient whereas the subindex on the x-axis values indicates the block size. The points within each box indicate the mean whereas the line reflects the expected value obtained from the optimization algorithm assuming independence. 123

Figure 4.9. Sensitivity of two-step tests for correlated hypotheses as function of the sample size in the first step. N=200 simulation with π_0=0.99, f=0.05, $\delta_j^{(1)}$=1, n=30, $R^{(1)}/m$=0.01. The arrow indicates the optimal sample allocation obtained assuming uncorrelated tests. •: ρ=0.4, block size=40, ■: ρ=0.4, block size=10. 124

Figure 4.10 Minimum number of replicates per treatment in a validation experiment as function of proportion of null hypotheses ($FDR^{(1)}$) and effect sizes. $R^{(1)}$=25 genes putatively differentially expressed at different levels of FDR , target sensitivity = 0.9, FDR after validation = 0.05. 125

FIGURE 4.11. Power of a single two-test for different correlations (0, 0.56, and 0.66) between the first and second stage data as a function the ratio of second stage to first stage effect sizes. The effect size of the first stage ($\delta^{(1)}$) is 0.6 whereas the number of replicates per treatment is 25 for each test..... 126

Figure 4.12. Box-Plot of the microarray and qRT-PCR based effect size estimates and their ratios. 127

Figure 4.13. Histogram of the estimated correlation coefficient between microarray and qRTPCR measured expressions. 128

INTRODUCTION

Transcriptional profiling is a key aspect of functional genomics as it can provide invaluable insights into the function of genes by elucidating their pattern of expression or regulation. Drawing meaningful inferences from these studies given resource and funding constraints depends on optimizing the experimental designs and statistical analysis models and methods according to restrictions imposed by the currently available technologies. The most commonly used techniques for gene expression studies are microarrays and quantitative reverse transcription polymerase chain reaction (qRT-PCR).

A two-color microarray, such as a cDNA or long-oligonucleotide microarray, can measure the relative expression of thousands of genes simultaneously between a pair of samples. This expression profiling technology seems particularly useful for individuals studying a wide range of less well represented organisms in their research programs and thus has become the platforms of choice for high throughput expression profiling in livestock functional genomics. Some of the challenges using these two color arrays pertain to experimental design (sample size and optimal sample allocation for a given experimental setup) and statistical analyses.

In general, the statistical issues of a microarray experiment include the following items (Allison et al., 2006): experimental design, preprocessing, inference and validation. Many statisticians have recently played a key role in the development of experimental designs and statistical methods for microarray experiments. For example, mixed models that account for multiple sources of variation in a statistical analysis have greatly improved the inferential power of two color microarrays (Cui and Churchill, 2003).

However, in some areas, statistical developments are still sorely lacking. For example, there has been no extensive work on acceptable statistical practice for qRT-PCR validation of microarray results, even though such validation is encouraged or even required by many journals.

Now qRT-PCR is a gene expression profiling technique that is usually assumed to be more accurate and precise than microarrays (largely due to increased dynamic range) and is commonly used as a validation tool as aforementioned. The characteristics of qRT-PCR assays make them more suited to simultaneously test a lower number of genes from a larger set of samples, and consequently are more limited as a high throughput screening tool compared to microarrays. Nevertheless, as with microarrays, qRT-PCR data presents interesting challenges for inferring upon differential gene expression, including aspects of choosing appropriate experimental designs and declaring statistical significance on differential expression, particularly when this methodology is used for validation of microarray results.

1. Experimental design and analysis of two-color microarray experiments.

For two-color or (two-channel) microarrays, the experimental design characterizes how particular samples are paired within each slide and the total number of biological replicates used in the experiment. A two-color microarray experiment actually includes two blocking factors. One factor is the array (or slide) that can accommodate two samples (i.e. one per channel) whereas the other factor is the dye labeling that typically involves assignment of one of two states or colors to each sample within an array. Given the restriction of allowing a comparison of only two treatments within

slides for a two channel array, a number of incomplete block (e.g. loop) designs have been proposed when several treatments are of interest (Dobbin and Simon, 2002; Kerr and Churchill, 2001).

Despite it not being the most efficient design (Dobbin and Simon, 2002; Kerr and Churchill, 2001; Tempelman, 2005), the reference design is still pervasively used for two color microarray studies. In a common reference design, aliquots from a uniform single reference sample are always hybridized with a different sample from each experimental group on each array. There are several possible deviations on this design, e.g. biological vs. technical replicates from the experimental groups and the labeling of reference with one or both dyes across arrays. Now whereas a common reference design consistently uses a single sample as a reference, a classical reference design uses different biological replicates from the reference group across all arrays (Tempelman, 2005). While the reference is not of interest per se, the efficiency might be improved by using the reference as one of the treatments of interest in some settings. Under those circumstances with the correct sample allocation to arrays, the reference design has the potential to be a very powerful and flexible experimental layout. Several authors have compared reference designs to other alternatives (Dobbin and Simon, 2002; Kerr, 2003a; Kerr and Churchill, 2001; Tempelman, 2005; Yang and Speed, 2002) in two color systems. However, competing deviations on reference designs have not been considered exhaustively. Furthermore (Kerr and Churchill, 2001; Yang and Speed, 2002), researchers often do not make the distinction between different levels of variability that arise in these designs. Appropriately accounting for such hierarchical replication in two color microarray experiments has been shown to be crucial to obtain appropriate inferences (Rosa et al.,

2005).

There are several experimental situations that can benefit from the improvement of reference designs. One such situation is a screening experiment whereby several treatments are compared to a control upon which significantly expressed genes are selected for further study. Likewise, if the number of treatments or classes to be compared is not known a priori, the reference design is extendible provided there is a sufficiently large source of reference sample to be repeatedly drawn from for future hybridizations with new experimental samples.

Typically, fluorescence intensity data (e.g. Cy3 and Cy5) from reference designs is further reduced to a ratio of these two values for each spot on an array before subsequent statistical analysis. The common reference design lends itself, in particular, to the quotient of fluorescence intensities between the "unknown" sample and the reference sample. In general, these ratios are subsequently logarithmically transformed to facilitate data normalization and distributional normality assumptions as required for parametric statistical methods; furthermore, ratios are easily compared across arrays. Analyses based on log-ratios are simpler for designs with one level of replication (such as a common reference design); however, log-ratio models are not easy to generalize to more complex experimental designs with hierarchical or technical levels of replication. Furthermore, even for a reference design, the use of ratios may lead to the loss of statistical information, and this should be investigated further.

Linear fixed and mixed effect models based on the analysis of log intensities have recently gained popularity (Kerr et al., 2000; Rosa et al., 2005; Smyth, 2004; Wolfinger et al., 2001). These models offer a flexible way to account for multiple sources of

variation that arise from hierarchically replicated microarray experiments. Log-ratio and log-intensity linear models have been compared only for simple experimental designs (Kerr, 2003b). Moreover, there is very limited research on how to implement linear mixed models for log-ratio data in order to account for multiple layers of replication (Smyth et al., 2005).

Given the wider availability of statistical analysis software for microarray experiments that usually accommodate only one or the other of the two different expression measures (log-ratios or log-intensities), a better understanding of the circumstances under which the two models will yield the same results and the limitations of each analysis would provide invaluable insights into the interpretations and comparison of results from independent investigations. Also, a better understanding of the extent of the loss of information incurred when resorting to a simpler analysis strategy would be important when evaluating the ramifications of such a strategy in a particular experiment. A general computational framework to assess these issues would contribute to a better understanding of the similarity and differences of alternative analysis strategies.

2. Analysis of relative quantification of RT-PCR data.

The primary response variable for gene expression from qRT-PCR consists of the fractional cycle number (C_T) at which threshold intensity is reached and is directly proportional to the negative logarithm of the mRNA concentration. The corresponding constant of proportionality may be estimated from a standard curve thereby providing absolute quantification of the mRNA present in a sample. For most comparative

transcriptional profiling experiments, however, the relative level between samples rather than actual abundance is sufficiently informative for comparing several treatments.

Relative quantification aims at comparing relative levels of target mRNA across samples by standardizing the total RNA quantity to an internal control gene assumed to be constant in its expression across different experimental conditions (Pfaffl, 2001). The most common summary statistic for this type of quantification is the ΔΔCT method (Livak and Schmittgen, 2001). This measure is easily computed for simple experimental designs where several groups are compared to a single control group. A series of ad-hoc parametric or non-parametric approaches (Pfaffl et al., 2002) are used to obtain p-values and standard errors associated with the treatment comparisons of interest using the ΔΔCT measure. Nonetheless, more general statistical models have not been specified to analyze this measure in more efficient and elaborate designs.

In more complex experimental designs, such as those that commonly appear in animal functional genomics, hypotheses that are more complex than all pairwise comparisons are likely to be of interest. For example, two-factor interaction hypotheses, time trend contrasts and hypotheses related to the significance of variance components are increasingly common. However, there has been limited formal development of statistical models and methods adopted for these questions using relative quantification of RT-PCR data. (Fu et al., 2006)

Linear mixed models offer a promising approach to the statistical analysis of qRT-PCR data as they have had with microarray data. Even though linear models have been used for qRT-PCR data in the past, these uses have generally been restricted to control gene selection (Szabo et al., 2004) or they do not consider multiple random

sources of variability (Cook et al., 2004). A general formulation of a linear mixed model for relative quantification of RT-PCR data will provide a flexible and powerful analysis methodology to accommodate arbitrarily complex designs and allow researchers to draw accurate inferences from hierarchically replicated experiments.

3. Design of screening and validation experiments.

A general consensus has been established on the necessity to independently validate results from microarray experiments (Allison et al., 2006) although recently the utility of such validation has been questioned (Rockett, 2003; Rockett and Hellmann, 2004). Most genomicists will attempt to validate results from a microarray assay using qRT-PCR, inmunohistochemistry or protein analysis techniques. Such validation is typically conducted for only a small fraction of the genes declared to be differentially expressed by a microarray experiment. The most common practice consists of selecting a few genes for qRT-PCR using mRNA from the same samples included in the microarray experiment. This practice of using the same samples again has been criticized because it provides only "technical validation" and does not preclude the possibility that a few unusual samples might have been selected by chance for the study (Allison et al., 2006), thereby biasing statistical tests for both microarrays and qRT-PCR in the same direction. Even though a better recommendation would to use different samples for validation by qRT-PCR that are completely independent from the samples used in the microarray experiment, there are yet no clear guidelines on how to assess appropriate sample size calculations and multiple testing considerations for a second stage validation.

Multiple testing considerations for controlling false positive rates are generally

ignored in qRT-PCR validation experiments because of the limited number of tests (i.e. selected genes) considered in the validation step. However, qRT-PCR is being increasingly automated such that there is now more research where several dozen genes are assayed simultaneously with these techniques (Perreard et al., 2006; Szabo et al., 2004). With many simultaneous multiple tests, improper attention to experiment wise error rates could yield an unacceptably large number of false positive results even in the validation stage. For example, the conventional comparison wise Type I error rates that were eventually realized to be too liberal for gene-specific tests in large microarray experiments might likewise not be suitable for a validation experiment. For a validation experiment using qRT-PCR, controlling the false discovery rate may be more useful to control for the rate of false positives without unduly sacrificing sensitivity as has been demonstrated for microarray experiments (Verhoeven et al., 2005). In other words, the decision rules for validating genes as being differentially expressed using qRT-PCR need to be more carefully studied. There is also a need to evaluate experimental designs for validation experiments and to provide a framework for helping to guide researchers design initial screening (microarray) and subsequent validation (qRT-PCR) experiments.

4. General hypothesis

The efficiency and power of comparative transcriptional profiling can be increased by further optimizing experimental designs and statistical analysis models and methods for microarray and validation (qRT-PCR) studies.

5. Specific Aims

This dissertation intended to propose more efficient experimental designs and statistical analysis models and methods for expression profiling experiments. In particular, linear fixed and mixed effects models are developed for the analysis of both microarray and qRT-PCR data, and attention is particularly directed to the modeling of multiple levels of variation.

The questions that will be addressed are of interest to experimental genomicists, and the proposed developments will be potentially useful not only to researchers working in functional genomics research with animals, but also with other model organisms. More efficient designs for the microarray experiments coupled with more powerful second-stage designs and analysis for the validation experiments will facilitate greater sensitivity and specificity for discovering genes that are differentially expressed between treatments of interest.

The overall aim of this dissertation was to investigate optimization possibilities for the design and statistical analysis of expression profiling experiments, including the two stage process of screening and further validating differentially expressed genes using linear model methods.

The specific objectives were:

1) To compare alternative reference designs for statistical efficiency of two color microarray experiments considering multiples sources of variation.
2) To investigate the ramifications of log-ratio versus log-intensity modeling in two color microarrays using linear mixed effects models
3) To develop linear mixed models for the analysis of relative quantification of RT-

PCR data.

4) To propose a general framework for determining the false discovery rate and sensitivity of gene expression studies when jointly designing microarray screening experiments linked to subsequent and selective qRT-PCR validation.

CHAPTER ONE

Juan P. Steibel and Guilherme J.M. Rosa. 2005. On reference designs for microarray experiments. Statistical Applications in Genetics and Molecular Biology. (4) 1:A36.

Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 36

On Reference Designs For Microarray Experiments

Juan P. Steibel* Guilherme J. M. Rosa†

*Department of Animal Science, Michigan State University, steibelj@msu.edu

†Department of Animal Science and Department of Fisheries & Wildlife, Michigan State University, rosag@msu.edu

On Reference Designs For Microarray Experiments*

Juan P. Steibel and Guilherme J. M. Rosa

Abstract

We compare four variants of the reference design for microarray experiments in terms of their relative efficiency. A common reference sample across arrays is the most extensively used variation in practice, but independent samples from a reference group have also been considered in previous works. The relative efficiency of these designs depends of the number of treatments and the ratio between biological and technical variances. Here, we propose another alternative of reference structure, denoted by blocked reference design (BRD), in which each set (replication) of the treated samples is co-hybridized to an independent experimental unit of the control (reference) group. We provide efficiency curves for each pair of designs under different scenarios of variance ratio and number of treatments groups. The results show that the BRD is more efficient and less expensive than the traditional reference designs. Among the situations where the BRD is likely to be preferable we list time course experiments with a baseline and drug experiments with a placebo group.

KEYWORDS: Microarrays, Experimental design, Reference design, Relative efficiency

*This project was supported in part by National Research Initiative Grant no. 2004-35604-14580 from the USDA Cooperative State Research, Education, and Extension Service.

1 INTRODUCTION

The reference design is one of the simplest and most commonly used designs in microarray experiments (Churchill, 2002). Despite being generally less powerful than other alternatives such as the loop design (Dobbin and Simon, 2002; Kerr, 2003a; Kerr and Churchill, 2001; Yang and Speed, 2002), the reference design presents some advantages (Dobbin and Simon, 2002), as it simplifies the statistical analysis and facilitates the comparison of results from different experiments within a meta-analysis context.

The reference design consists of the hybridization of each test sample with a common type of sample (Kim et al., 2002). In the statistical literature on design of microarray experiments, however, the term "reference design" is used for two different experimental layouts. Kerr and Churchill (2001) and Glonek and Solomon (2004) implicitly use a "replicated reference", where the reference sample includes replication at the same level as the treatment samples. Hereinafter we refer to this design as "classical reference design" (ClRD). Conversely, Dobbin and Simon (2002) consider a common reference sample ("common reference design", CRD), which refers to a single (biological) sample that is co-hybridized to each of the test samples. Both variants of reference design are used in practice but the CRD seems to be preferred (Alizadeh et al., 2000; Lin et al., 2002; Papp et al., 2003; Perou et al., 2000).

In general, in the CRD and ClRD the reference group is not of interest per-se. Nevertheless, the reference group can be represented by one of the treatments of interest, such as the initial time in a time course experiment (Yang and Speed, 2002) or a wild type strain (Wolfinger et al., 2001). To differentiate this design from the ClRD described above, which also presents biological replication for the reference group, we will denote it by replicated reference design (RRD). The aim of this paper is two fold: first to compare these alternative designs from both statistical and technical points of view, and second to propose a new variant of the reference design, which is shown to present higher power and lower cost than the traditional alternatives.

The paper is structured as follows: in Section 2, the CRD, ClRD and RRD are compared to each other in terms of total amount of resources required and their average variance of contrasts. In Section 3, a new reference design alternative is proposed and compared to the traditional reference designs (CRD, ClRD and RRD). In Section 4, the results are discussed and general recommendations on the use of each design are provided.

2 THE TRADITIONAL REFERENCE DESIGNS

2.1 GENERAL LAYOUTS

Schematic representations of CRD, ClRD and RRD experiments are provided in Figure 1 (Panels a, b and c, respectively). The experiments involve three treatments (A, B and C) and 12 arrays; the indexes represent biological replication within each treatment group. The CRD uses a single reference sample (denoted by R) without biological replication on it, which is co-hybridized with each of the four biological replications of each treatment group. Conversely, the ClRD uses an independent reference sample in each array. The RRD, on the other hand, considers one of the treatments as a reference (the group A in Figure 1c), from which a different biological replication is hybridized in each of the 12 arrays. In this case, group A has then 12 biological replications whereas the remaining treatment groups have six replications each. In the CRD and ClRD, any gene specific dye bias can be confounded with the reference sample effect, as it will cancel-out in the indirect comparisons, therefore the dye swap approach is not mandatory. The RRD however, should include both labeling directions so dye and treatment effects can be disentangled.

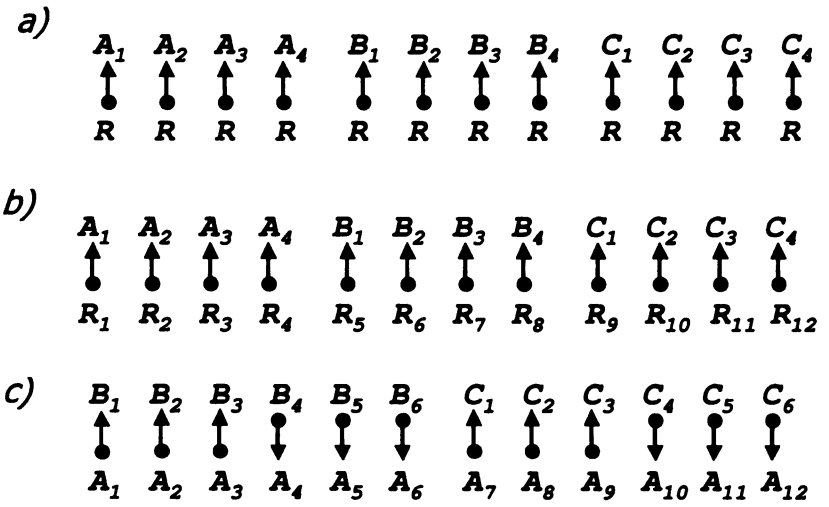


Figure 1: Alternative reference designs with three treatments (A, B and C) and 12 arrays: a) Common reference design (CRD) with four replicates (subindexes); reference sample (R) is the same in all arrays; b) Classical reference design (ClRD) with four replicates (subindexes) in 12 arrays; reference sample (R) is replicated; and c) Replicated reference design (RRD), in which six replications of treatments B and C are hybridized together with independent replicates of a control treatment (A).

2.2 Required resources: arrays and mRNA samples

In terms of resources needed for running a specific experiment, we can compare the three designs by fixing either the total number of arrays or the number of replications in each treatment group. Specifically, consider a CRD or ClRD with K treatments and r biological replicates per treatment, such that $r = c(K-1)$, where c is an arbitrary integer value considered here to constrain the design space to balanced cases only. In this case, $n = rK$ arrays are necessary for running the experiment. On the other hand, it is seen that a RRD can accommodate $r^* = cK = r+r/(K-1)$ biological samples with the same number of arrays (Table 1; compare also Figures 1a and 1b). Likewise, fixing the number of non-reference samples, only $n= r(K-1)$ arrays will be needed in a RRD as compared to rK arrays in the CRD or ClRD.

Table 1. Number of biological replicates in non-control treatments (r), for a given number of arrays (n) and treatments (K), for each alternative reference design.

		Design		
K	n	CRD	ClRD	RRD
2	6	3	3	6
	10	5	5	10
	20	10	10	20
3	12	4	4	6
	24	8	8	12
4	12	3	3	4
	24	6	6	8
6	30	5	5	6

2.3 Linear models and contrast variance

The analysis of log-ratio of Cy3 to Cy5 intensities is considered here through the following general gene-specific linear model:

$$\boldsymbol{r}_g=\boldsymbol{X}\boldsymbol{\beta}_g+\boldsymbol{\varepsilon}_g, \qquad [1]$$

where $\boldsymbol{r}_g$ is a vector of log ratios (log Cy3 – log Cy5) for gene g in each array, after suitable data normalization (Lonnstedt and Speed, 2002). Here we assume a single spot per gene on each array, but in the case of multiple spots, the elements of $\boldsymbol{r}_g$ may represent the average (or median) normalized log-ratio from all the

spots corresponding to gene g in each array. In addition, $\boldsymbol{X}$ is an incidence matrix (see example and details in the Appendix); and $\boldsymbol{\beta}_g$ is a vector of linear coefficients, taken to be $\boldsymbol{\beta}_g=[\tau_{(1-R)} \ldots \tau_{(K-R)}]^T$ or $\boldsymbol{\beta}_g=[\alpha\ \tau_{(2-1)} \ldots \tau_{(K-1)}]^T$ for the CRD (or ClRD) and RRD, respectively, where $\tau_{(i-j)}=\mu_i-\mu_j$ represents the contrast between treatments i and j (i,j= 1,2,...,K and R), and α is a dye bias term. The elements of the vector $\boldsymbol{\varepsilon}_g$ are assumed independently normally distributed with variances σ_ε^2 for the CRD, and $\sigma_{\varepsilon'}^2$ for ClRD and RRD cases. Assuming that the variance components on the log intensity scale are σ_a^2 (biological variance) and σ_e^2 (residual variance), it is seen that $\sigma_\varepsilon^2 = \sigma_a^2 + 2\sigma_e^2$ and $\sigma_{\varepsilon'}^2 = 2\sigma_a^2 + 2\sigma_e^2$ (please refer to Appendix for details).

While the CRD and the ClRD have only indirect comparisons between treatment groups, the RRD comprises K-1 direct and $(K-1)(K-2)/2$ indirect comparisons. As each kind of contrast has a different variance, we compare the designs using the A-optimality criterion discussed in Kerr and Churchill (2001). According to this criterion, the average variance of all contrasts is computed considering all the comparisons equally relevant. If this is not the case, the use of A-optimality criterion may not be appropriate for comparing experimental designs (Yang and Speed, 2003). The variances of each kind of contrast (direct and indirect), as well as the overall average variances are presented in Table 2.

Table 2. Total number and specific variances of direct and indirect contrasts between treatments groups, and overall average variance of contrasts for replicated (RRD), classical (ClRD) and common (CRD) reference designs.

	RRD	ClRD	CRD
Number of Direct Comparisons	$(K-1)$	0	0
Var[direct comparison]	$\frac{2(\sigma_a^2+\sigma_e^2)}{r}$	-	-
Number of Indirect Comparisons	$\frac{(K-2)(K-1)}{2}$	$\frac{K(K-1)}{2}$	$\frac{K(K-1)}{2}$
Var[indirect comparison]	$\frac{4(\sigma_a^2+\sigma_e^2)}{r}$	$\frac{4(\sigma_a^2+\sigma_e^2)}{r}$	$\frac{2(\sigma_a^2+2\sigma_e^2)}{r}$
Overall Average Variance	$\frac{(K-1)4(\sigma_a^2+\sigma_e^2)}{Kr}$	$\frac{4(\sigma_a^2+\sigma_e^2)}{r}$	$\frac{2(\sigma_a^2+2\sigma_e^2)}{r}$

The efficiency of the RRD relative to the CRD may be computed using the following expression:

$$E_{RRD:CRD} = \frac{\overline{Var}(CRD)}{\overline{Var}(RRD)}, \quad [2]$$

where $\overline{Var}(\cdot)$ is the overall average variance of the contrasts in each design. Similar equations can be used to compare any pair of designs.

For the designs presented so far, the relative efficiency is a function of the number of treatments (K) and (or) the biological-to-technical variance ratio (σ_a^2 / σ_e^2), as discussed by Dobbin and Simon (2002). Figures 2a and 2b present the relative efficiency as a function of the variance ratio for a given number of biological replicates (r) or number of arrays (n), respectively. The comparison for fixed number of arrays is restricted to cases in which $r=c(K-1)$, as discussed before.

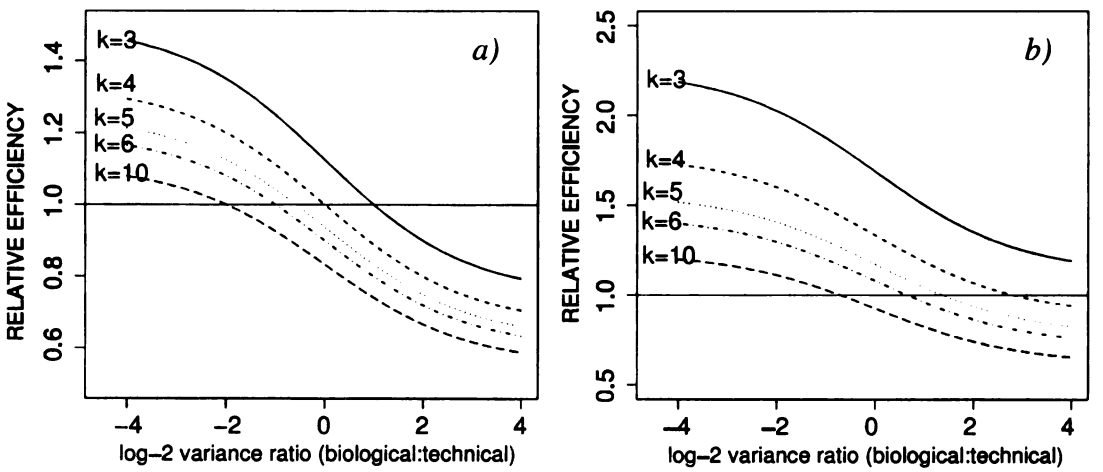


Figure 2. Relative efficiency of the replicated reference design (RRD) as compared to the common reference design (CRD), considering: a) the same number of replicates (r), and b) the same number of arrays (n). X-axis is the log2 of the variance ratio; curves correspond to different numbers of treatments (K).

The relative efficiency of the CRD as compared to the ClRD is a function of the variance ratio only (Figure 3a), and the relative efficiency of the RRD to the ClRD depends only on the number of treatments. Similar results for the relative comparison between CRD and ClRD were discussed by Tempelman (2005).

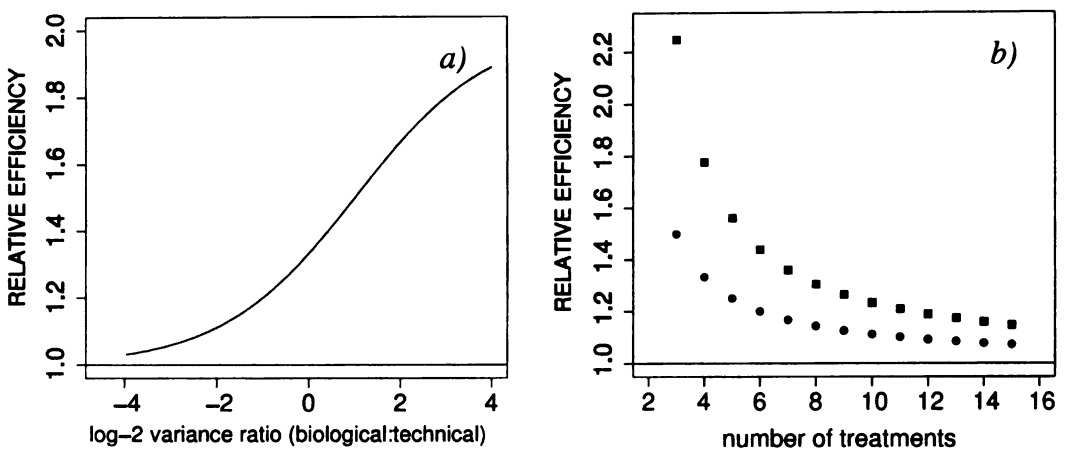


Figure 3. Efficiency of the: a) common reference design (CRD), and the b) replicated reference design (RRD), relative to the classical reference design (ClRD). ●: fixed number of biological replicates, ■: fixed number of arrays.

3 AN IMPROVED REFERENCE DESIGN

3.1 LAYOUT

An alternative layout for the reference design structures discussed so far is represented in Figure 4, which will be denoted hereinafter by "blocked reference design" (BRD). In this case, a reference treatment is used similarly to the RRD but with a set of r biological samples hybridized with every treatment. Here we deliberately consider a single dye labeling for each subject in the reference group (treatment A) but with a dye balanced across subjects within each treatment. Note that for K=2, the BRD is the same as the RRD, which in turn are balanced complete block designs.

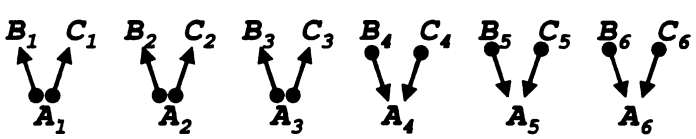


Figure 4. Blocked reference design (BRD) with 3 treatments (A, B and C) and 6 replicates (subindexes) in 12 arrays. Each test sample is hybridized together with a replicate from the control treatment (A); the same set of replicates of the control group is used with the other treatments.

3.2 Required resources

Technically, all choices of reference designs (CRD, ClRD, RRD and BRD) can be compared in terms of the number of replications or arrays required, as well as the sample quantity (mRNA quantity) necessary for a given number of biological replication (Table 3).

Table 3. Resources required in each of the four variants of reference design for comparing K treatments, with r biological replicates for the non-control groups. q: necessary aliquot of mRNA for each hybridization.

	CRD	ClRD	RRD	BRD
Number of non-ref. samples	r	r	r	r
Number of arrays	rK	rK	$r(K-1)$	$r(K-1)$
Number of ref. samples	1	rK	$r(K-1)$	r
mRNA quantity of non-ref.	q	q	q	q
mRNA quantity of ref.	rKq	q	q	$(K-1)q$

3.3 Statistical efficiency

Model [1] with correlated residuals (see details in the Appendix) can be used for the statistical analysis of the BRD. In this case, the residuals from log-ratios taken with respect to the same reference sample will have the following covariance structure:

$$\Sigma_{(K-1)\times(K-1)} = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_a^2 & \cdots & \sigma_a^2 \\ \sigma_a^2 & \sigma_\varepsilon^2 & \cdots & \sigma_a^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_a^2 & \sigma_a^2 & \cdots & \sigma_\varepsilon^2 \end{bmatrix}, \qquad [3]$$

where $\sigma_\varepsilon^2 = 2\sigma_a^2+2\sigma_e^2$. The variance of an indirect contrast will be $2(\sigma_a^2+2\sigma_e^2)/r$, whereas the variance of a direct comparison will be $2(\sigma_a^2+\sigma_e^2)/r$. The BRD has the same number of direct and indirect comparisons, as well as the same variance for the direct contrasts, as compared to the RRD. The variance of the indirect contrasts, however, is equal to the variance in a CRD. Consequently, fixing the number of biological replicates (of the non reference samples) in a specific experiment, the variance of any contrast under BRD will be equal to or smaller

than the variance under either of the other designs. The average variance of contrasts in a BRD is:

$$\overline{Var}(BRD) = \frac{2}{r}\left[\sigma_a^2 + \frac{2}{K}(K-1)\sigma_e^2\right], \qquad [4]$$

and the efficiency of this design is the highest under any variance ratio scenario (Figures 5a and 5b). In addition, as the BRD uses fewer biological replicates of the reference sample than the RRD and fewer arrays than the CRD, its cost will be always the smallest among the three variants.

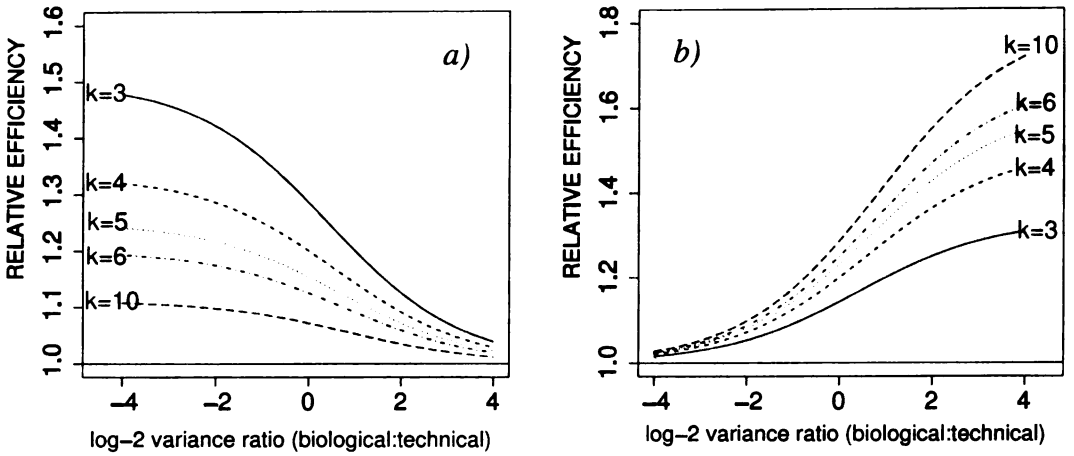


Figure 5. Relative efficiency of the blocked reference design (BRD) compared to: a) the replicated reference design (RRD), and to b) the common reference design (CRD). A fixed number of replicates (r) is considered. X-axis is the log2 of the variance ratio. Each curve corresponds to a different number of treatments (K).

Similarly to the RRD, the BRD does not spend half of the hybridizations with a group that is not of direct interest. This allows incorporating more replicates in a BRD as compared to the CRD, for a given number of arrays. Consequently, the advantage in relative efficiency of the BRD as compared to the CRD is even larger than what is shown in Figure 5b, for a fixed number of replicates (r). Relatively to the RRD, the efficiency curves presented in Figure 5a for the BRD are the same if the comparisons are made at fixed number of replicates or fixed number of arrays.

4 RESULTS AND DISCUSSION

4.1 BIOLOGICAL AND TECHNICAL REPLICATION

This paper compares four variants of the reference design for microarray experiments, considering both the technical and the biological variability. We also distinguish between designs where the reference is not of interest per se (CRD and ClRD) and designs where the reference is one of the groups under comparison (RRD and BRD).

Some of these designs have been compared in previous works. Yang and Speed (2002) described the RRD, in which one treatment is used as the reference, and compared it to a CRD for time-course experiments. The authors, however, did not discuss the issue of biological and technical replication. Even though the error variance in their work is always directly comparable to $\sigma^2_{\varepsilon'}$ discussed here for the RRD, the error variance for their CRD is not so unless we assume zero biological variance in both designs. In a later paper, Yang and Speed (2003) incorporated the effect of different levels of replications in design comparisons, but the relative efficiency of the RRD to the CRD was not reassessed by them. Tempelman (2005) compared the CRD and ClRD in terms of statistical power and efficiency, at different levels of technical and biological variability. Our paper is more general than the previous ones, as it compares additional reference designs and makes also a clear distinction between biological and technical variability in the analyses.

The methodological approach of this paper consists of analyzing the log-ratio of Cy3 and Cy5 using a linear model parameterized in terms of the variances of log intensities. Such model is equivalent to the intensity model (Kerr, 2003b) if arrays are considered as fixed effects. Our procedure may also be considered a variant of the one presented by Yang and Speed (2003), where the variance and covariances for the log-ratios are expressed in terms of the variance components of the intensity model. This parameterization allows a general and straightforward comparison of the variants of reference design for different biological-to-technical variance ratios and number of treatments.

4.2 RELATIVE EFFICIENCY OF TRADITIONAL REFERENCE DESIGNS

Figure 3 shows that the ClRD is the least efficient of the designs. It will be equivalent to the CRD only in the hypothetical situation of null biological variance. Likewise, the ClRD is also less efficient than the RRD, and will tend to be as efficient as the RRD only if the number of treatments is very large. Consequently, from a statistical standpoint, it makes little sense to use the classical reference design. Similarly to the ClRD, the RRD also uses a replicated

reference, but such a reference is one of the treatments of interest. This strategy increases its overall efficiency and allows for an increase of sample size (more replicates for a fixed number of arrays).

Figure 2 shows the tendency of the RRD to be more efficient than the CRD for lower variance ratios (smaller biological variance relative to technical variance) and smaller number of treatments. With four treatments, if the number of non-control replicates is held fixed (Figure 2a), the relative efficiency tends to an asymptotic value (for $\sigma_a^2=0$) of 1.33, which equals to the value (2.00/1.50) reported in Table 2 of Yang and Speed (2002). Here, we also performed the comparisons by holding fixed the number of arrays (Figure 2b). For instance, with three treatments, the RRD always outperforms the CRD, while $\sigma_a^2 < \sigma_e^2$ guarantees relative efficiency above one for up to six treatments.

4.3 AN ALTERNATIVE REFERENCE DESIGN

In this paper, we propose a new variant of the reference design (denoted here as BRD) and compare it to the traditional reference designs commonly used in the literature. We show that the BRD is more efficient and less expensive than the other three variants of reference design, for experiments involving three or more treatment groups. As shown in Figures 4 and 5, the average variance of the contrasts for the BRD is always the lowest among the four alternatives. For lower biological variances, the differences in efficiency between the BRD and RRD vanish, whereas at higher biological variance the BRD and the CRD tend to be equivalent. Consequently, for any microarray experiment the BRD may have similar efficiency to the RRD for some genes and to the CRD for others, but on average the BRD will always outperform both designs. As compared to the ClRD, the BRD will always be more efficient, as the ClRD is outperformed by the CRD and the RRD as well.

4.4 TECHNICAL AND LOGISTIC ISSUES

As far as cost is concerned, at a fixed number of biological replicates, the BRD uses fewer arrays as compared to the CRD or ClRD and fewer biological replicates for the control treatment than the RRD. Consequently, the BRD is always the least expensive among the three alternatives.

In spite of the reference design being less efficient than other designs under certain circumstances (Bueno Filho et al., 2005; Dobbin and Simon, 2002; Kerr, 2003a; Kerr and Churchill, 2001; Yang and Speed, 2002), it is still advocated by some authors, especially in the applied genomic science. For example, the CRD is suggested by some authors to be the most convenient design for sample clustering applications (Dobbin and Simon, 2002) or genetical genomics studies (Jansen and

Nap, 2001), where the aim of the research is to extract the individual variation in the expression profile. In addition, even if the aim of the experiment is to compare average profiles of gene expression among populations, the reference design may be preferable in some cases. For example, if the comparison of each treatment to a control group is more important than the other contrasts, the reference design can be as good as any other design (Glonek and Solomon, 2004). This situation may occur in preliminary or screening experiments including a large number of treatments that are compared to a reference group (e.g. several drugs against a placebo group or a time course experiment with a baseline). If the comparisons of treatments against a control group are the most relevant, the RRD and BRD are the only experimental designs that guarantee direct comparisons for those contrasts and consequently should be the most efficient designs. Additionally, the BRD will be more efficient than the RRD for the rest of the comparisons (among no control treatments).

With more than three treatments, the BRD will be more efficient than the balanced incomplete block design for a fixed number of biological replicates, under almost any variance ratio scenario (refer to formulas in Supplementary Material of Dobbin and Simon, 2002). Fixing the number of arrays, the direct comparisons under the RRD and BRD cases will have the same variance as in the balanced incomplete block design. The balanced incomplete block design, however, will still show some significant advantages over the reference designs because it provides the same variance for every contrast, while the BRD will have smaller variance for the comparisons against the control and larger variances for the contrasts among non-control treatments.

Lastly, another situation where the reference design may be preferred over other alternatives is when the number of groups or classes is uncertain at the beginning of the experiment, for example within a class discovery context (Dobbin and Simon, 2002). In these cases, the balanced incomplete block design is not indicated, as one needs to anticipate the number of groups (as well as each experimental unit membership) to be able to distribute them accordingly across the slides. Under these circumstances, the CRD is the most indicated design, especially if large quantities of reference samples are available.

The literature also shows a different aspect of the comparison of the reference design to other designs based on direct comparisons (Belbin et al., 2004; Konig et al., 2004). These authors compare the reliability of an indirect measurement of fold change (through a reference) to a direct measurement. Using a CRD, both works reached the same conclusion: the direct and indirect measures of log-ratio or fold change tend to agree if some kind of filtering criteria is applied to the data. This is expected because keeping only the most reliable spots from each array will diminish the overall technical variability and increase the reliability of indirect measurements. The results from Belbin et al. (2004) and Konig et al. (2004),

however, should be interpreted with some care as they included only limited technical replication, and no biological replication.

From a practical point of view, the researcher should examine not only the statistical power and the cost of the arrays and experimental units when designing a microarray experiment. Sample quantity and availability are also important issues. The CRD requires a large quantity of mRNA from a single sample (reference). The RRD, on the other hand, requires limited mRNA quantity from a large number of biological samples, which may be unfeasible when working with expensive or limited experimental units. Lastly, the BRD is an intermediate alternative in this aspect as it requires the same number of experimental units for all treatments (including the reference group), but the reference requires $(K-1)$ times more mRNA than the non-control groups. This last point may be prohibitive when working with limited mRNA quantity and a large number of treatments (e.g. experiments involving tissue samples from small animals or embryos).

Another issue to consider is the suitability of the control treatment as a reference. In general, a treatment with a large number of non-expressed genes will not be a good reference group. In these cases, a pool from several individuals, or a synthetic reference sample should be used in a CRD layout. This issue is out of the scope of this paper and the reader should refer to the existing literature (Gorreta et al., 2004; Kim, et al., 2002; Sterrenburg et al., 2002) for further discussion on this matter.

4.5 Concluding remarks

In summary, this paper proposes an alternative design for microarray experiments (denoted as BRD), which exploits the flexibility of the reference design, but minimizes the loss of efficiency of its indirect comparisons. Specifically, the use of one of the treatments as the reference saves arrays and provides a subset of direct comparisons, while the proper allocation of samples across arrays may yield lower variance than a RRD for the indirect contrasts. The BRD presents both of these advantages with a minimal increase in the required sample quantity (as discussed above) for the reference treatment. Among the situations where the BRD is likely to be implemented we list time course experiments (time zero as a control, for example) and experiments including competing drugs and a placebo group.

APPENDIX

Consider the following gene-specific model:

$$r_i = Y_{i1kl} - Y_{i2k'l},$$

where r_i is the normalized log-ratio relative to the i^{th} array, Y_{i1kl} is the log-intensity corresponding to the Cy3-labeled l^{th} sample from treatment k and $Y_{i2k'l}$ is the log-intensity of the Cy5-labeled l^{th} sample from treatment k'. Assume also that (Rosa et al., 2005):

$$Y_{ijkl} = \mu + A_i + D_j + T_k + B_{l(k)} + e_{ijkl},$$

where μ is the general mean, A_i, D_j and T_k are array, dye and treatment effect terms (considered as fixed effects here), respectively; $B_{l(k)}$ is a subject-specific effect, $i.i.d\ N(0,\sigma_a^2)$; and e_{ijkl} is a residual term, $i.i.d\ N(0,\sigma_e^2)$. In general, a model on the log-ratio scale will be derived from the formula above as follows:

$$\begin{aligned} r_i &= Y_{i1kl} - Y_{i2k'l} \\ &= \mu + A_i + D_1 + T_k + B_{l(k)} + e_{i1kl} - \left(\mu + A_i + D_2 + T_{k'} + B_{l(k')} + e_{i2k'l}\right) \\ &= (D_1 - D_2) + (T_k - T_{k'}) + (B_{l(k)} - B_{l(k')} + e_{i1kl} - e_{i2k'l}) \\ &= \alpha + \tau_{(k-k')} + \varepsilon_i, \end{aligned}$$

where $\alpha = D_1 - D_2$ is the difference between Cy3 and Cy5 labeling effects, $\tau_{(k-k')} = T_k - T_{k'}$ is the contrast between treatments k and k', and $\varepsilon_i = B_{l(k)} - Bl(k') + e_{i1kl} - e_{i2k'l}$ is a residual term containing both biological and technical components.

1. Model for CRD and ClRD (refer to Figure 1a; array 1 is the left-most array):

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \\ r_{11} \\ r_{12} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \tau_{(A-R)} \\ \tau_{(B-R)} \\ \tau_{(C-R)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}$$

Here $T_{k'}$ is R (the reference), which is always labeled with Cy5. Consequently the dye effect (intercept) can not be included in the model, but the effects of interest (τ) can still be estimated.

The variance of each error term ε_i is given by:

a) CRD:

$$\begin{aligned} Var(\varepsilon_i) &= Var(B_{l(k)} - B_{l(k')} + e_{i1kl} - e_{i2k'l}) \\ &= 2\sigma_e^2 + \sigma_a^2, \end{aligned}$$

and the covariance between any pair of elements is null, $Cov(\varepsilon_i, \varepsilon_{i'})=0$, so the (co)variance matrix relative to the residual vector $\boldsymbol{\varepsilon}$ can be expressed as $Var(\boldsymbol{\varepsilon}) = I_{(12)}(\sigma_a^2 + 2\sigma_e^2)$. Note that $Var(B_{l(k')}) = 0$ as there is no biological replication in the reference sample.

b) ClRD:

$$\begin{aligned} Var(\varepsilon_i) &= Var(B_{l(k)} - B_{l(k')} + e_{i1kl} - e_{i2k'l}) \\ &= 2\sigma_e^2 + 2\sigma_a^2. \end{aligned}$$

Likewise, the covariance between any pair of elements is null, $Cov(\varepsilon_i, \varepsilon_{i'})=0$. The (co)variance matrix relative to the residual vector $\boldsymbol{\varepsilon}$ can be expressed as $Var(\boldsymbol{\varepsilon}) = I_{(12)}(2\sigma_a^2 + 2\sigma_e^2)$.

2. Model for RRD (refer to Figure 1b; array 1 is the left-most array):

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \\ r_{11} \\ r_{12} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \tau_{(B-A)} \\ \tau_{(C-A)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}$$

As in this case the control treatment A ($T_{k'}$) is labeled with both Cy3 and Cy5, the dye effect (dye bias) can be modeled by including an intercept into the model. In a RRD, as the reference treatment is also replicated (at the biological level), the variance of each residual term is given by:

$$\begin{aligned} Var(\varepsilon_i) &= Var(B_{l(k)} - B_{l(k')} + e_{i1kl} - e_{i2k'l}) \\ &= 2(\sigma_e^2 + \sigma_a^2), \end{aligned}$$

which are independent to each other, so the residual (co)variance matrix can be expressed as $Var(\boldsymbol{\varepsilon}) = I_{(12)}2(\sigma_a^2 + \sigma_e^2)$. Notice that this matrix is equal to that of the ClRD.

3. Model for BRD (refer to Figure 4; array 1 is the left-most array):

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \\ r_{11} \\ r_{12} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \tau_{(B-A)} \\ \tau_{(C-A)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \end{bmatrix}$$

Similarly to the RRD, an intercept is included into the model to estimate (and to account for) the dye effect. Here also the residual variances are given by

$$\begin{aligned} Var(\varepsilon_i) &= Var(B_{l(k)} - B_{l(k')} + e_{i1kl} - e_{i2k'l}) \\ &= 2(\sigma_e^2 + \sigma_a^2). \end{aligned}$$

The residuals, however, are not independent. The (co)variances are given by:

$$\begin{aligned} Cov(\varepsilon_i, \varepsilon_{i'}) &= Cov(B_{l(k)} - B_{l(k')} + e_{i1kl} - e_{i2k'l}, B_{l(k'')} - B_{l(k''')} + e_{i'1k''l} - e_{i'2k'''l}) \\ &= Cov(-B_{l(k')}, -B_{l(k''')}), \end{aligned}$$

which can be expressed as:

$$Cov(-B_{l(k')}, -B_{l(k''')}) = \begin{cases} \sigma_a^2 & if\ B_{l(k')} = B_{l(k''')} \\ 0 & otherwise \end{cases}.$$

The residual (co)variance matrix is then written as

$$Var(\varepsilon) = I_{(6)} \otimes \begin{bmatrix} 2(\sigma_a^2 + \sigma_e^2) & \sigma_a^2 \\ \sigma_a^2 & 2(\sigma_a^2 + \sigma_e^2) \end{bmatrix},$$

where $\otimes$ is the Kronecker product operator.

REFERENCES

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**(6769):503-511.

Belbin, T. J., J. Gaspar, M. Haigentz, R. Perez-Soler, S. M. Keller, M. B. Prystowsky, G. Childs, and N. D. Socci (2004). Indirect measurements of differential gene expression with cDNA microarrays. *Biotechniques* **36**(2):310-314.

Bueno Filho, J. S. S., S. G. Guilmour, and G. J. M. Rosa (2005). Design of microarray experiments for genetical genomics studies. *Submitted.*

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat Genet* **32 Suppl**:490-495.

Dobbin, K. and R. Simon (2002). Comparison of microarray designs for class comparison and class discovery. *Bioinformatics* **18**(11):1438-1445.

Glonek, G. F. V. and P. J. Solomon (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics* **5**(1):89-111.

Gorreta, F., D. Barzaghi, A. J. VanMeter, V. Chandhoke, and L. Del Giacco (2004). Development of a new reference standard for microarray experiments. *Biotechniques* **36**(6):1002-1009.

Jansen, R. C. and J. P. Nap (2001). Genetical genomics: the added value from segregation. *Trends Genet* **17**(7):388-391.

Kerr, M. K. (2003a). Design considerations for efficient and effective microarray studies. *Biometrics* **59**(4):822-828.

Kerr, M. K. (2003b). Linear models for microarray data analysis: Hidden similarities and differences. *J Comput Biol* **10**(6):891-901.

Kerr, M. K. and G. A. Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**(2):183-201.

Kim, H., B. Zhao, E. C. Snesrud, B. J. Haas, C. D. Town, and J. Quackenbush (2002). Use of RNA and genomic DNA references for inferred comparisons in DNA microarray analyses. *Biotechniques* **33**(4):924-930.

Konig, R., D. Baldessari, N. Pollet, C. Niehrs, and R. Eils (2004). Reliability of gene expression ratios for cDNA microarrays in multiconditional experiments with a reference design. *Nucleic Acids Research* **32**(3):e29.

Lin, S. J., M. Kaeberlein, A. A. Andalis, L. A. Sturtz, P. A. Defossez, V. C. Culotta, G. R. Fink, and L. Guarente (2002). Calorie restriction extends Saccharomyces cerevisiae lifespan by increasing respiration. *Nature* **418**(6895):344-348.

Lonnstedt, I. and T. Speed (2002). Replicated microarray data. *Statistica Sinica* **12**(1):31-46.

Papp, B., C. Pal, and L. D. Hurst (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**(6945):194-197.

Perou, C. M., T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown, and D. Botstein (2000). Molecular portraits of human breast tumours. *Nature* **406**(6797):747-752.

Rosa, G. J. M., J. P. Steibel, and R. J. Tempelman (2005). Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comparative and Functional Genomics* **6**(3):123-131.

Sterrenburg, E., R. Turk, J. M. Boer, G. B. van Ommen, and J. T. den Dunnen (2002). A common reference for cDNA microarray hybridizations. *Nucleic Acid Research* **30**(21):e116.

Tempelman, R. J. (2005). Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. *Vet Immunol Immunopathol* **105**(3-4):175-186.

Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* **8**(6):625-637.

Yang, Y. H. and T. Speed (2002). Design issues for cDNA microarray experiments. *Nat Rev Genet* **3**(8):579-588.

Yang, Y. H. and T. P. Speed (2003). Design and analysis of comparative microarray experiments. *in Statistical Analysis of Gene Expression Microarray Data.* T. P. Speed, ed. CRC Press, Boca Raton, FL.

CHAPTER TWO

COMPARISON OF LOG-INTENSITY AND LOG-RATIO LINEAR MODELS FOR TWO COLOR MICROARRAYS.

ABSTRACT: In this paper, we compare log-ratio and log-intensity models for two-color microarray experiments with technical and biological replication. A linear transformation of the log-intensity model is used to derive the equivalent log-ratio model for a variety of designs. We demonstrate that some designs, such as a dye swap or connected loop designs, require the specification of random sample pair effects in the ratio. For a split plot design where arrays are experimental units, log-ratios are not convenient as this makes some effects unestimable in the model. Furthermore, analyses using log-intensity values generally allow for more efficient recovery of inter-slide information in design involving more than two treatments as we determine for different scenarios.

Introduction.

Differentially labeled fluorescence intensities are used to compare two samples for mRNA abundance of a particular transcript for any one spot on a long oligonucleotide or cDNA microarray. Statistical analyses of such two color microarray experiments have often been based on using logarithms of ratios (log-ratios) of these spot-specific fluorescence intensities as the response variables (Yang and Speed, 2003) although models based on the use of logarithm of the corresponding intensities (log-intensities) for each dye have also been used (Kerr et al., 2000; Wolfinger et al., 2001). Logarithms, typically to base 2, are often used to transform microarray fluorescence intensities and their corresponding ratios in the anticipation that the resulting data will facilitate analysis by parametric statistical methods that depend upon normality assumptions. Current statistical models for the analysis for log-intensities typically include fixed and random effects (i.e. mixed models) to account for different levels of replication (Rosa et al., 2005), whereas fixed effects linear models are more commonly used for log ratios (Smyth, 2004). Multilevel models have been proposed to analyze log-ratios although the most commonly used implementation requires a common intra class correlation assumption across transcripts (Smyth et al., 2005). Other generalized least squares (GLS) or mixed model analyses have been used to compare experimental designs using log-ratios measures of expression (Steibel and Rosa, 2005). The literature on the comparison of log-ratio versus log intensity models has focused its attention on those models that are readily specified for either response variable (Kerr, 2003b), but not in more complex designs that demand modeling different hierarchical levels of replication.

Modeling log-intensities allows specifying the array effect as random such that for some incomplete block designs, it is possible to recover inter array information (Graybill and Deal, 1959). On the other hand, an analysis based on ratios is strictly an intra array analysis such that there is a potential loss of information.

In this paper we demonstrate how to derive a linear model for log-ratios starting with linear mixed model specifications for log-intensities. We further compare log-ratio and log-intensity based models for several experimental designs. In the first section, general methodology for deriving the log-ratio model starting from a log-intensity model is presented. In the second section we apply the method to several designs and show in which cases and under which circumstances the results will be equivalent. In the last section we elaborate on the possibility of recovering inter-array information using a log-intensity model as opposed to a log-ratio model.

Linearly transforming a log intensity model to a log ratio model.

For any two color microarray design, a linear (mixed) model can be developed to analyze the log-intensity values (Wolfinger et al., 2001). These designs usually include fixed effects like the two dyes, treatments and other covariates as well as random effects such as arrays and other terms to account for different levels of replication. A log-ratio statistical model could be derived from a log-intensities model by simply applying a linear transformation. After applying this transformation, the original intercept and the array effects vanish. The new intercept for the log-ratios model is simply the dye effect (Vinciotti et al., 2005) whereas the treatment effect is parameterized in terms of a set of

treatment contrasts (Smyth, 2004). Furthermore, the variance components in the new log-ratios model will be a linear combination of the variance components in the log-intensity scale. As a result, random effects often still persist in the statistical model for log-ratio data in some designs such that they should be included. We further illustrate this property for some designs in the next section.

Ratio models for specific designs

Consider the common reference design based on, for example, n biological replicates per treatment for each of t treatments. The design then requires a total of nt arrays and $2nt$ intensity measurements, including nt measurements on the commonly labeled reference sample, if each gene is spotted only once on each array and there is no technical replication (i.e. no multiple measures per biological replicate within treatments). Let's suppose that the $2nt$ x 1 vector of log-intensities $\mathbf{y}$ is sorted by treatment and by array, as in the order provided from left to right in Figure 2.1, such that the first dye labeled sample is specified before the second dye labeled sample within each array. A suitable mixed effects model for the log intensity model could then be written as follows:

$$\mathbf{y} = \mathbf{1}_{nt}\mu + \mathbf{X}_D\begin{bmatrix}\gamma_1\\ \gamma_2\end{bmatrix} + \mathbf{X}_T\begin{bmatrix}\tau_1\\ \tau_2\\ \vdots\\ \tau_t\\ \tau_R\end{bmatrix} + \mathbf{Z}_A\begin{bmatrix}a_1\\ \vdots\\ a_{nt}\end{bmatrix} + \mathbf{Z}_B\begin{bmatrix}b_{1,1}\\ b_{1,2}\\ \vdots\\ b_{t,n}\\ b_{R,1}\end{bmatrix} + \mathbf{e} \qquad [1]$$

Here, $X_D = (\mathbf{1}_{nt} \otimes \mathbf{I}_2)$ and $\mathbf{X}_B = \left[\mathbf{1}_n \otimes \mathbf{I}_t \otimes \begin{bmatrix}1\\0\end{bmatrix} \mid \mathbf{1}_n \otimes \begin{bmatrix}0\\1\end{bmatrix}\right]$ define the design matrices for the fixed effects of dye and treatment respectively. Similarly, $Z_A = (\mathbf{I}_{nt} \otimes \mathbf{1}_2)$ and $\mathbf{Z}_B = \left[\mathbf{I}_{nt} \otimes \begin{bmatrix}1\\0\end{bmatrix} \mid \mathbf{1}_{nt} \otimes \begin{bmatrix}0\\1\end{bmatrix}\right]$ define the design matrices for the random effects. Note that $\mathbf{1}_n$ denotes an unitary row vector of dimension n whereas $\mathbf{I}_t$ denotes the identity matrix of order t. Also $\otimes$ denotes the Kronecker or direct product (Searle, 1982) whereas we use the notation | to denote the horizontal concatenation of two matrices. Hence μ denotes the overall mean, γ_1 and γ_2 denote the fixed effects of the two dyes, $\tau_1, \ldots \tau_R$ denote the treatment effects, a_1, a_2, ..., a_{nt} denote the random effects of arrays m = 1,2,...,nt, respectively, whereas $b_{k,l}$ denotes the random effect of the lth biological replicate within the kth treatment, noting that there is only one common replicate in the reference group having effect $b_{R,1}$. Finally, $\mathbf{e}$ denotes the vector of residuals which are presumed to be distributed as $N(\mathbf{0}, \mathbf{I}_{2nt}\sigma_e^2)$; similarly, $a_m \sim NIID\left(0, \sigma_a^2\right) \; \forall\, m$ and $b_{k,l} \sim NIID\left(0, \sigma_b^2\right) \; \forall\, k,l$.

Suppose that the vector of log-intensities $\mathbf{y}$ is sorted by array and dye (e.g. Cy3 and Cy5) within array such that the vector of log ratios, $\mathbf{r}$, can be determined as $\mathbf{r} = \mathbf{L}\mathbf{y}$, for $\mathbf{L} = \mathbf{I}_{nt} \otimes [1 \quad -1]$. Then, as expected, the dimension nt x 1 of $\mathbf{r}$ is half that of $\mathbf{y}$. Premultiplying also the right side of Equation [1] by $\mathbf{L}$, we derive the following equilavent log-ratio model for $\mathbf{r}$:

$$\mathbf{r} = \mathbf{1}_{nt}(\gamma_1 - \gamma_2) + (\mathbf{1}_n \otimes \mathbf{I}_t)\begin{bmatrix} \tau_1 - \tau_R \\ \tau_2 - \tau_R \\ \vdots \\ \tau_t - \tau_R \end{bmatrix} + \boldsymbol{\varepsilon} \qquad [2]$$

where $\boldsymbol{\varepsilon} \sim N\left(0, \mathbf{I}_{nt}\sigma_\varepsilon^2\right)$ for $\sigma_\varepsilon^2 = \sigma_{b*}^2 + 2\sigma_e^2$ with σ_{b*}^2 denoting the biological variation for the difference between the treated and common reference biological replicates.

Either of the two model specifications ([1] or [2]) will produce identical results using conventional ordinary least squares (OLS) estimation as appropriate when all effects are assumed to be fixed as they are in equation [1] if we momentarily assumed array effects to be fixed. However, assuming random array effects in equation [1] would potentially allow for the recovery of inter-block information (Kerr, 2003a) if data is missing.

Now consider a dye swap balanced block (t=2) design (Dobbin and Simon, 2002; Kerr, 2003a) with n biological replicates without technical replication such that n also defines the number of arrays as in Figure 2.2. Again, ordering the vector of log intensities $\mathbf{y}$ by array and then by dye within array we have the following intensity model:

$$\mathbf{y} = \mathbf{1}_{2n}\mu + \mathbf{X}_D\begin{bmatrix}\gamma_1 \\ \gamma_2\end{bmatrix} + \mathbf{X}_T\begin{bmatrix}\tau_1 \\ \tau_2\end{bmatrix} + \mathbf{Z}_A\begin{bmatrix} a_1 \\ \vdots \\ a_{tn}\end{bmatrix} + \mathbf{e},$$

where $\mathbf{X}_D = (\mathbf{1}_n \otimes \mathbf{I}_2)$, $\mathbf{X}_T = \mathbf{1}_{n/2} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$, and $\mathbf{Z}_A = (\mathbf{I}_n \otimes \mathbf{1}_2)$.

Invoking the same distributional specifications on the random and residual effects and developments analogous to those used in progressing from Equation [1] to [2] for the common reference design, a fixed effect linear effects model may be specified for the log-ratios in this balanced block design using **r** = **Ly** such that:

$$\mathbf{r} = \mathbf{1}_n(\gamma_1 - \gamma_2) + \left(\mathbf{1}_{n/2} \otimes \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)(\tau_1 - \tau_2) + \boldsymbol{\varepsilon} \qquad [3]$$

Here $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{\varepsilon})$ which can be shown to be expressed as a function of the residual and biological random effects of the log intensity model since $\sigma^2_{\varepsilon} = 2\sigma^2_b + 2\sigma^2_e$.

Let's now consider the balanced block design with *n* biological replicates per each of the two treatments but such that now a dye swap is conducted on the same two biological replicates hybridized twice (*m*=2) against each other. Then the total number of arrays is *nm*. Figure 2.3 illustrates the case for *m* = 2. The intensity model is specified as follows:

$$\mathbf{y} = \mathbf{1}_{4n}\mu + \mathbf{X}_D \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \mathbf{X}_T \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_t \\ \tau_R \end{bmatrix} + \mathbf{Z}_A \begin{bmatrix} a_1 \\ \vdots \\ a_{tn} \end{bmatrix} + \mathbf{Z}_B \begin{bmatrix} b_{1,1} \\ b_{2,1} \\ \vdots \\ b_{1,n} \\ b_{2,n} \end{bmatrix} + \mathbf{e},$$

with $\mathbf{X}_D = (\mathbf{1}_{2n} \otimes \mathbf{I}_2)$, $\mathbf{X}_T = \mathbf{1}_n \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$; $\mathbf{Z}_A = (\mathbf{I}_{2n} \otimes \mathbf{1}_2)$ and $\mathbf{Z}_B = \mathbf{I}_n \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$.

Invoking again our transformation $\mathbf{r} = \mathbf{L}\mathbf{y}$ and the same distributional assumptions on the random and residual effects as before, it becomes obvious that we need to specify the random effects of biological pair replicates in the log-ratios model for this design:

$$\mathbf{r} = \mathbf{1}_{nm}(\gamma_1 - \gamma_2) + \left(\mathbf{1}_n \otimes \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)(\tau_1 - \tau_2) + \left(\mathbf{I}_n \otimes \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)\begin{bmatrix} b_1^* \\ \vdots \\ b_n^* \end{bmatrix} + \boldsymbol{\varepsilon}. \qquad [4]$$

Here $b_l^* = b_{1,l} - b_{2,l}$ are the random biological pairing effects or the differences in the random biological replicate effects from the two treatments (1 and 2) paired against each other on the same array. Therefore, the distribution of these random effects as a function of the biological replicate variance component σ_b^2 on the log intensity scale is $b_l^* \sim NIID(0, 2\sigma_b^2)\forall l$; similarly, it could be readily shown $\boldsymbol{\varepsilon} \sim N(0, 2\mathbf{I}_{nm}\sigma_e^2)$ where σ_e^2 is the residual variance on the log intensity scale.

Note that the log-intensity model [3] and log-ratio model [4] will yield identical results if a mixed model analysis that properly specifies the random effects is used in both cases. However, only one random effects factor with half as many biological effects $(b_1^*, b_2^*, ,\ldots, b_n^*)$ needs to be specified in the log-ratios model [4]. If a mixed model software is not available (e.g. using OLS thereby ignoring $b_1^*, b_2^*, ,\ldots, b_n^*$), one might treat these biological pair effects as fixed and then compute a contrast based on those effects:

$$\mathbf{r} = \mathbf{1}_{2n}(\gamma_1 - \gamma_2) + \left(\mathbf{I}_n \otimes \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} + \boldsymbol{\varepsilon} \qquad [5]$$

Then, the treatment difference would be estimated by $\hat{\tau}_1 - \hat{\tau}_2 = \frac{1}{n}\sum_k \hat{b}_k$. This contrast gives an unbiased estimate of the treatment difference, however, it will not lead to the correct inference (i.e. standard errors and *P*-values) because this contrast along with [5] ignores biological variability.

Using model [5], the OLS statistic used to test the treatment effect contrast would fail to partition biological from technical variability. This is, under the null hypothesis $H_0 : \tau_1 - \tau_2 = 0$, the *F*-statistic from model [4] is not distributed following a central *F*-distribution (as expected) if there is biological variability for gene expression. Conversely because of the simpler experimental design, model [3] will provide an *F* statistic that will indeed test for treatment effect, even in the presence of non-null σ_b^2. In this design, treating the array as fixed or random has no effect because each array represents a complete block.

A connected loop design is an incomplete block design that can be shown to be particularly effective for comparing $t > 2$ treatments in a two color microarray system (Tempelman, 2005). Figure 2.4 illustrates such a design for $t = 3$. The corresponding log intensity model for $t = 3$ as described in Rosa et al. (2005) is presented below:

$$\mathbf{y} = \mathbf{1}_{6n}\mu + \mathbf{X}_D\begin{bmatrix}\gamma_1\\ \gamma_2\end{bmatrix} + \mathbf{X}_T\begin{bmatrix}\tau_1\\ \tau_2\\ \vdots\\ \tau_t\\ \tau_R\end{bmatrix} + \mathbf{Z}_A\begin{bmatrix}a_1\\ \vdots\\ a_{3n}\end{bmatrix} + \mathbf{Z}_B\begin{bmatrix}b_{1,1}\\ b_{2,1}\\ b_{3,1}\\ \vdots\\ b_{1,n}\\ b_{2,n}\\ b_{3,n}\end{bmatrix} + \mathbf{e},$$

where,

$$\mathbf{X}_D = (\mathbf{1}_{3n} \otimes \mathbf{I}_2),\ \mathbf{X}_T = \mathbf{1}_n \otimes \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 1\\ 1 & 0 & 0\end{bmatrix},\ \mathbf{Z}_A = (\mathbf{I}_{3n} \otimes \mathbf{1}_2) \text{ and } \mathbf{Z}_B = \mathbf{I}_n \otimes \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 1\\ 1 & 0 & 0\end{bmatrix}.$$

This model could also be transformed to a log-ratio model using the same linear transformation (i.e. $\mathbf{r} = \mathbf{L}\mathbf{y}$) described previously. Equation [6] represents the corresponding log-ratio model for the t=3 as from Figure 2.4.

$$\mathbf{r} = \mathbf{1}_{3n}(\gamma_1 - \gamma_2) + \mathbf{X}_t\begin{bmatrix}\tau_1 - \tau_3\\ \tau_2 - \tau_3\end{bmatrix} + \mathbf{Z}_b\begin{bmatrix}b_{1,1}\\ \vdots\\ b_{3,n}\end{bmatrix} + \boldsymbol{\varepsilon} \qquad [6]$$

where $\mathbf{X}_t = \mathbf{1}_n \otimes \begin{bmatrix}1 & -1\\ 0 & 1\\ -1 & 0\end{bmatrix}$ and $\mathbf{Z}_b = \left[\mathbf{I}_n \otimes \begin{bmatrix}1\\ 0\\ -1\end{bmatrix} \middle| \mathbf{I}_n \otimes \begin{bmatrix}-1\\ 1\\ 0\end{bmatrix} \middle| \mathbf{I}_n \otimes \begin{bmatrix}0\\ -1\\ 1\end{bmatrix}\right]$ are known design matrices. Again, the distribution of random effects may be parameterized in terms

of the log-intensity model variance components $b_{k,l} \sim NIID(0, \sigma_b^2)$ and $\boldsymbol{\varepsilon} \sim NIID(0, 2\sigma_e^2)$.

Unlike that of any of the other previously discussed designs, the assumption on the specification of the array effects as fixed or random in the log-intensity model is important in the connected loop design. Specifying array effects as fixed leads to the log-ratio model specified in [6] above. The residual variance estimated with model [4] will be twice the residual variance of the intensity model, while the biological variance estimates will be identical to that from the log-intensity model. However, if array effects are assumed to be random in the log-intensity model, there will be recovery of inter-array information and the estimates of the treatment difference (and their standard errors) will be different between the two models.

There are some microarray designs where there is no equivalent model between the log-intensity and log-ratio scales, where array effects are treated as fixed or random. This is true for the split plot design, as illustrated in Figure 2.5, where there are at least two different treatment factors, say S and T. Consider the simplest case where each factor has two levels. For the subplot factor, for example T, comparisons are made between two different levels or treatments, T_1 and T_2 for biological replicates within each array, analogous to that considered for the balanced block design in Figure 2.2. Levels of the wholeplot factor S are further superimposed on this design such that s arrays are assigned to S_1 and s arrays to S_2. In other words, arrays serves as the experimental units for factor S whereas biological replicates serve as the experimental

units for T. Suppose that the data vector **y** is sorted by the two levels of S, by array and then by dye. Then the linear mixed model for the log-intensity model is presented below.

$$\mathbf{y} = \mathbf{1}_{4a}\mu + (\mathbf{I}_{2s} \otimes \mathbf{I}_2)\begin{bmatrix}\gamma_1 \\ \gamma_2\end{bmatrix} + (\mathbf{I}_2 \otimes \mathbf{1}_{2s})\begin{bmatrix}\alpha_1 \\ \alpha_2\end{bmatrix} + \mathbf{1}_s \otimes \begin{bmatrix}1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0\end{bmatrix}\begin{bmatrix}\tau_1 \\ \tau_2\end{bmatrix} +$$

$$\left(\mathbf{I}_2 \otimes \mathbf{1}_3 \otimes \begin{bmatrix}1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0\end{bmatrix}\right)\begin{bmatrix}\alpha\tau_{11} \\ \alpha\tau_{12} \\ \alpha\tau_{21} \\ \alpha\tau_{22}\end{bmatrix} + (\mathbf{I}_{2s} \otimes \mathbf{1}_2)\begin{bmatrix}a_{1,1} \\ \vdots \\ a_{2,s}\end{bmatrix} + \mathbf{e},$$

where α_1 and α_2 denote the fixed effects of the main plot factor, τ_1 and τ_2 denote the fixed effect of the sub plot factor, and $\alpha\tau_{11}$ ...$\alpha\tau_{22}$ denote the fixed effect of the interaction between factors, $a_{k,l} \sim NIID\left(0, \sigma_a^2\right)$ and $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$. Given that array serves as the experimental unit for S, array effects must be treated as random. By pre-multiplying the incidence matrices of this model by **L**, the columns corresponding to the main plot factor vanish. In other words, the following log-ratio model cannot test for main plot factor effects.

$$\mathbf{r} = \mathbf{1}_{2s}\delta + \left(\mathbf{1}_a \otimes \begin{bmatrix}1 \\ -1\end{bmatrix}\right)(\tau_1 - \tau_2) + \left(\mathbf{I}_2 \otimes \mathbf{1}_3 \otimes \begin{bmatrix}1 \\ -1\end{bmatrix}\right)\begin{bmatrix}\alpha\tau_{11} - \alpha\tau_{12} \\ \alpha\tau_{21} - \alpha\tau_{22}\end{bmatrix} + \boldsymbol{\varepsilon}, \qquad [7]$$

where $\boldsymbol{\varepsilon} \sim NIID(0, 2\sigma_e^2)$.

For the other fixed effects of interest (the sub-plot factor, interaction and simple effects of subplot factor within main plot factor), the tests based on the log-ratio model and log intensity model will be equivalent.

Recovery of inter-array information.

The analysis of log-ratios is equivalent to performing an intra-array analysis; in other words, there is no other choice than to treat array effects as fixed when analyzing log-ratios. On the other hand, one may treat array effects as fixed (intra-array analysis) or random (combined inter-intra-array analysis) for the analysis of log intensity data. In this section we describe a method for estimating the amount of recovery of inter-array information using a mixed model analysis on log-intensities (i.e. treating arrays as random).

Consider, for example, the reference design of Figure 2.1 using Equation [1] as the analysis model. Assuming array effects to be fixed would be equivalent to analyzing the difference (log-ratio) between the two treatments at each spot. The (intra-array) estimate of the treatment effect is:

$$\widehat{\tau_1 - \tau_2} = \frac{1}{n}\sum_{i}\left(y_{i1l} - y_{iR1}\right) - \frac{1}{n}\sum_{i'}\left(y_{i'2l} - y_{i'R1}\right) \qquad [8]$$

Equation [8] shows that the intra-array estimate of the treatment difference is actually equivalent to a log-ratio analysis. Specifically, instead of analyzing the intensities we take the log ratio for each array: $r_i^{(k)} = y_{ikl} - y_{iR1}$ and then we fit equation

[2]. Conversely, if the array effects are considered random, two estimates of a treatment difference are available:

a) Inter-array estimate: $\widehat{\tau_1 - \tau_2}^{(1)} = \frac{1}{n}\sum_{i}(y_{il} + y_{iR1}) - \frac{1}{n}\sum_{i'}(y_{i'2l} + y_{i'R1})$

b) Intra-array estimate: $\widehat{\tau_1 - \tau_2}^{(2)} = \frac{1}{n}\sum_{i}(y_{il} - y_{iR1}) - \frac{1}{n}\sum_{i'}(y_{i'2l} - y_{i'R1})$

Note that in these expressions, the choice of treatments ($k=1, 2$ in this example) will determine the set of arrays (i or i') used in the computation of contrasts. Also, note that the subindex l for subject is entirely determined by the array subindex i. As before, there is only one subject (R_1) for the reference.

A single, minimum variance estimate can be obtained for the treatment difference with:

$$\widehat{\tau_1 - \tau_2}^{*} = \frac{\left(\sigma_e^2\right)\left(\widehat{\tau_1 - \tau_2}\right)^{(1)} + \left(\sigma_e^2 + 2\sigma_a^2\right)\left(\widehat{\tau_1 - \tau_2}\right)^{(2)}}{2\left(\sigma_e^2 + \sigma_a^2\right)} \qquad [9]$$

A mixed model (equation [1] with random array) may be used to obtain the combined intra-inter array estimate of the treatment difference in equation [9]. Kerr (Kerr, 2003a; 2003b) present expressions to compute the variance of a treatment contrast in this design assuming either fixed or random array effects. Specifying arrays as random facilitates a more efficient estimation of the treatment difference. For more complex designs, the analytical derivation of the contrast variance under a mixed model may be unwieldy, although a numerical approach is possible using linear mixed model software

(Rosa et al., 2005; Tempelman, 2005). This procedure has been used to compute the relative efficiency of two designs, but can also be applied to compute the relative efficiency of two models to estimate the treatment difference.

Assume an arbitrary linear mixed model (in the log-intensity scale):

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is a vector containing the elements y_{gijlk}, $\boldsymbol{\beta}$ is a vector of fixed effects (e.g. dye and treatment effects), $\mathbf{u}$ is a vector of random effects (e.g. array and biological replicate), $\boldsymbol{\varepsilon}$ is a vector of residuals, and $\mathbf{X}$ and $\mathbf{Z}$ are design-specific incidence matrices. For $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ we assume:

$$\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim MVN\left(\begin{bmatrix}0\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\right),$$

where $\mathbf{G}$ is a diagonal variance-covariance matrix for the random effects with elements σ_a^2 and σ_b^2 and $\mathbf{R}=\mathbf{I}\sigma_e^2$, so $\mathbf{y}\sim N(\mathbf{X\beta},\mathbf{V})$, with $\mathbf{V}=\mathbf{ZGZ'}+\mathbf{R}$. In this context, the comparison of experimental designs may be performed in terms of their relative efficiency to estimate a treatment difference. Given two alternative designs, A and B, their relative efficiency may be expressed as:

$$RE_{A:B} = \frac{\widehat{Var(\tau_1 - \tau_2)_B}}{\widehat{Var(\tau_1 - \tau_2)_A}} \qquad [10]$$

If $RE_{A:B}$ is > 1 implies that design A is more efficient than design B to estimate the treatment difference. The variances under each design may be estimated analytically (Kerr, 2003a) or numerically (Stroup, 2002) depending on the particular designs and

models under consideration. In general, given design A with fixed and random effects incidence matrices $\mathbf{X}_A$ and $\mathbf{Z}_A$, respectively, and a contrast vector $\mathbf{K}'$, such that, for example, $\mathbf{K}'\hat{\boldsymbol{\beta}} = \widehat{\tau_1 - \tau_2}$ specifies the estimated treatment difference, then

$$Var(\widehat{\tau_1 - \tau_2})_A = \mathbf{K}'\left[\mathbf{X}'_A(\mathbf{Z}_A\mathbf{G}\mathbf{Z}'_A + \mathbf{R})^{-1}\mathbf{X}_A\right]^{-}\mathbf{K}.$$

To estimate the increase in efficiency due to recovery of inter-array information, the equation [8] is used replacing $Var(\widehat{\tau_1 - \tau_2})_A$ with the variance of the estimated treatment difference from the fixed array effects model and $Var(\hat{\tau}_2 - \hat{\tau}_1)_B$ with the contrast variance derived from the mixed model as follows:

$$Var(\widehat{\tau_1 - \tau_2})_B = \mathbf{K}'\left[\mathbf{X}'_B(\mathbf{Z}_B\mathbf{G}\mathbf{Z}'_B + \mathbf{R})^{-1}\mathbf{X}_B\right]^{-}\mathbf{K}.$$

For example, given a common reference design, the relative efficiency of the mixed model to the fixed array effects model is illustrated in Figure 2.6. The maximum increase in efficiency occurs when the biological and array variance are small compared to the residual variances. For any variance ratio close to 1, we expect a moderate increase in efficiency on the order of 20% to 30% whereas for biological or array variances more than four times the residual variance, the gain in efficiency will be negligible.

Analytical calculations are not possible in the connected loop design comparison such that we use a numerical approach to estimate the contrast variances (Table 2.2). The mixed model procedure (PROC MIXED) of SAS was used to compute the contrast variances for both models conditional on the variance components ratios $\sigma_a^2 : \sigma_e^2$ and

$\sigma_b^2 : \sigma_e^2$. For this design, the maximum value for the relative efficiency is 1.33 and occurs when both biological and array variance components are zero. If the magnitude of the residual variance is on the same order of magnitude of the other variance components (variances ratios equal to one), the recovery of inter block information will be less that a 5% increase in efficiency.

Discussion

Linear mixed models are commonly used for the analysis of two color microarray data. Software packages are either oriented towards the analysis of log-intensities or log-ratios. Previous studies have compared log-ratio and log-intensity models for designs with only one level of replication (Kerr, 2003a, b; Vinciotti et al., 2005). We have extended these comparisons to experiments with technical and biological replication. An appropriate log-ratio model can be derived from a linear transformation of the log-intensity model specifying various levels of technical and biological replication. By pre-multiplying the model components by a contrast matrix, we have shown that for some designs (connected loop, dye swap) there is an equivalent linear model in the log-ratio scale.

For the balanced block design, the two approaches are always equivalent. In particular, a balanced block design without technical replication can be analyzed in either of the two scales using fixed effects models with identical results. If technical replication is present, the random effects of the sample (in the log-intensity model) or sample pair (in

the log-ratio model) should also be included. Moreover, treating the biological replicates as fixed effects does not produce the correct tests as we have shown in Table 2.1.

If more than two treatments are included in the design (for example the loop or reference designs), the derived log-ratio model will be equivalent to the log-intensity model based on fixed array effects. However, we have also demonstrated that the analysis of log-ratios is sometimes not useful as it is not possible to estimate certain effects (for example in a split-plot design). This situation can be considered an extreme case of information loss in the intra-array analysis.

Analyzing log-ratios rather than log-intensities may involve a loss of information, particularly if array effects are properly treated as random in the latter. The mixed model with random array effects will tend to be more efficient than the fixed array effects model for log-intensities (equivalent to the log-ratio model) as we have illustrated in a common reference design and in a connected loop design. In those cases, the amount of recovery of inter array information is a function of the variance ratios. Large residual variance relative to both array and biological variances will lead to the maximum information recovery using mixed model analyses in the common reference design. Simmilarly, for the connected loop design, if either the array or biological variances are small compared to the residual variance, the recovery of inter-array information will be maximized using the mixed model analysis.

It is commonly perceived that two color microarrays yield a measure of relative expression and consequently the ratios are to be preferred over absolute intensities. However as true this argument is, it does not invalidate the use of log-intensity models,

provided that array effects are specified. Moreover, for a simple design (dye swap design without technical replication), the log-ratio and log-intensity models are completely equivalent. This is analogous to the comparison of a paired *t*-test and ANOVA for a randomized complete block design with two treatments. The paired *t*-statistic is the square root of the F test from the ANOVA and both will yield identical *P*-values.

Another point to consider is the effect of normalization on the model comparison. In this paper we assumed data to be properly normalized as well as log-transformed. In general, the normalization is done in the log-ratio scale using intra-array LOESS (Yang et al., 2002) from which normalized log intensities are derived. Consequently there is no difference induced by the normalization procedure applied before fitting either log-intensity or log-ratio models. Other methods of normalization, for example quantile normalization, may apply differently to ratios or intensities and that could be an additional source of differences.

To conclude, we have shown that in complex designs a log-ratio model could be derived from a linear transformation of the corresponding log-intensity model. However, the log-intensity models are, in general, easier to elicit, more flexible, and eventually more efficient than log-ratio models. Analyses of log-ratios, however, may still be useful for simple designs, especially for slightly improving computational tractability, if there is no expected recovery of inter-array information from analyzing log-intensity data.

Table 2.1. Expected mean squares of relevant terms in models [4] and [5]. In both cases, the model [4] is assumed to be the data generation model. σ_e^2 Is the residual variance in the log-intensity scale, σ_B^2 is the individual variance in the log-intensity scale.

	Model [4]	Model [5]
EMS(Trt)	$\sigma_e^2 + 2\sigma_b^2 + Q(Trt)$	$2\sigma_e^2 + 4\sigma_b^2 + Q'(Trt)$
EMS(B)	$\sigma_e^2 + 2\sigma_b^2$	-
EMS(error)	σ_e^2	$2\sigma_e^2$

Table 2.2. Relative efficiency of intra-inter-slide (random array) analysis compare to the intra-slide (fixed array) analysis in the loop design (Figure 2.4). σ_e^2 Is the residual variance in the log-intensity scale, σ_B^2 is the individual variance in the log-intensity scale.

$\lambda=$	$\rho=\sigma_B^2:\sigma_e^2$						
$\sigma_A^2:\sigma_e^2$	*0*	*1/4*	*1/2*	*1*	*2*	*4*	*→∞*
0	*1.33*	*1.22*	*1.17*	*1.11*	*1.07*	*1.04*	*1.00*
¼	*1.22*	*1.15*	*1.12*	*1.08*	*1.05*	*1.03*	*1.00*
½	*1.17*	*1.12*	*1.09*	*1.06*	*1.04*	*1.02*	*1.00*
1	*1.11*	*1.08*	*1.06*	*1.04*	*1.03*	*1.01*	*1.00*
2	*1.07*	*1.05*	*1.04*	*1.03*	*1.02*	*1.01*	*1.00*
4	*1.04*	*1.03*	*1.02*	*1.01*	*1.01*	*1.01*	*1.00*
→∞	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*

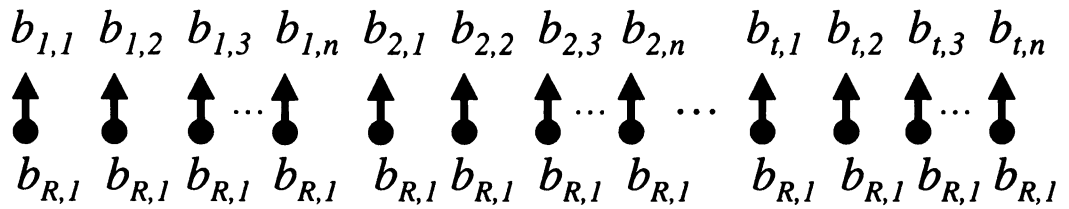


Figure 2.1. Common reference design. B_{kl} is the l^{th} biological replicate within the k^{th} group.

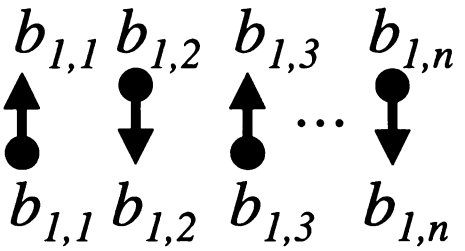


Figure 2.2. Complete block design with dye swap . B_{kl} is the l^{th} biological replicate within the k^{th} group.

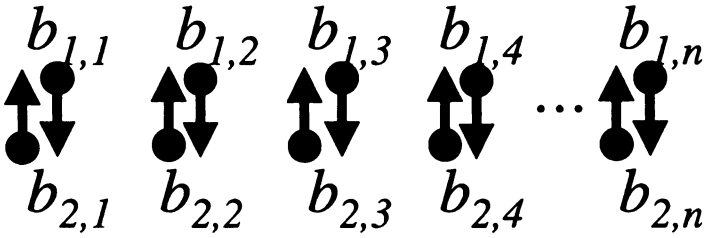


Figure 2.3. Dye swap design with technical replication. B_{ij} is the j^{th} biological replicate within the i^{th} group.

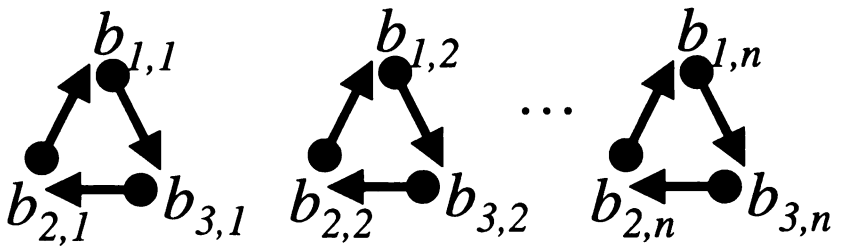


Figure 2.4. Connected loop design. B_{kl} is the l^{th} biological replicate within the k^{th} group

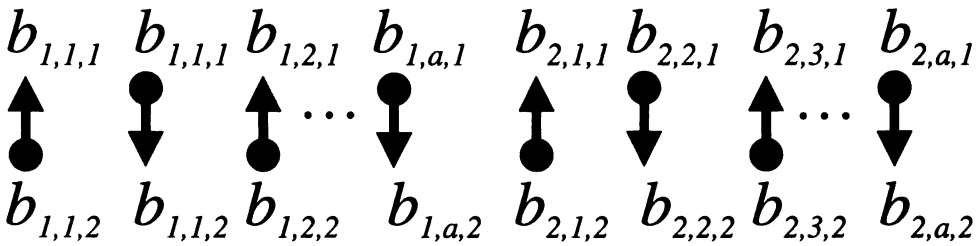


Figure 2.5. Split-plot design. B_{kls} is the l^{th} biological replicate within the k^{th} group of factor T receiving the s^{th} level of factor S. The same experimental unit ($\boldsymbol{B_{ij}}$) is assigned to two levels of the factor S (subplot factor) in the same array. Independent experimental units are used for the two levels of the factor T (whole plot factor).

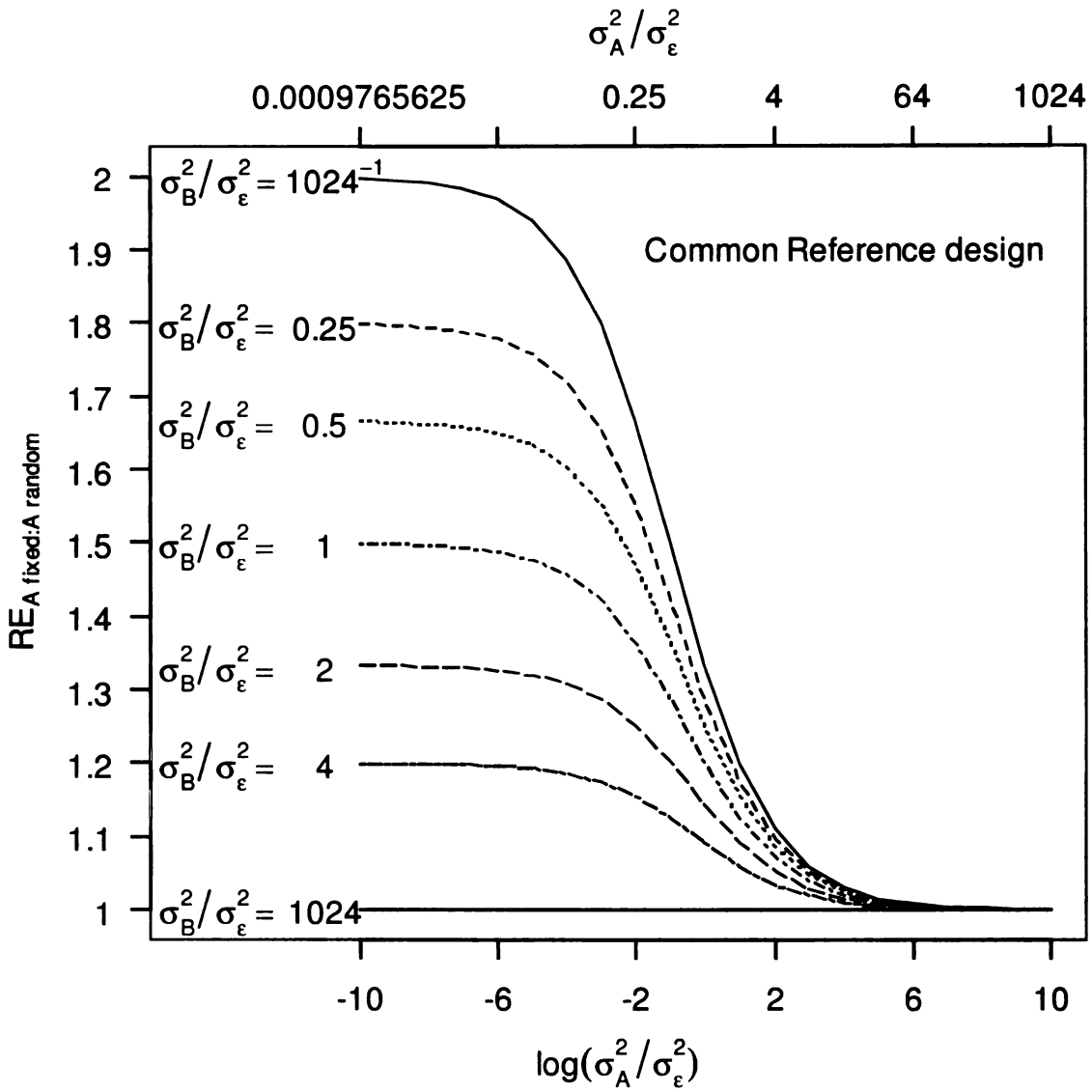


Figure 2.6. Relative contrast variance under different assumptions for the array effect in reference designs. σ^2_B is the biological variance, σ^2_A is the array variance and σ^2_ε is the residual variance.

CHAPTER THREE

LINEAR MIXED MODELS FOR THE ANALYSIS OF RELATIVE QUANTIFICATION OF RT-PCR DATA

ABSTRACT: Quantitative reverse transcription polymerase chain reaction (qRT-PCR) is currently viewed as the most precise technique to quantify levels of messenger RNA. Relative quantification compares the expression of a target gene under two or more experimental conditions normalized to the measured expression of a control gene. The statistical methods and software currently available for the analysis of relative quantification of RT-PCR data lack the flexibility and statistical properties to produce valid inferences in a wide range of experimental situations. In this paper we present a novel method for the analysis of relative quantification of RT-PCR data, which consists of the analysis of cycle threshold values (C_T) for a target and a control gene using a general linear mixed model. Our method allows testing of a broader class of hypotheses than traditional analyses such as the comparative C_T. For all possible pairwise comparisons, the estimated fold change was the same using either linear mixed models or a comparative C_T method, but a simulation study indicated that the linear mixed model approach is more powerful. In summary, the method presented in this paper is more accurate, powerful and flexible than the traditional analysis methods for analysis of RT-PCR data. This new method will be especially useful for studies involving more than two treatments and multiple experimental factors.

1. Background

Reverse transcription (RT), followed by quantitative polymerase chain reaction (qPCR), is currently viewed as the most accurate, sensitive, and specific technique to quantify levels of messenger RNA (Bustin, 2000). At present, there are several instrumentations and chemistries available for implementation of this technique, all of which rely on the same fundamental principle (Bustin, 2000). This principle consists of the specific amplification of cDNA from a target transcript in several cycles of PCR coupled with measurement of a fluorescence intensity assumed to be directly proportional to the amount of product in each cycle (Giulietti et al., 2001). This methodology has been extensively validated, and its accuracy and specificity have been proven for the different chemistries available (Winer et al., 1999).

The quantitative output of the RT-PCR consists of an amplification curve, which is composed of a set of cycle numbers and associated fluorescence intensities that are ulteriorly summarized in a single value called cycles to threshold (C_T). The C_T is a unitless value defined as the fractional cycle number at which the sample fluorescence signal passes a fixed threshold above the baseline. Because the threshold is arbitrarily set within the exponential amplification phase, the C_T is inversely proportional to the log of the initial transcript copy number (or log-transcript concentration) of the assayed sample. The constant of proportionality of the C_T to the log-concentration is the amplification efficiency (E).

Absolute and relative quantification strategies can be applied to measure mRNA abundance using qRT-PCR (Giulietti et al., 2001; Johnson et al., 2000). Absolute

quantification allows a direct comparison of expression between different treatments, but it is more costly than relative quantification. This is specially the case when several genes are profiled in one experiment, requiring the fit of a standard curve for each target gene. On the other hand, relative quantification is more practical to be implemented on a large scale but its statistical analysis remains a challenge.

Relative quantification compares the expression of a target gene under various conditions (treatments) normalized to the measured expression of an internal control (Pfaffl, 2001) (assumed to be constantly expressed across samples). In general, the numerous mathematical expressions available for this calculation (Gentle et al., 2001; Liu and Saint, 2002; Livak and Schmittgen, 2001; Marino et al., 2003; Muller et al., 2002; Ramakers et al., 2003; Swillens et al., 2004; Tichopad et al., 2004; Tichopad et al., 2003; Tichopad et al., 2002) may be summarized by equation (1) (Pfaffl, 2001):

$$FC_{trt_1:trt_2} = \frac{\left(E_{\text{Target}}\right)^{\Delta C_{\text{T(target)}}(trt_2 - trt_1)}}{\left(E_{Control}\right)^{\Delta C_{T(control)}(trt_2 - trt_1)}}, \quad (1)$$

where, $FC_{trt_1:trt_2}$ is the relative expression (fold-change) of the target gene in a sample from treatment 1 compared to a sample from treatment 2, E_{Target} and E_{Control} are the amplification efficiencies of the target and the control genes, respectively, and $\Delta C_{\text{T(target)}}(trt_2 - trt_1)$ is the C_T of the treatment 2 minus the C_T of the treatment 1. If both amplification efficiencies take the maximum possible value (E=2), expression (1) becomes the familiar $2^{-\Delta\Delta Ct}$ expression (Livak and Schmittgen, 2001). Moreover, almost any other mathematical expression or method available in the literature to calculate the fold change is a variant of the expression (1). The differences among the variants of

equation (1) refer mainly to the estimation of the efficiency either from a relative standard curve (Johnson et al., 2000; Pfaffl, 2001) or from the individual amplification curves (Gentle et al., 2001; Swillens et al., 2004).

The methods based on expression (1) are mathematical equations devised to calculate the fold change between two samples. These equations, however, lack the statistical formalism needed to draw valid inferences, especially when multiple biological replicates from each experimental group are assayed (Gentle et al., 2001; Marino et al., 2003; Pfaffl et al., 2002). Moreover, many ad-hoc approaches associated with formulas similar to (1) have been used with the objective of generating a set of "companion" p-values or standard errors (Livak and Schmittgen, 2001; Muller et al., 2002). However, few of them are valid in the presence of both biological and technical replication. Currently, the REST® software (Pfaffl et al., 2002) is one of the few programs that implements a valid statistical analysis to test hypotheses and estimate the fold changes using expression (1).

However, such software is limited to the simplest case of an experimental design because it can analyze only pair-wise comparisons among groups under a completely randomized design. A linear mixed model (Cook et al., 2004) was recently proposed for the implementation of the so-called analytical method (Marino et al., 2003). Such a model is potentially more flexible than the existing alternatives, but it makes the strong assumption that there is a common random effect for the control and test genes in each biological replicate. Consequently, there is a necessity for a formal statistical method for the analysis of the relative quantification of RT-PCR data that allows the accommodation

of more complex experimental designs (such as blocking factors) and the testing of general hypotheses (including interactions, pairwise and group contrasts).

The objective of this paper is to present a novel, flexible method for the analysis of relative quantification of RT-PCR data using linear mixed models. A variety of approaches are used to validate the proposed methodology, to compare it with existing methods, and to illustrate its flexibility. First, our model is compared to other alternatives using a real dataset; second, a model-free simulation based on that dataset is used for comparative validation of the methodology; third, several datasets are analyzed and different linear models are compared; and lastly, the results from comparative C_T and linear models are compared in those datasets.

2 Results

2.1 Motivating example

Quantitative RT-PCR was used to study the expression of the gene diazepam binding inhibitor (DBI) in the brain of piglets subject to weaning and social isolation (Poletto et al., 2006). The experimental layout followed a completely randomized block design (n=3 litters) and the treatments consisted of a 2×2 factorial combination of weaning (early weaned or non-weaned) and social isolation (isolated or control).

Preliminary assays indicated that *Sus scrofa* 18S ribosomal RNA (18S) was suitable for use as an endogenous control gene. The estimated amplification efficiency of primers of the two genes (18S and DBI) was close to two (Table 3.1 Supplementary material). All reactions were performed in triplicates but some observations were excluded from the analysis because of evidence of non-specific amplifications (as

revealed by dissociation curve analyses) (Giglio et al., 2003). Model (2) was used for the analysis of the expression of DBI normalized to the expression of 18S:

$$y_{gijkr} = TG^{*}_{gi} + l_{gj} + B_{gijk} + D_{ijk} + e_{gijkr}, \quad (2)$$

where y_{gijkr} is the C_T obtained from the thermocycler software for the g^{th} gene from the r^{th} well, corresponding to the k^{th} animal in the j^{th} litter subjected to the i^{th} treatment, TG^{*}_{ig} is the mean of treatment i in the expression of gene g (18S and DBI), l_{gj} ~N(0, σ^2_{lg}) is a gene-specific random effect of the j^{th} litter, B_{gijk} ~N(0, σ^2_{Bg}) is a gene-specific random effect of the k^{th} piglet in the j^{th} litter, D_{ijk} ~N(0, σ^2_D) is a random sample specific effect (common to both genes) and e_{gijkr} ~N(0, σ^2_e) is a residual term. The sample specific effect, D_{jk}, captures differences among samples that are common to both genes, particularly those that affect total mRNA concentration, such as differential extraction or amplification efficiencies among samples. The treatments consisted of the combination of two factors, and the sub-index *i=1, 2, 3, 4* corresponds to: early weaning + control (EWC), early weaning + isolation (EWI), non-weaning + control (NWC) and non-weaning + isolation (NWI), respectively.

Model (2) was fit to the data using the SAS mixed procedure (Littell, 1996) and a residual analysis was performed to check the parametric assumptions of the model. Tests of differential expression among groups were performed for the interaction of weaning by isolation and for pair-wise treatment differences (simple effects). Point and interval estimates of the fold changes were approximated from the linear contrasts (in the log

scale) by back transformation. The fold changes were also estimated with the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001) (ΔC_T) using the two procedures presented in the original work (Poletto et al., 2006). In the first variant ($\Delta CT_{[1]}$) , equation (1) (assuming E=2) was applied to each pair of littermates and the resulting pair-wise fold change was then averaged. In the second variant ($\Delta CT_{[2]}$), equation (1) was applied on the averaged pair-wise C_T difference among littermates. A paired t-test was used to assess the significance of the fold change calculated with $\Delta CT_{[2]}$; no statistical tests were performed on the fold change obtained with $\Delta CT_{[1]}$ because of insufficient replications to implement the suggested non-parametric test (Ben Ahmed et al., 2004; Kishimoto et al., 2004; Martell et al., 1999; Pellagatti et al., 2003; Rhoden et al., 2004).

Finally, model (3) was also used to analyze the data:

$$y_{gijkr} = TG^{*}_{gi} + l_j + D_{ijk} + \varepsilon_{gijkr}. \tag{3}$$

Model (3) is a simplified version of model (2) without the random sample and litter effects, and is equivalent to a previously published model for analysis of amplification curve data (Cook et al., 2004).

2.2 Testing and estimating differential expression

There was no evidence of interaction effect between isolation and weaning on the expression of DBI (P=0.829), but there was a significant three-fold decrease in the DBI expression due to isolation (P=0.003). The traditional analysis method ($\Delta CT_{[2]}$) does not

allow testing of this interaction, but it may be still used to estimate the fold change of pair-wise comparisons (Fig. 3.1).

The estimates of fold change were similar using Model (2), (3) and $\Delta CT_{[2]}$, but the estimates were slightly larger (in absolute terms) using $\Delta CT_{[1]}$ (Fig. 3.1). For instance, in the NWI-NWC comparison, the fold change from Models (2), (3) and $\Delta CT_{[2]}$ was 0.35 (i.e., suppression of 2.9 fold) while $\Delta CT_{[1]}$ yielded an estimate of 0.40 (i.e., suppression of 2.5 fold).

The confidence intervals for the fold changes based on $\Delta CT_{[2]}$ were wider than those based on Models (2) and (3), and the general conclusions were not equivalent. For example, Models (2) and (3) indicated a significant decrease in the expression of DBI in response to social isolation in both early-weaned and non-weaned animals (P=0.013 and P=0.019 respectively from model (2)), While $\Delta CT_{[2]}$ only detected EWI - EWC as significant (P=0.03). At a significance level of α=5%, Model (2) and Model (3) yielded the same conclusions, but the confidence intervals were narrower for Model (3). Then, depending on the significance level adopted, the conclusions might differ.

2.3 Validation through simulation

A simulation experiment was used to validate and compare alternative analysis methods. A fair simulation study in this case precluded the use of any of the analysis models as the data generation process. Alternatively, we permuted the real data to generate a population of 1000 datasets with known fold changes, while keeping the original data structure,

distribution and variability. The simulated datasets were analyzed using four methods: Model (2), $\Delta CT_{[1]}$, $\Delta CT_{[2]}$ and Model (3).

Table 3.1 summarizes the point estimates of the fold change obtained for each contrast using the four analysis methods. Model (2), Model (3) and $\Delta CT_{[2]}$, yielded unbiased estimates of the fold change and the three methods had roughly the same mean square error. On the other hand, $\Delta CT_{[1]}$ produced biased estimates and also showed a larger mean square error. The increase in the mean square error for the $\Delta CT_{[1]}$ based estimates was due to both their bias and their larger variance. Moreover, the bias was always upwards and, consequently, the over expression was exaggerated and the down regulation was understated. For example, the expected fold change in the contrast EWI-EWC was 2.0 but the average estimate from $\Delta CT_{[1]}$ was 2.2. Conversely, for the contrast EWC-NWC the true fold change was 0.707, but $\Delta CT_{[1]}$ produced an average fold change estimate of 0.805.

Confidence intervals (95%) for the fold changes were computed using Model (2), Model (3) and $\Delta CT_{[2]}$ (Table 3.2). The narrowest confidence intervals corresponded to Model (3), followed by Model (2) and $\Delta CT_{[2]}$. Nevertheless, a further analysis of the real coverage of these "nominal" 95% confidence intervals (values within parenthesis in Table 3.2) revealed that Model (3) yielded intervals with significantly less coverage than the other two methods. The real coverage of confidence intervals obtained from Model (3) was well below the nominal 95% confidence level. Confidence intervals calculated from Model (2) exhibited the closest coverage to the nominal level. The coverage

obtained from $\Delta CT_{[2]}$ derived confidence intervals was also close to 95%, but the width of the confidence intervals were sensibly larger.

Model (2) and Model (3) allowed testing general linear hypotheses related to the interaction and the main effects (including pair-wise comparisons of treatments). Conversely, $\Delta CT_{[2]}$ only allowed single effect contrasts between pairs of treatments (Table 3.3).

Under the null hypothesis (contrast NWI - NWC in Table 3.3), Model (2) and $\Delta CT_{[2]}$ yielded a type I error rate very close to the nominal 5% test value, and the discrepancies observed in Table 4.4 for the $\Delta CT_{[2]}$ method were within the expected simulation error (based on extensive simulations not shown in this paper). The realized type I error rate of tests from Model (3) was clearly above the nominal error level (0.07 for α=0.01 and 0.18 for α=0.05). The histograms of the p-values (Figure 3.2) show in more detail the anomalous distribution of the p-values from model (3). The expected distribution of p-values for a series of independent tests under null hypothesis is uniform over the interval [0,1]. While the p-values from Model (2) and $\Delta CT_{[2]}$ exhibited the typical (uniform) distribution expected under the null hypothesis, the inverted-J shape in Panel (a) of Figure 3.2 reveals an excess of smaller p-values.

Under the alternative hypothesis (Table 3.3, all comparisons except NWI - NWC), Model (3) showed the highest probability of declaring significant a fold change larger than 1, but part of this apparent power comes from an inflated type I error rate as shown

before. Model (1) was more powerful than $\Delta CT_{[2]}$ and, in absolute terms, the increase in power was more evident for larger fold changes.

2.4 Model checking and result comparison in experimental datasets.

Although Model (2) is suitable only for the analysis of the described dataset, an equivalent model can be elicited for any specific data structure or design layout. The main components of Model (2) are the random sample effects and the random interaction between sample and gene factors. Moreover, gene specific variances are assumed for the sample-gene interaction. The measurement error term (residual effects) is assumed homoskedastic with respect to genes.

To test the adequacy of these assumptions in a broader set of experimental data (Abruzzo et al., 2005; Coussens et al., 2003; Coussens et al., 2004; Peirson et al., 2003; Poletto et al., 2006; Szabo et al., 2004), six different datasets where analyzed. Details of the datasets are presented in Table 4.4. The datasets included one to 63 test genes and four to 80 biological samples. All datasets but one included technical replicates (assay replicates). Twelve alternative models were compared using the Akaike information criteria and the Bayesian information criteria (Schwarz, 1978). The models represented the combinations of different assumptions: sample specific random effect (included or not), sample-gene random interaction (homoskedastic, heteroskedastic, or not included in the model), and residual variance (homogeneous or heterogeneous across genes).

The effects included in the best-fit model for each dataset are shown in the Supplementary Material. A random sample effect was present in all models. Similarly, the gene by sample interaction with heterogeneous variances among genes was selected for almost all the datasets. The only exception was the MRD (Peirson et al., 2003) dataset

where the model with homogeneous variances was preferred. A possible explanation for this is that the dataset included only two genes (control and test) that may have not shown significant heterogeneity of variances to select the alternative model, while the inclusion of more test genes increases the chances of having differential variances of gene expression. Gene specific residual variances were also generally favored by the model selection criteria. In the TLD (Abruzzo et al., 2005) dataset, heterogeneous residual variances could not be fit due to convergence problems. In the SHK (Szabo et al., 2004) dataset, the residual term included both the residual and the sample-gene interaction effects of the other models because the dataset lacks technical replicates. Consequently, the model with heterogeneous residuals indicates the presence of a gene by sample interaction with heterogeneous variances, heterogeneity of variance in the measurement errors, or both. In the remaining datasets, the heterogeneity of residual variances was caused by different (gene specific) precisions for the measurement of gene expression. Disentangling the sources of such heterogeneity is beyond the scope of this paper, but we anticipate that differential amplification efficiencies may be one of such causes.

Except for a few subtle differences among the models selected for each dataset, the general model including sample and gene-sample random effects was always preferred. The inclusion of heterogeneous residual variances had only a marginal effect on the tests for differential expression (results not shown). In contrast, omitting the sample-gene effect from the models increased the type I error rates over the nominal value (as shown in the previous sections).

2.5 Comparison of results with $2^{-\Delta\Delta CT}$ methods

The datasets were used to compare the estimation of ratios of expression between groups using the linear mixed model and the $2^{-\Delta\Delta CT}$ methodology (ΔCT). For each comparison of interest (Supp. Material), the point estimates and 95% confidence intervals of the log-ratio of gene expression were obtained using the linear mixed model and the ΔCT.

Figure 3.3a presents the point estimates. As shown, there was an almost perfect agreement between the two model specifications. This result can actually be generalized for other gene expression datasets so that the linear mixed model and the ΔCT methodology are expected to produce very similar point estimates of fold changes.

Figure 3.3b presents the results for the width of the confidence intervals. In general, there are more points above the indifference line than below it. This implies that the confidence intervals for log-ratios obtained using the linear mixed model tended to be shorter than those obtained with the ΔCT method. Nevertheless, there were differences among datasets. In general, the datasets involving only two groups (MRD (Peirson et al., 2003) and SHK (Szabo et al., 2004)) showed very similar length of confidence intervals for the two methods. In contrast, datasets involving multiple groups tended to show markedly shorter confidence intervals with the linear mixed models. An exception was the SPL (Coussens et al., 2003) dataset, for which the width of the confidence intervals did not show a clear pattern between the linear model and the ΔCT for the various contrasts (treatment pairs), possibly due to an apparent heterogeneity of variances among groups.

3. Discussion

In this paper we presented a novel method for the analysis of relative quantification of RT-PCR data. Our approach consists of the analysis of (raw or efficiency corrected) C_T values for a target and a control gene using a general linear mixed model. Currently, the use of qRT-PCR is pervasive in functional genomics studies and the complexity of experimental designs or sampling schemes have increased considerably (Hamalainen et al., 2001; Martinez et al., 2004; Recinos et al., 2004). However, the statistical and mathematical approaches available for the analysis of such data lack the flexibility and statistical properties necessary to produce useful and valid inferences in complex experimental layouts. Conversely, our method is flexible and allows the incorporation of an arbitrarily complex experimental protocol in both the treatment structure (factorial, time courses, etc.) and the sampling scheme (blocks, split-plots, etc.). Furthermore, the linear mixed model allows testing any general linear hypothesis; for instance, in the first real data example presented in this paper, we could test the hypothesis of interaction between social isolation and early weaning in the expression of DBI in the brain of the piglets. In contrast, the traditional analysis method ($2^{-\Delta\Delta CT}$) could not test the same hypothesis and its application was restricted to pair-wise comparisons of treatments. With other datasets, we could test for linear and quadratic trends in time course experiments and for interactions and main effects in a 2×3 split-plot design (results of these contrasts not shown).

For pair contrasts, our method (Model (2)), Model (3) and $\Delta CT_{[2]}$ produced similar estimates of the fold change, but another implementation of the $2^{-\Delta\Delta CT}$ method

($\Delta CT_{[1]}$) produced slightly different estimates. The $\Delta CT_{[1]}$ implementation of the $2^{-\Delta\Delta CT}$ is frequently used in practice, and its significance is usually assessed with non parametric tests (Ben Ahmed et al., 2004; Kishimoto et al., 2004; Martell et al., 1999; Pellagatti et al., 2003; Rhoden et al., 2004). Such a procedure could not be applied in this case because of limited sample size. For hypothesis testing and interval estimation the methods yielded divergent results: Model (3) showed the highest significance (lower p-values and narrower confidence intervals) and $\Delta CT_{[2]}$ showed the least significant results; Model (2) produced intermediate results.

These differences are consequence of the assumptions behind each procedure. Particularly, our model assumes a Gaussian distribution of the log expression. It also assumes heterogeneous variances in the expression of the target gene and the control gene, and the presence of sample-specific effects related to the measurement protocol. A priori, all these assumptions are plausible. The assumption of normally distributed log-expression levels has been extensively used (Andersen et al., 2004; Brunner et al., 2004; Szabo et al., 2004). Also, the heteroskedastic models for the analysis of several candidate control genes presented better fit than homoskedastic alternatives (Szabo et al., 2004).

The aforementioned assumptions were reasonable in this experiment; however, we also conducted a simulation study to evaluate the performance of our model. We avoided the use of a parametric model for the simulation by creating a population of datasets using resampling methods. In other words, neither the original distribution of the C_T values nor the relative technical and biological variabilities were altered. From this simulation, Model (2) emerged as the best model for the analysis showing a correct type I

error rate and confidence intervals coverage. Contrarily, Model (3) overstated the significance of the comparisons and the coverage of the confidence intervals. While these results are specific for these data, we think that the inclusion of a gene-sample specific effect separated from a sample specific effect (as in Model 2) is more plausible than solely inclusion of a sample effect common to both genes (as it is implicit in Model 3).

Comparing the simulation results from Model (2) and $\Delta CT_{[2]}$, it is evident that the simultaneous analysis of all groups (Model (2)) provides more power than the independent pairwise comparisons ($\Delta CT_{[2]}$). This is not surprising because a pair contrast within Model (2) had 6 degrees of freedom while the paired t-test associated with $\Delta CT_{[2]}$ had only 2 degrees of freedom. Moreover, the advantage of the mixed model methodology over the $\Delta CT_{[2]}$ would be larger if more treatments were included. For instance, in a completely randomized design with 10 treatments or groups and 3 biological replicates in each, any t-test between a pair of treatments will have 4 degrees of freedom, while an ANOVA based F-test will have 20 degrees of freedom. On the other hand, if a certain experiment is restricted to two treatments or groups, both methods will yield identical results.

We validated these conclusions by analyzing a set of experimental data that included different numbers of genes and biological replications. The assumptions regarding random effects were confirmed by model selection in every dataset. Additionally, we found heterogeneous residual variances in most of the datasets. Using the same datasets, the increased overall power of the linear mixed model analysis over the comparative C_T for large numbers of groups was illustrated using the width of the

log-ratio confidence interval. In general, comparisons from experiments involving more groups tended to show significantly shorter confidence intervals using the linear mixed model, while comparisons from experiments with two groups yielded equivalent confidence intervals with any methodology.

4. Conclusions

In summary, we have shown the importance of proper modeling of qRT-PCR data for correctly controling the type I error, and we have provided a general method for the analysis. The most important feature of our modeling approach is the use of (raw or efficiency corrected) C_T data as response variables to conduct a joint analysis of target and control gene expression, modeling simultaneously the biological and technical variation. Furthermore, Model (2) is a single alternative implementation of the linear mixed model approach (the most appropriate for our real data example), but it was easily expanded to fit data from other designs. Finally, our method is more accurate, powerful and flexible than any existing analysis method and it is especially useful in studies involving more than two treatments or time points and multiple experimental factors.

5. Methods

5.1 Materials and RT-PCR reactions

Sample collection, mRNA extraction, cDNA synthesis and PCR protocols are described in detail in the Supplementary Material.

5.2 Model derivation

We assume that the expression z_{gijk} (copy number or concentration of mRNA) of gene g in sample k of litter (block) j and experimental group i can be described by:

$$\log\left(z_{gijk}\right) = TG_{gi} + l_{gj} + B_{gijk}, \tag{4}$$

where TG_{ig} is the mean expression of gene g in the i^{th} treatment, l_{gj} is the random effect of litter on each gene [l_{gj}~N(0, σ^2_{lg})] and B_{gijk} is the gene and sample specific effect [B_{gijk} ~N(0, σ^2_{Bg})].

If mRNA is isolated, cDNA is synthesized and qRT-PCR is conducted in several independent wells for each sample in the presence of primers for each of the genes, the generated data may be analyzed with the following model:

$$y_{gijkr} = TG_{gi} + l_{gj} + B_{gijk} + D_{ijk} + e_{gijkr}, \tag{5}$$

where y_{gijkr} is a measured expression level in the log scale (for example: C_T), D_{ij} is a sample specific effect introduced by the experimental protocol and ε_{gijkr} is a well-specific measurement error.

In Model (5), D_{ij} represents a measurement artifact that is sample specific, and it is assumed to be $D_{ijk} \sim N(d_i, \sigma^2_D)$. This implies that the experimental protocol affects the measurement on the sample for all assayed genes in the same way, but it may generate a treatment bias (d_i). These assumptions (apart from the specific Gaussian distribution) are

standard in relative quantification analyses. Moreover, the existence of the D_{ij} effects is supposedly the reason to include a control gene in such assays.

If we assume that the $TG_{gi} = \mu + \tau_{gi}$ and that $\tau_{gi} = 0$ for the control gene, and we fit Model (5) assuming $D_{ij} \overset{i.i.d}{\sim} N(0, \sigma_D^2)$ (i.e. Model(2)), the following values for the TG^*_{gi} effects are expected:

$$\begin{cases} TG^*_{gi} = \mu + \tau_{gi} + d_i, \text{ for } g = \text{target} \\ TG^*_{gi} = \mu + d_i, \text{ for } g = control \end{cases} \tag{6}$$

5.3 Hypothesis testing and estimation

Suppose that the interest is to estimate the fold change between EWI (*i=2*) and EWC (*i=1*) for the target gene (*g=2*) normalized to the control gene (*g=1*). This is equivalent to estimate the log-difference (or log of the fold change) using:

$$dif_{(EWI-EWC)} = (TG^*_{22} - TG^*_{21}) - (TG^*_{12} - TG^*_{11}). \tag{7}$$

It is clear that if (6) holds (i.e. there is no differential expression of the control gene), the expectation of (7) is:

$$E\left[dif_{(EWI-EWC)}\right] = TG_{22} + d_2 - TG_{21} - d_1 - (d_2 - d_1) = TG_{22} - TG_{21}, \tag{8}$$

which is the quantity of interest. Furthermore, point estimates, hypothesis tests and confidence intervals of (7) are readily available, and the fold-change estimates may be approximated by transforming point and confidence interval limits to the correct scale. For example, if the data z are C_T values, the fold change estimation formulae would be:

$$FC_{(EWI-EWC)} = 2^{-diff_{(EWI-EWC)}}. \tag{9}$$

5.4 Data and response variable

The response variable for Model (5) may be any measure proportional to the log-mRNA concentration in the samples. In our particular example, the amplification efficiency was close to the optimal value (E=2), and consequently the C_T values constituted a suitable response. The Supplementary Material includes a detailed explanation of an alternative response variable when the amplification efficiency is smaller than two.

5.5 Programs for analysis

The models implemented in this paper can be readily fit using mixed model software. Particularly, a SAS code is available in the Supplementary Material.

5.6 Simulation study

From the expression data (raw C_T) from the target (DBI) and the control (18S) genes, we computed the arithmetic mean of each treatment (averaging out all the available biological and technical replicates), subtracted the corresponding average treatment value from each individual observation and added the general mean. The result of this procedure is a dataset that keeps the original variability among litters and among technical replicates, but it has a common mean for all treatments. Subsequently, the data were reshuffled to create 1000 datasets. Within each litter, the treatment memberships were permutated among the four treatments, but the technical replicates were kept together. Then, the observations corresponding to EWC animals were increased by the value 0.5 and the observations corresponding to EWI animals were decreased by the value 0.5. Consequently, the resulting population of trials had roughly the same biological and technical variability of the original data, but known fold change for each treatment pair (second column in Table 3.1).

The analysis of this population of datasets provided a set of 1000 p-values for each of the hypotheses tested. The type one error rate (α) was estimated by the rate of rejections (for certain nominal α) in the comparison EWC-NWC. Conversely, power was estimated counting the number of rejections in any non-null hypothesis. The coverage of the confidence intervals was estimated from the proportion of intervals that contained the true fold change value.

5.7 Model selection

A detailed model description used for each dataset of Table 3.4 for model selection and results comparison is presented in the Supplementary Material.

Contrast	$E(FC)^a$	$FC[2]^b$	$FC[\Delta CT1]^c$	$FC[\Delta CT2]^d$	$FC[3]^e$
EWI - EWC	2.000	2.059 (0.3922)	2.220 (0.4958)	2.059 (0.3959)	2.054 (0.3720)
EWI - NWI	1.414	1.442 (0.1588)	1.558 (0.2004)	1.443 (01602)	1.440 (0.1503)
EWC - NWC	0.707	0.744 (0.0470)	0.805 (0.0618)	0.744 (0.0474)	0.742 (0.0443)
NWI - NWC	1.000	1.054 (0.0925)	1.140 (1.3200)	1.054 (0.0937)	1.051 (0.0875)

Table 3.1 Properties of the point estimates of the fold change. [a]expected (true) value of the fold change for each comparison. The other columns present the mean estimates with their mean squared errors (in parenthesis) from 1000 simulations. [b]Model (2); [c]$\Delta CT_{[1]}$; [d] $\Delta CT_{[2]}$; and [e]Model (3).

Contrast	$CI_{[2]}$ [a]	$CI_{[\Delta CT]}$ [b]	$CI_{[3]}$ [c]
EWI - EWC	3.03 (93.7 %)	6.03 (91.0 %)	1.63 (79.8 %)
EWI - NWI	2.12 (95.4 %)	4.34 (94.1 %)	1.14 (84.2 %)
EWC - NWC	1.09 (94.7 %)	2.24 (93.0 %)	0.58 (81.4 %)
NWI - NWC	1.55 (94.7 %)	3.22 (92.9 %)	0.83 (82.0 %)

Table 3.2. Properties of the interval estimates of the fold change. Each column presents the mean width of the 95% confidence interval for the fold change with the actual coverage of the interval in parenthesis. Desirable properties are coverage close to 95% and small interval width. [a]Model (2); [b]$\Delta CT_{[2]}$ and [c]Model (3).

Contrast	α= 1%			α= 5%		
	$p_{[2]}$[a]	$p_{[\Delta CT2]}$[b]	$p_{[3]}$[c]	$p_{[2]}$[a]	$P_{[\Delta CT2]}$[b]	$p_{[3]}$[c]
EWI - EWC	0.157	0.058	0.700	0.439	0.257	0.872
EWI - NWI	0.030	0.026	0.246	0.131	0.098	0.415
EWC - NWC	0.031	0.026	0.276	0.152	0.104	0.429
NWI - NWC	0.010	0.029	0.072	0.053	0.071	0.180
wean. x isol.	0.075	-	0.416	0.249	-	0.584

Table 3.3. Properties of the hypothesis tests. p: proportion of rejected tests from the 1000 simulated datasets in each contrast at two significance levels (α). The NWI-NWC contrast corresponded to the null hypothesis (no differential expression) and the expected value is $p= \alpha$. For contrasts different from NWI-NWC, a larger value of p implies more power. The subindex indicates the analysis method. [a]Model (2); [b]$\Delta CT_{[2]}$ and [c]Model (3).

Dataset name	Number of genes	Experimental Design	Biological samples	Assay replicates
SPLIT (Coussens et al., 2003)	2	Split plot design. Main plot factor: disease status. Subplot factor: infection	12	2
TC (Coussens et al., 2004)	5	Longitudinal time course experiment. Timepoints: 0, 2, 4, 8, 16 hours after infection.	Four	2
PFC (Poletto et al., 2006)	5	2 x 2 Factorial in a randomized complete block design. Factors: weaning; isolation	12	3
MRD (Peirson et al., 2003)	2	Completely randomized design with two groups.	Eight	3
TLD (Abruzzo et al., 2005)	64	Completely randomized design with two groups.	Nine.	4
SHK (Szabo et al., 2004)	6	Completely randomized design with one group.	80	1

Table 3.4. Experimental data for model checking and result comparison. The letters in parenthesis are the abbreviations used in the text to refer to each dataset.

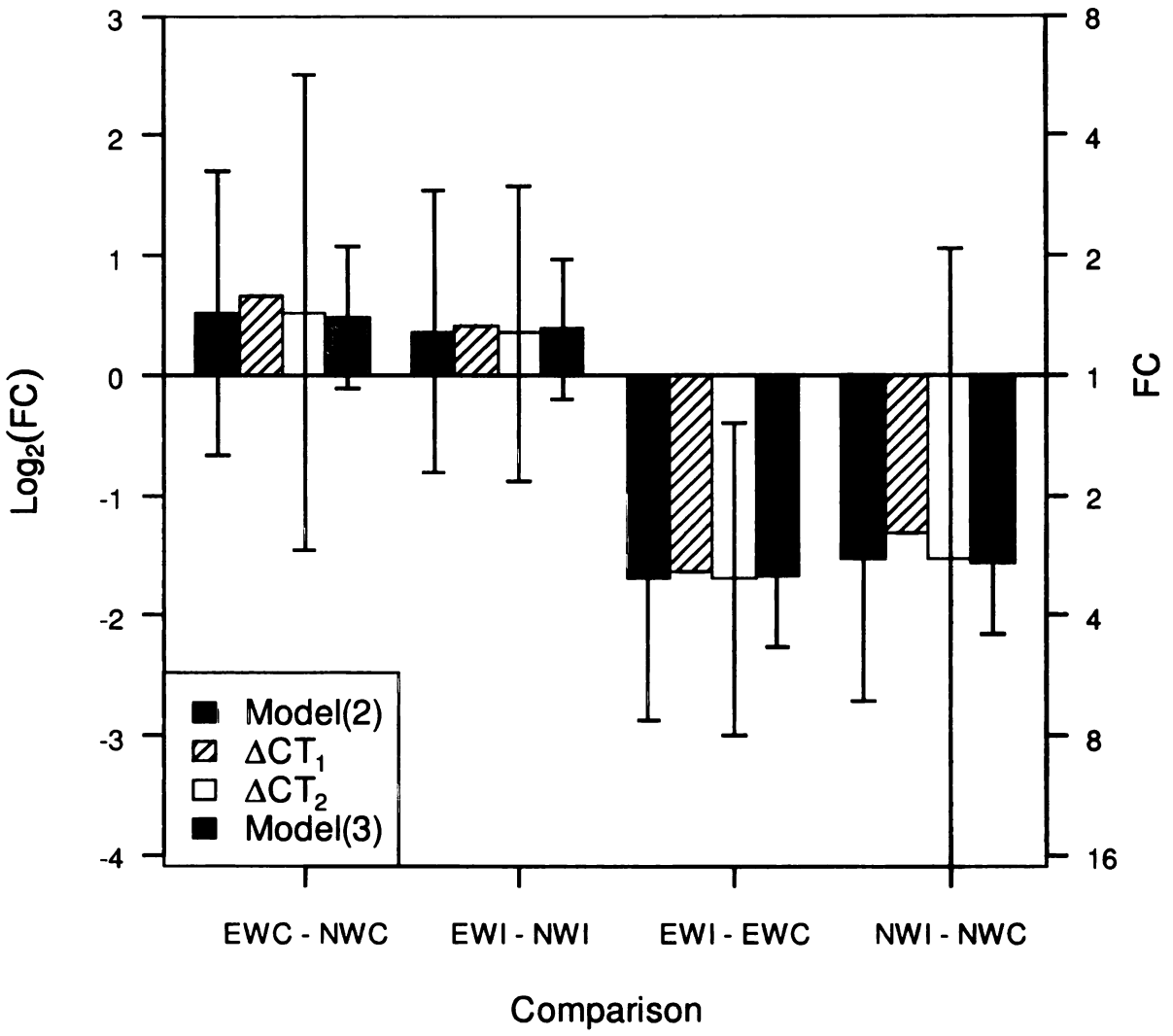


Figure 3.1. Fold change estimates. The $\log_2$ fold changes for four pairs of contrasts (abscissa) are presented. Fold change scale is included on the right axis. Segments indicate the 95% confidence interval. Comparisons whose confidence interval include the value 0 (1 in fold change scale) are not significant at α=5%. A confidence interval or significant level for $\Delta CT_{[1]}$ could not be calculated.

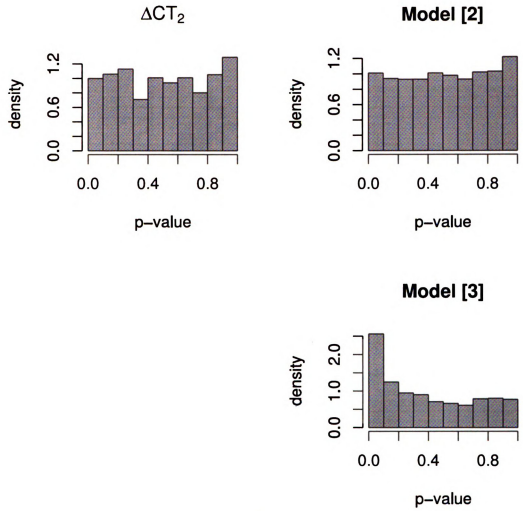


Figure 3.2.Histograms of p-values under null hypothesis. Under null hypothesis, the p-values are expected to follow a uniform distribution over the [0, 1] interval. The departure from that distribution in model (3) indicates an excess of false positives (higher frequencies for smaller p-values).

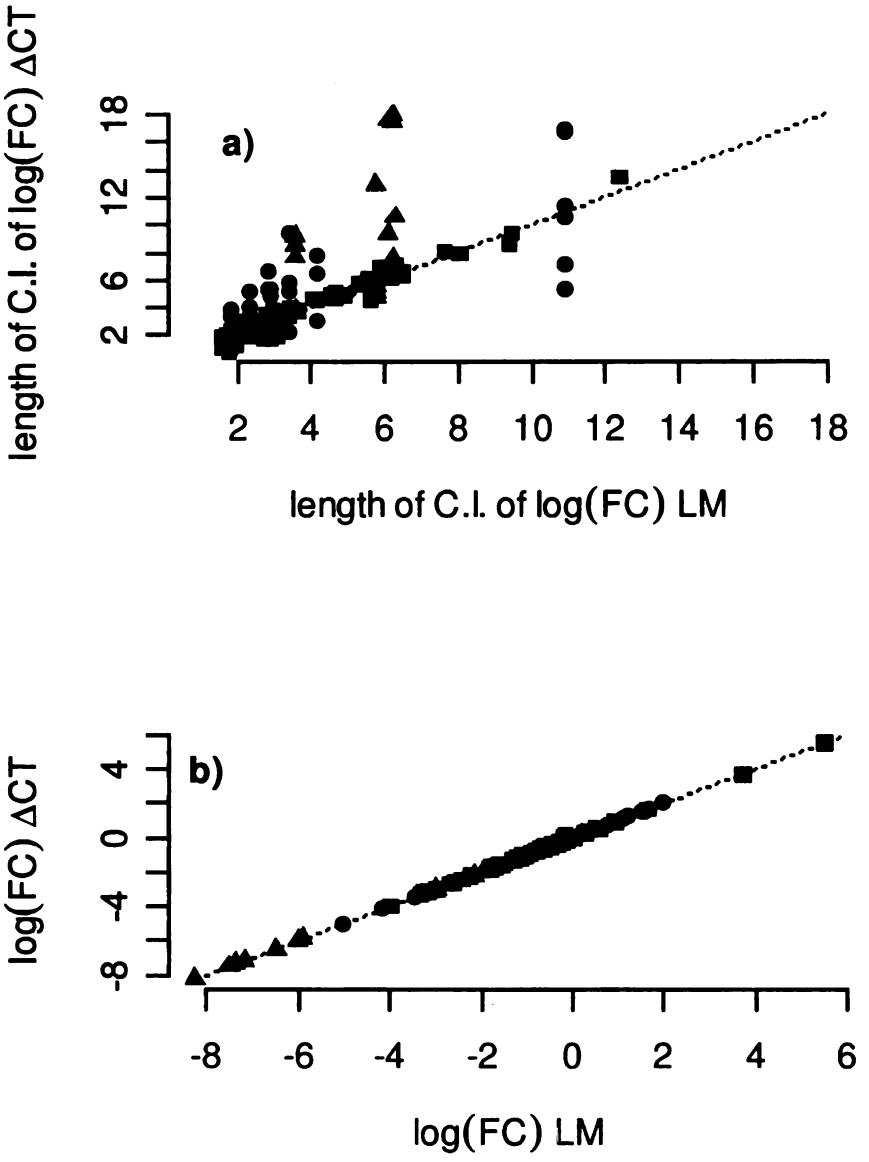


Figure 3.3. Result comparison between ΔCT (vertical axis) and linear mixed model (horizontal axis). (a) log fold changes. (b) length of the confidence interval for log fold changes. The 1:1 line is represented by dots. Comparisons in each dataset are indicated by symbols. ■: TLD, ●: SPLIT, ▲: TC, ◆: MRD and ●: PFC.

TH

2c

CHAPTER FOUR

DESIGN AND ANALYSIS OF TWO-STAGE EXPERIMENTS FOR TRANSCRIPTIONAL PROFILING

1. Introduction

Gene expression microarrays are powerful tools for screening the transcription profile for thousands of genes simultaneously on any particular sample. Despite the continuous sensitivity and increased dynamic range of microarrays, this technology is still regarded as lacking some precision. This limitation has prompted genomics researchers to validate their microarray results using an independent technique, typically quantitative reverse transcription polymerase chain reaction (qRT-PCR), although some authors have expressed skepticism about the necessity of doing so (Rockett, 2003; Rockett and Hellmann, 2004). At any rate, in some scientific journals, such validation is mandatory, for example in Circulation Research and in Arthritis and Rheumatism (Rockett, 2003).

Allison et al. (2006) has recently posed the important question of what is validation and how should it be performed. We assume that the ultimate objective of gene expression technologies is to efficiently search for genes that are differentially expressed (DE) between two or more conditions. Generally, the expression of a set of genes that were declared DE by a microarray experiment is further screened using qRT-PCR. If a gene is also concluded to be DE using qRT-PCR, presumably in the same direction as the microarray results, then that gene is believed to be validated. In other words, the testing procedure is two stages; the screening or pilot phase is performed using

microarray technology followed by a confirmatory or validation phase using qRT-PCR for only those gene-specific null hypotheses rejected in the first stage.

Two-stage tests and designs have previously been proposed for association studies (Lin, 2006; Satagopan and Elston, 2003; Satagopan et al., 2004). Those procedures control the family wise error rate (FWER) and have potentially lower power when a large number of tests are considered. More recently, two-stage designs controlling false discovery rates (FDR) have also been proposed for association studies (Zehetmayer et al., 2005). Miller (2001) described a very simple procedure to control the FWER in the second stage for gene expression studies.

In practice, a researcher may be interested in assessing the number of samples necessary to obtain certain power with a given FDR. This problem has been addressed for single stage experiments based on microarrays (Jung, 2005). Conversely, the power and experimental design of the confirmation phase in a two-stage test have not been addressed in the literature. In this paper, we propose methods for helping design two-stage experiments and provide guidelines for conducting the two-stage tests. We work under two possible design scenarios. In the first case, the total number of samples available is fixed and the main objective is to allocate samples to either the first stage (i.e. microarray) or the second stage (qRT-PCR). In the second case, the necessary number of samples to attain a given power for the second stage is determined conditional on a previous microarray experiment. We also investigate the ramifications of using independent samples as opposed to the same samples in both stages.

The structure of the paper is constructed as follows: In Section 2, we formulate the problem and elicit the theory for FDR control and power calculation in a two-stage

experiment involving multiple tests. Section 3 is divided into three parts. First, we assess the effects of the proportion of genes that are non-DE and the proportion of genes chosen to be validated in the second stage on overall sensitivity and FDR in the joint design of a screening and validation experiment. Second, we study these same effects for the design of a validation experiment upon completion of a microarray experiment. Finally we estimate effect sizes and correlation between expression measures on the same gene in an experimental dataset and demonstrate how to empirically infer upon sensitivity and FDR.

2. Methods

2.1 Type I error rate in a two-stage single test.

Consider an expression profiling experiment comparing just two experimental groups or treatments. Let X_{i1j} (X_{i2j}), with $j=1...n$ in both treatments, denote the measured expression level of a gene i in subject j within Group 1 (2).

We momentarily assume a known within-group variance σ_i^2 to be common to the two treatments and (log) expression levels to be normally distributed. Specifying d_i=E(X_{i1j})-E(X_{i2j}) as the true mean difference, the classical z-test statistic has distribution:

$$z_i^{(1)} = \frac{\bar{X}_{i1}^{(1)} - \bar{X}_{i2}^{(1)}}{\sigma_i^{(1)}\sqrt{\frac{2}{n_1}}} \sim \begin{cases} N(0,1), d_i^{(1)} = 0 \\ N(\delta_i^{(1)}\sqrt{\frac{n_1}{2}}, 1), d_i^{(1)} > 0 \end{cases}.$$

Here, the superscript index (1) indicates data generated under Stage 1 (microarray) such that $\delta_i^{(1)} = \frac{d_i^{(1)}}{\sigma_i^{(1)}}$ defines the effect size for Stage 1. Similarly, superscript index (2) pertains to data derived from Stage 2 (qRT-PCR).

The joint distribution of the first stage and second stage test statistics is specified as bivariate normal with correlation ρ^*:

$$\begin{bmatrix} z_i^{(1)} \\ z_i^{(2)} \end{bmatrix} \sim N\left(\begin{bmatrix} \delta_i^{(1)}\sqrt{\frac{n_1}{2}} \\ \delta_i^{(2)}\sqrt{\frac{n_2}{2}} \end{bmatrix}, \begin{bmatrix} 1 & \rho^* \\ \rho^* & 1 \end{bmatrix} \right).$$

Now if independent samples are used for the first and second stage testing, then $\rho^*=0$; however, if the same samples are used in both stages, we would expect $\rho^*>0$ to be a monotonic function of the correlation in gene expression measurements between the two techniques.

In general, the overall Type I error rate α for any particular gene over the two-stage test can be readily determined as

$$\alpha = \int_{Zc_1}^{\infty} 1 - \Phi\left(\frac{Zc_2 - \rho^* z_1}{\sqrt{1-\rho^{*2}}} \right) \varphi(z_1)\, dz_1 , \qquad [1]$$

where $Zc_1 = \Phi^{-1}(\alpha_1)$ and $Zc_2 = \Phi^{-1}(\alpha_2)$ are the critical values for declaring a gene to be DE at the first and second stage tests respectively, based on stage-specific Type I error rates α_1 and α_2, respectively. Under a true alternative hypothesis, the power (1-β) of the two-stage test is:

$$1-\beta = \int_{Zc_1}^{\infty} 1-\Phi\left(\frac{Zc_2 - \left(\delta_i^{(2)}\sqrt{\frac{n_2}{2}} + \rho^*\left(z_1 - \delta_i^{(1)}\sqrt{\frac{n_1}{2}}\right)\right)}{\sqrt{1-\rho^{*2}}}\right)\varphi(z_1)\,dz_1 \,. \qquad [2]$$

We momentarily assume that different samples are used for each stage. Note that under these circumstances, $\alpha = \alpha_1\alpha_2$ and $(1-\beta) = (1-\beta^{(1)})(1-\beta^{(2)})$.

2.2 FDR and sensitivity in two-stage tests.

Table 4.1 delineates all possible combinations of outcomes of the true state of nature with conclusions drawn from statistical analysis using a two-stage gene expression profiling experiment. The first stage (i.e. microarray) is conducted using the same n_1 experimental units per group from each treatment for all $m = m_0 + m_1$ hypotheses or genes, m_0 of which are non DE and the remaining m_1 are then DE. Hence a proportion $\pi_0 = m_0/(m_1+m_0)$ of the m gene specific hypotheses is truly null (non-DE). The first stage rejection set involves the $R^{(1)}$ hypotheses declared significant by a statistical test on the first stage experiment; these genes are subsequently tested in a second stage (i.e. qRT-PCR) using a set of n_2 samples per group being independent from the first n_1 samples. In other words, a total of $n = n_1 + n_2$ biological replicates per treatment are used in the two stage study. Among these $R^{(1)}$ genes, there are $R^{(2)}$ null hypotheses rejected again in the second stage thereby creating a list of putatively validated DE genes. The remainder of

the second stage tests ($A^{(2)}$ hypotheses or genes) are then generally concluded to be non-DE. Now the correctly accepted hypotheses (true negatives) across the two stages sum to $A_0^{(1)} + A_0^{(2)}$ whereas the number of incorrectly accepted hypotheses (false negatives) sum to $A_1^{(1)} + A_1^{(2)}$; again, here the superscripts characterize the testing stage. Similarly, the number of correctly rejected second stage (true positives) tests is $R_1^{(2)}$ whereas the number of incorrectly rejected second stage hypotheses (false positives) is $R_0^{(2)}$. Among the $R^{(1)}$ samples tested in the second stage, $R_0^{(1)}=A_0^{(2)}+R_0^{(2)}$ determines the number of true negatives whereas $R_1^{(1)}=A_1^{(2)}+R_1^{(2)}$ is the number of rejected hypotheses. Hence, the first stage false discovery rate ($FDR^{(1)}$) is defined by the proportion $R_0^{(1)} / R^{(1)}$. Among the various definitions of power for multiple tests, the sensitivity is probably the most meaningful in microarray experiments (Pawitan et al., 2005). Table 4.2 illustrates that applying a second stage test will generally decrease the FDR and the sensitivity. Nevertheless, the allocation of the n samples to first stage (n_1) and second stage (n_2) should be optimized in order to attain the maximum sensitivity on identifying DE genes across the two stages.

To control the FDR and maximize the sensitivity we assume independence of the test statistics and define the following two stage expressions:

$$FDR^{(2)} = \frac{\pi_0 \alpha}{\pi_0 \alpha + (1-\pi_0) Sensitivity^{(2)}}, \text{ and} \qquad [3]$$

$$Sensitivity^{(2)} = \frac{\sum_{j \in m_1} (1-\beta_j)}{m_1}, \qquad [4]$$

From [4], the sensitivity is the average power of the single (non-null) tests and represents the proportion of DE genes that we expect to detect with the two-stage experiment. A two stage design for expression profiling should only have a sensitivity advantage over the single stage design (i.e., using only microarrays) if the effect size of the second stage tends to be larger that the effect size of the first stage for any particular gene. Moreover, for a given sample partition (n_1:n_2) between the first and second stages, the maximum sensitivity of the two stage design will always correspond to the largest possible $\alpha^{(1)}$. However, the largest value of $\alpha^{(1)}$ is generally constrained by the maximum number of first stage rejections (declared DE genes) that can be further validated in the second stage because of budget limits. For example, if there are 4000 hypotheses tested with a microarray experiment and at most 40 of those could be validated using high throughput qRT-PCR, then the value of $\alpha^{(1)}$ should be determined such that

$$\frac{R^{(1)}}{m} = \pi_0 \alpha^{(1)} + (1-\pi_0)\frac{\sum_{j \in m_1} (1-\beta_j^{(1)})}{m_1} \leq 0.01 . \qquad [5]$$

The second stage Type I error rate, $\alpha^{(2)}$, for statistical significance should be chosen such that the FDR in Equation [2] is controlled at a desired level. Finally, for a given total sample size n, the optimization should be done over the alternative sample partitions (n_1:n_2).

3. Results

3.1 Sample size and sample allocation.

3.1.1 Known variance and constant effect sizes across independent genes.

The question of how many samples are needed to simultaneously control the FDR and attain a certain level of specificity with single-stage microarray experiments has been addressed previously (Jung, 2005). Here we optimize the design of a two-stage experiment by applying such a method to estimate the number of samples per group in a single stage experiment. Subsequently, we maximize equation [4] with respect to the sample partitions and $\alpha^{(1)}$ subject to restrictions in FDR and proportion of genes selected in equations (Recinos et al.) and [5].

In summary, the two-stage design procedure proceeds as follows:

1) Specify the input parameters

a) f=maximum FDR level

b) m= total number of hypotheses

c) s= minimum target sensitivity

d) $R^{(1)}/m$= maximum proportion of hypotheses declared DE in the first stage to be validated in the second stage

e) π_0= proportion of hypotheses that are null

f) $\delta_j^{(1)}$= First stage effect sizes

g) $\delta_j^{(2)}$= Second stage effect sizes

2) Obtain overall sample size, n, for a single stage design, based on 1a) through g)

3) Apply the optimization over all possible partitions of n in n_1 and n_2.

Let's consider a couple of cases:

Case 1. f=0.05, m=4000, π_0=0.99, s=0.6, $\delta_j^{(1)}$=1. Using equation (8) in (Jung, 2005) along with n=28. $\delta_j^{(2)}$=1.4, $R^{(1)}/m$= 0.01 (i.e., maximum of 40 genes to be validated), the optimal sample allocation is n_1=21 replicates per treatment in the first stage and n_2=7 replicates per treatment in the second stage. With these design parameters, the sensitivity

of the two-stage design and the single stage design are both very close to 0.6. Hence, there would be little advantage in using a two-step design in this situation.

Case 2. If twice as many hypotheses were to be validated as in Case 1, i.e., $R^{(1)}/m$= 0.02 with all other input parameters being identical, the optimal sample allocation would then be: n_1=18 samples for the first stage and n_2=10 samples for the second stage. Furthermore, the sensitivity for the two-stage design would be 0.7 thereby substantially exceeding the sensitivity of 0.63 for the single stage design.

Using the same procedure, the optimal sample allocation and resulting sensitivity for a two-stage experiment for different proportions of genes to be validated (i.e. $R^{(1)}/m$) in the second stage are presented in Figure 4.1. In general, most of the samples are allocated to the first stage (Figure 4.1a), but this allocation changes with the proportion of declared DE first stage hypotheses to be validated and the effect size ratio ($\delta^{(2)}:\delta^{(1)}$). As more declared DE hypotheses are allowed to be retested in the second stage, less samples are proportionately assigned to the first stage. That is, if only a limited number of hypotheses are retested in the second stage, the first stage needs to be as sensitive as possible in order to carry over as many of the truly DE genes in the second stage. With respect to the consequence of the effect size ratios, the minimum proportion of samples allocated to the first step occurred at $\delta^{(2)}:\delta^{(1)}=1.4$ regardless of the values for the proportion of genes to be carried over from the first stage. From Figure 4.1b, it is further evident that the proportion of all genes to be carried over the second stage should be at

least equal to the proportion of genes that are DE in order to have greater sensitivity with the use of a two-stage design. For example, if 40 out of 4000 hypotheses are true positives ($\pi_1 = 0.01$) and we choose to validate only 20, the sensitivity could never exceed 0.5.

3.1.2 Varying effect sizes

In the previous section, a constant effect size was assumed for each gene within both stages; however, we know that this assumption is generally not valid for gene expression studies. Subsequently, we studied three different distributions for effect sizes as illustrated in Figure 4.2: a) constant, b) symmetric and c) asymmetric decreasing.

Figures 4.3 and 4.4 display the optimal sample allocation and sensitivity determinations based on two different across-gene means for effect size in the first stage $\mathrm{E}\left(d_j^{(1)}\right) = 0.4$ and $\mathrm{E}\left(d_j^{(1)}\right) = 1.0$ for each of the three distribution types. In general, a constant effect size for the first stage always had greater sensitivity (Figures 4.3 and 4.4 b). However, the sample allocation strategy changes only marginally between the three distribution types, especially when there is a large number of samples (Figure 4.4 a). With a larger effect size and hence smaller number of replicates to attain the desired power, the proportion allocated to the first stage appeared to be more sensitive to the distribution of the first stage effect size, but this was only attributable to the reallocation of one or two samples per treatment.

We already demonstrated that different fixed ratios of the second stage to first stage effect sizes across all genes influenced optimal sample size allocations between the two stages (Figures 4.1, 4.3 and 4.4). We also decided to investigate variability in these

ratios based on similar distribution specifications as considered for first stage effect size (Figure 4.5)

It should be noted, however, that these ratios had no impact on the optimal sample size allocation and overall power as indicated in Figure 4.6.

3.1.3 Unknown variances.

Thus far, we have assumed a known variance in all of these tests (i.e. *z*-tests), but the procedure can be extended to a *t*-test for the more typical cases where variances are not known and must be inferred from the data. Figure 4.7 shows the optimal sample size allocation and sensitivity for various effect size ratios and proportion of genes to be validated when the variance is assumed unknown. We adjusted the overall n to maintain the same target minimum sensitivity level in the single stage case. For example, for $\delta^{(1)}=1$, for the t-test we have $n=30$ while for the z-test we needed $n=28$ for the same effect size using a *z*-test. In general, the sample size allocation is very similar to that obtained in the known variance case, but a slightly smaller proportion is assigned to the first stage.

3.1.4 Implementing two stage tests.

In order to determine the optimal sample allocation between the two stages, the optimization procedure used in the previous sections also yielded the optimal significance level cutoffs in both stages; i.e. $\alpha^{(1)}$ and $\alpha^{(2)}$. Zehetmayer et al. (Zehetmayer et al., 2005) proposed to use $\alpha^{(1)}$ calculated from equation [5] assuming a priori values for effect sizes and π_0 and then set $\alpha^{(2)}$ such that the desired FDR is attained. Such a strategy could

certainly be used in our work with the additional restriction that the proportion of hypotheses selected from the first stage is not larger than a specified limit. Alternatively, a fixed proportion could be selected in the first stage and $\alpha^{(2)}$ then set to control the FDR. Furthermore, the estimation of the FDR in the first and second stages requires the specification of π_0, the proportion of all hypotheses that are truly null. At present, there are several methods to estimate π_0, including the procedure proposed by Storey (2002) that we subsequently use. The estimation of π_0 is more critical in the second stage where only a limited number of hypotheses are tested. In such situation, the FDR from the Stage 1 is an estimate of the π_0 in the validation set.

In this section we use a simulation study to assess the properties of both testing procedures based on an optimal design calculated in the previous section. For all cases, variances were assumed to be unknown such that inferences are based on *t*-tests. We use four procedures to determine $\alpha^{(1)}$ and $\alpha^{(2)}$ in order to control FDR:

1) Estimates of $\alpha^{(1)}$ and $\alpha^{(2)}$ from the optimization procedure.

2) Estimates of $\alpha^{(1)}$ from the optimization but restricted to a maximum number of rejections and $\alpha^{(2)}$ set to control FDR. Furthermore, π_0 is estimated from the validation experiment.

3) Similar to 2) but π_0 is estimated from the FDR of the first stage.

4) Fixed proportion of hypotheses selected in the first stage with second stage similar to 3).

The design was optimized for the following input parameter values: π_0=0.99, f=0.05, $\delta_j^{(1)}$=1, n=30, $R^{(1)}/m$=0.01. The optimization resulted in 23 samples for the first stage and 7 samples in the second stage. The true FDR and sensitivity of the two-stage test are presented in panels a) and b), respectively, of Figure 4.8.

All testing methods provided good control of the FDR and similar sensitivity. The first two procedures, however, are not expected to perform well (control FDR and give maximum sensitivity) if the actual parameters are different from the values assumed in the optimization because they do not account for uncertainty in the values of effect sizes and π_0.

3.1.5 Correlated genes.

The optimization has been performed assuming that the expression of the genes is not correlated; i.e. expression of each gene is independent from each other. Therefore, we used simulation to further investigate the sensitivity obtained with the selected design when the statistical tests are correlated between genes. We specified a block-diagonal compound symmetry correlation structure between all genes based on individual blocks of either size 10 or 40 genes and within-block correlation coefficients of either 0.4 or 0.75. Figure 4.9 shows the sensitivity and FDR of the two-stage procedure for these specifications.

It can be noted from Figure 4.9 that, on average, the FDR is controlled at the nominal level and the target sensitivity is obtained, but the variance of FDR and sensitivity increases with the correlation.

A set of simulation studies was run over alternative partitions of n_1:n_2 to evaluate the sample optimal allocation under correlated tests (Figure 4.10). For a moderate correlation structure (block size=10, ρ=0.4), the optimal partition coincided with that obtained in the optimization assuming independent tests. However, for a more extreme correlation (block size=40, ρ=0.4)), the optimal sample partition was slightly different, assigning one more replication to the first stage. In practice these differences may not be relevant and the actual gain in sensitivity might be minimal for low to moderate n_1.

Even for highly correlated gene expression, the optimal sample partition (n_1=*26*, n_2=*4*) is very close to the sample partition assuming uncorrelated tests (n_1=*23*, n_2=*7*). Moreover, any of these sample partitions yielded practically the same sensitivity, i.e., very close to the target value of 0.6.

3.2 Design of a validation experiment.

3.2.1 Sample size calculation of an independent validation experiment.

We previously considered the joint design of a screening and validation experiment. A simpler case may be the optimization of a validation experiment upon completion of an observed screening experiment. Several authors have addressed the issue of sample size calculation in microarray experiments (Hu et al., 2005; Jung, 2005).

Jung (2005) has proposed a simple algorithm for sample size calculation to attain a target FDR and sensitivity that could be easily applied to the design of validation experiments. Given a list of putatively differentially expressed genes from the microarray, the associated FDR is an estimate of the proportion of null hypotheses in the set. Moreover, from the microarray data, an estimate of the effect sizes for each gene could be obtained. With these elements and given a target FDR and sensitivity for the validation experiment, the required sample size can be readily computed.

Case 3:

A microarray experiment yielded a list of 25 genes declared DE, with an overall FDR of 0.20. The average effect size was δ=1.0. The target sensitivity of the validation experiment was 0.9 and the FDR was specified to be either 0.01 or 0.05. Following Jung (2005), the required sample size is $n_2=10$ or $n_2=20$ to control FDR at levels of 0.05 or 0.01, respectively. Moreover, to attain such FDR levels, we expect that the cutoffs for declaring statistical significance for the validation experiment to be $\alpha_2= 0.0364$ (FDR=0.01) or $\alpha_2=0.19$ (FDR=0.05).

These values in Case 3 are calculated based on the assumption that the samples used in the validation experiment are independent from those used in the microarray experiment. Also, note that if the traditional comparison wise level of significance cutoffs α_2=0.01 or α_2=0.05 are used (for fixed sample size), the sensitivity is actually expected to be smaller than the intended value.

Figure 4.11 presents the sample size (n_2) per treatment needed to validate an experiment with 25 genes declared to be DE at different levels of FDR and average different effect sizes. Here n_2 is small compared to the number of replicates (n_1) necessary for a microarray experiment, particularly for small FDR values. As expected, the number of samples increases if the FDR from the microarray increases or if the average effect size decreases.

3.2.2 Non-independent sample sets in screening and validation experiments.

Thus far we have assumed that different samples are used in the validation. But due to resource constraints, the same samples are often reused for the qRT-PCR validation experiment. Such validation might more properly be called technical validation. If the same samples are used, the two test statistics (based on the microarray and qRT-PCR data) on each gene are not independent and this should be accounted for when determining $\alpha^{(1)}$ and $\alpha^{(2)}$.

The correlation among the tests (ρ^*) will depend on the correlation ρ among repeated measures (i.e., correlation of the microarray and qRT-PCR expression measures on the same sample) and the number of samples in common in the two experiments (n_c); that is, it can be readily shown that $\rho^* = \rho \cdot \sqrt{\frac{n_c}{n}}$. Consequently, if the same biological samples are used for both tests, the correlation of the tests is the repeated measure correlation.

In the presence of correlated tests, the actual type I error rate of the two-step test, needs to be calculated with equation [1]. Alternatively, given a nominal type I error level of the first step (α_1) and a target overall α, $\alpha^{(2)}$ for the validation step can be obtained by solving equation [1] for Zc_2. The power of this test can subsequently be obtained from equation [2].

Case 4.

For a certain gene, the correlation ρ among repeated measurements with microarray and qRT-PCR is 0.66. If $\alpha^{(1)}$ =0.05 and the overall type I error rate required is α=0.0025, the second stage p-value cutoff $\alpha^{(2)}$ should be 0.0037898. On the other hand, if ρ =0.56, then $\alpha^{(2)}$ should be equal to 0.0051.

Figure 4.12 illustrates the power of a single test for n_1=n_2=25 and $\delta^{(1)}$=0.6 in the first stage for three different scenarios of correlation between measures of the same gene with different technologies (ρ =0, ρ =0.56, and ρ =0.66).

For a large relative effect size in the second stage, the power of the test is almost identical regardless of the degree of correlation in the two measures. But for moderate effect size ratios, there is a clear advantage in using independent samples for both stages.

The FDR of the two stage tests also depends on the correlation between sequential test statistics. In the previous case, if we assume for example ρ^*=0.66, $\delta^{(1)}$=0.6 and $\delta^{(2)}$=0.72, using the same value for $\alpha^{(1)}$ as for correlated tests compared to independent

tests, the FDR sensibly increases (Table 4.3). The value for $\alpha^{(2)}$ can also be adjusted in correlated tests to attain an overall comparison wise error rate equivalent to the independent test. In that case the FDR is only slightly larger than the FDR for independent tests, but the power is reduced from 0.57 to 0.42. Finally, adjusting $\alpha^{(2)}$ to obtain the same FDR as the uncorrelated test, yields a further reduction in power in the correlated tests, to a value $1-\beta$=0.37.

3.3 Estimated effect sizes and correlation for microarray and qRT-PCR assays.

Throughout this paper, it is assumed that effect sizes from different technologies are known. For illustration purposes, effect sizes in the range of 0.4 to 1.5 are commonly used in the literature, but in practice, a pilot dataset might be used to estimate these parameters for a similar future experiment. In this section, we estimated effect sizes using an experimental dataset available from the pubic domain (Perreard et al., 2006). The original dataset consisted of 123 samples of which 94 non-metastasised tumor samples were included in the study where the absolute expression level of 53 genes was measured using a qRT-PCR assay. Expression levels of the corresponding genes in the same samples were retrieved from a more comprehensive microarray study (GSE2607 at http://www.ncbi.nlm.nih.gov/geo/) The sample was divided into two groups according to estrogen receptor status (positive, $n=49$ and negative, $n=45$). The expression values were normalized relative to the arithmetic mean of three housekeeping genes (MRPL19, PSMC4, and PUM1). A summary of the distribution of the effect sizes for both technologies is presented in Figure 4.13.

The overall effect size ratio was slightly larger than 1.0 for the expression measured by qRT-PCR with respect to microarrays. In fact, for some cases, this ratio was as high as 1.5 whereas for some genes the effect size estimated from microarrays was somewhat larger than the estimated effect size for qRT-PCR. The average effect size from both technologies was just below 1.0, indicating a moderately large effect size as compared to other reports from gene expression studies (Pawitan et al., 2005). This result is not surprising considering that the genes were selected based on previous knowledge of their functional classification.

The effect sizes are a function of the experimental designs and different experimental layouts will produce different effect size estimates because of differences in efficiencies. In order to generalize these results to other designs, we need to specify the relative efficiency of the two designs (Timm, 2004). For example, suppose that design A has an effect size ES_A=1.0 and that design is 1.1 times more efficient than design A, the expected effect size of design B is ES_B=1.0×1.1.

A bivariate linear model was used to estimate the residual correlation between the two technologies for each gene. The median correlation coefficient was 0.66 and the central 50% of the estimated values was between 0.56 and 0.79 (Figure 4.14). This high level of correlation is on the order of the values used in the preceding section on correlated tests.

Generalizing the estimate of the correlation to other experimental layouts is not as straightforward as it was for effect size estimates as the task requires dissecting the technical and biological sources of variation associated with each gene expression assay platform.

4 Discussion

In this study, we proposed the use of two-stage tests for expression profiling experiments to couple results from microarrays and qRT-PCR assays. The proposal consists of using the microarray experiment as a screening step, selecting a reduced number of genes that are subsequently assayed using qRT-PCR. Two stage designs can be planned a priori by splitting a sample set for usage within the two stages and defining the number of hypothesis to carry from the first to the second stage. Previous studies controlled the experiment wise type I error rate (Satagopan et al., 2004; Satagopan et al., 2002), but more recently, the control of FDR in two stage experiments has been considered for association studies (Zehetmayer et al., 2005). Recently, the use of sensitivity has been proposed as a measure of power in multiple tests (Pawitan et al., 2005).

Our results showed that the sensitivity of a two stage experiment at a fixed FDR will depend on the sample size, the value and distribution of effect sizes in the first and second stage, and the proportion of genes selected in the first stage for subsequent testing in the second stage. For attaining a larger sensitivity in a two stage test compared to a single stage test, the proportion of genes selected should be equal or larger than the expected proportion of differentially expressed genes and the effect sizes of the second stage should be larger than the effect sizes in the first stage. We focused most of our calculations in the particular case where the proportion of selected genes is equal to the expected proportion of true alternative hypotheses. Pawitan et al. (2005) showed that in this circumstance, for a π_0=0.99, the FDR is expected to be very high, usually larger than 0.2. In our case, however, this is not completely detrimental since the second step is

designed to eliminate the false positives and lower the overall FDR to an acceptable level. Conversely, selecting a small proportion of genes in the first stage for further second stage validation will impose a constraint in the maximum sensitivity that can be attained.

Our results indicate that in general, more samples should be assigned to the first stage than to the second stage to maximize sensitivity. That is, generally 60 to 80 % of all biological samples should be utilized for the microarray experiment. Lower numbers of first stage replicates would be recommended for situations where a large proportion of genes is selected for further validation and the relative effect size for the second stage is around 1.4 times larger than the effect size in the first stage. We also showed that the relative sample allocation is similar for different combinations of total numbers of samples available, for different effect sizes, and for different distributions of the effect sizes. The magnitude and distribution of the effect sizes however, do affect the overall sensitivity and minimum sample size necessary to attain a certain target FDR and sensitivity which is in accordance with results from single stage experimental designs (Jung, 2005).

In certain studies, including a large number of arrays may be cost prohibitive. If less arrays than the optimal number are used, the sensitivity of the experiment will be reduced. A way to mitigate this is to use a more liberal p-value cutoff and include more genes in the validation study (simultaneously increasing sample size), but such strategy will also increase the overall cost of the experiment. Reducing sample size in the first stage while keeping a stringent significance criteria can not be compensated by increasing sample size in the second stage because the overall power of each test is (for independent

tests) the product of the first stage and second stage power. Moreover, in the optimized designs the power of the validation step is already close to one for all genes.

The overall sample allocation between the two stages is similar assuming known or unknown variances, but the total sample size is increased if there is uncertainty about the variance, especially if large effect sizes are expected. Under these circumstances, reduced sample sizes will lead to different total sample size if a *t*-test (unknown variances) is assumed instead of a *z*-test (known variances). Similar to the case of single stage experiments (Jung, 2005), the differences are minimal for effect sizes of the order of 1.0 or smaller.

We also compared different alternatives to implement the two-stage test. Our results suggest that an effective strategy is to select a fixed proportion of genes in the first stage and then estimate the p-value cutoff for statistical significance in the second stage to control the FDR at a desired level. In that case, the FDR of the selected genes from the first stage is a suitable estimate of the proportion of null hypotheses selected from the first stage. Interestingly, the p-value cutoffs for the second stage are much higher than the traditional 0.05 and 0.01 values commonly employed in validation experiments. The reason for this is that the proportion of null hypotheses in the validation set is usually small, i.e. lower than 0.5.

In the second section of this study, we addressed the issue of designing a validation experiment conditional on an existing microarray study and the use of the same samples in the validation experiment (i.e. technical validation). Design optimization of a validation experiment has not been considered before. Yet we have shown that if a sufficiently large list of genes is being validated, the general principles of design of

microarray experiments when controlling FDR (Jung, 2005) could be applied. Moreover, in that case, the microarray experiment can be used as a pilot study from which the proportion of differentially expressed genes in the validation set and effect sizes could be estimated. In general, if the FDR of the microarray experiment is small and the effect size is large, few additional samples and a liberal p-value cutoff for statistical significance would be sufficient for the second stage validation experiment.

Reusing the same samples in the validation experiment is called technical validation. This practice, though not recommended, is common (Allison et al., 2006). One reason for this protocol is the fact that some transcriptional profiling experiments are conducted using samples stored in the past from previous experiments and re-conducting the validation experiment with new samples might not be feasible. We have shown that given a false rejection (false positive) in the first stage, the chances of repeating this same decision error in the second stage would be higher when using the same samples. In order to control the overall type I error rate in these cases, the p-value cutoff for declaring statistical significance in the second stage would need to be reduced. Reducing this cutoff, however, will produce a less powerful test. In order to calculate these cutoffs, the correlation between the tests, which in turn is a function of the correlation between sequential measures in the same sample, needs to be determined.

Lastly, we used a publicly available dataset to estimate effect sizes and correlation of microarray and qRT-PCR based expression measures. We found that for this particular dataset, the effect size of the selected genes (declared as differentially expressed) was of the order of 1.0. The average ratio of effect size from qRT-PCR to microarray was close to 1.6 but showed a large variability with some genes having ratios

as low as 0.6 and more than half of the genes have ratios larger than 1.0. Also the repeated measures correlation was very high, with an average value of 0.64 and a more than half of the genes showing a correlation larger than 0.7. These results are specific to this particular experiment, but are within the range of parameters considered in our study.

TABLE 4.1 Outcomes of $m=m_0+m_1$ two-stage tests. A represents accepted hypotheses, R represents rejected hypotheses, the sub index 0 indicates true null hypotheses and 1 indicates true alternative hypotheses. The super index indicates the stage of the test.

	Conclusions from first stage test			
TRUE STATE OF NATURE	$A^{(1)}$	$R^{(1)}$ *Conclusions from 2nd stage test*		*Total*
		$A^{(2)}$	$R^{(2)}$	
H_0	$A_0^{(1)}$	$A_0^{(2)}$	$R_0^{(2)}$	m_0
H_1	$A_1^{(1)}$	$A_1^{(2)}$	$R_1^{(2)}$	m_1

TABLE 4.2 Error measures in two stage tests based on the quantities presented in Table 4.1.

Error measure	*Expression*	
	First Stage	*Two-Stage*
Sensitivity	$\frac{A_1^{(2)} + R_1^{(2)}}{m_1}$	$\frac{R_1^{(2)}}{m_1}$
FDR	$\frac{A_0^{(2)} + R_0^{(2)}}{A_0^{(2)} + R_0^{(2)} + A_1^{(2)} + R_1^{(2)}}$	$\frac{R_0^{(2)}}{R_0^{(2)} + R_1^{(2)}}$

Table 4.3. Sensitivity and FDR for two stage tests based on correlated test statistics. $\pi_0=0.99$, $\delta^{(1)}=0.6$, $\delta^{(2)}=0.72$, $n_1=n_2=25$.

ρ	Nominal α_2	α	Sensitivity	FDR
0.00	0.05	0.0025	0.57	0.31
0.66	0.05	0.0179	0.64	0.73
0.66	0.0038	0.0025	0.42	0.38
0.66	0.0022	0.0016	0.53	0.31

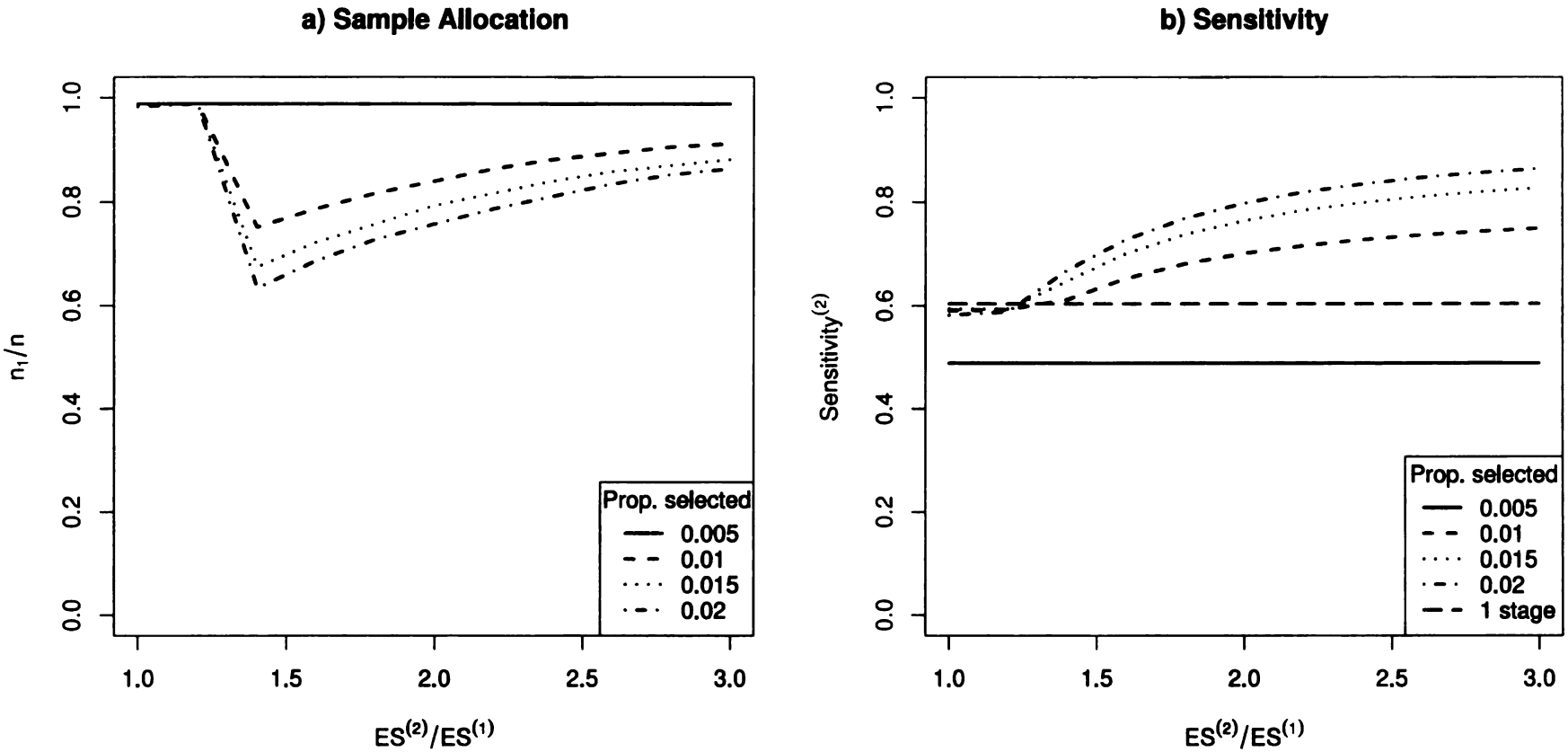


Figure 4.1. Sample allocation and sensitivity of optimal designs as a function of the relative effect size between the two stages. The lines represent different proportions (0.005, 0.01, 0.015, 0.02 and 0 =1 stage) of selected hypotheses to be carried to the second stage. The results were calculated for: π_0=0.99, f=0.05, $\delta_j^{(1)}$=0.4, and n=170 (see text).

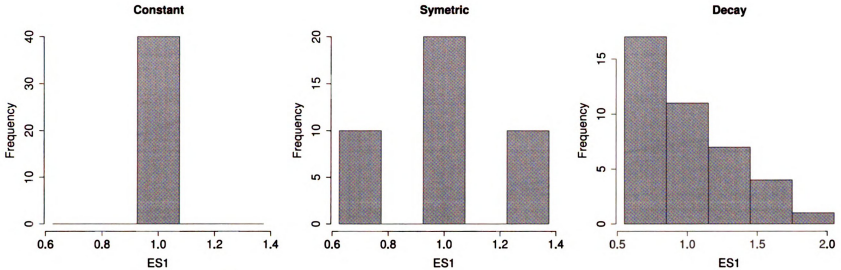


Figure 4.2. Distribution of the effect size for stage 1. a) constant, b) symmetric. c) decreasing. The three distributions were scaled to have mean 1.0 or 0.4.

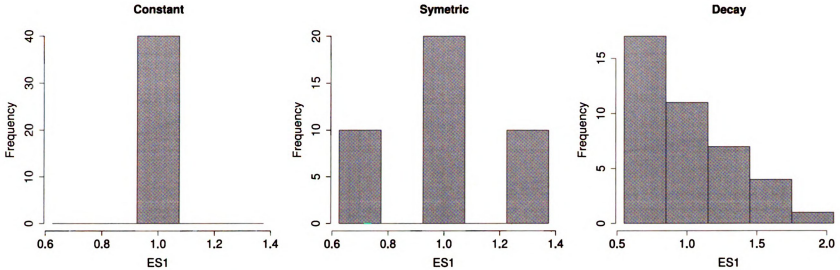


Figure 4.2. Distribution of the effect size for stage 1. a) constant, b) symmetric. c) decreasing. The three distributions were scaled to have mean 1.0 or 0.4.

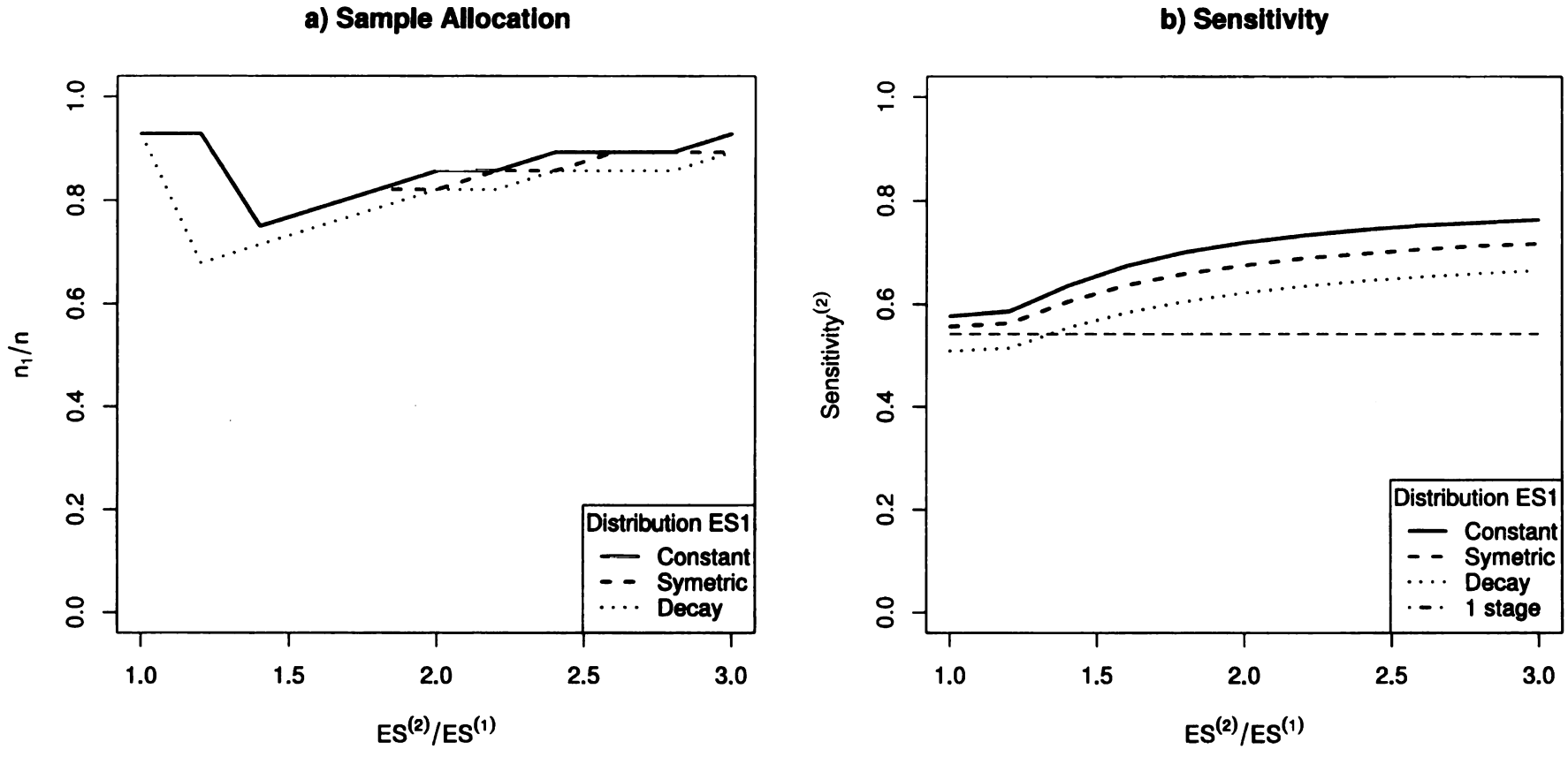


Figure 4.3. Sample size allocation and sensitivity as function of the distribution of the first and second stage effect size, average $\delta^{(1)}=1$, $n=28$. $\pi_0=0.99$, $f=0.05$. The single stage sensitivity corresponds to a decaying distribution.

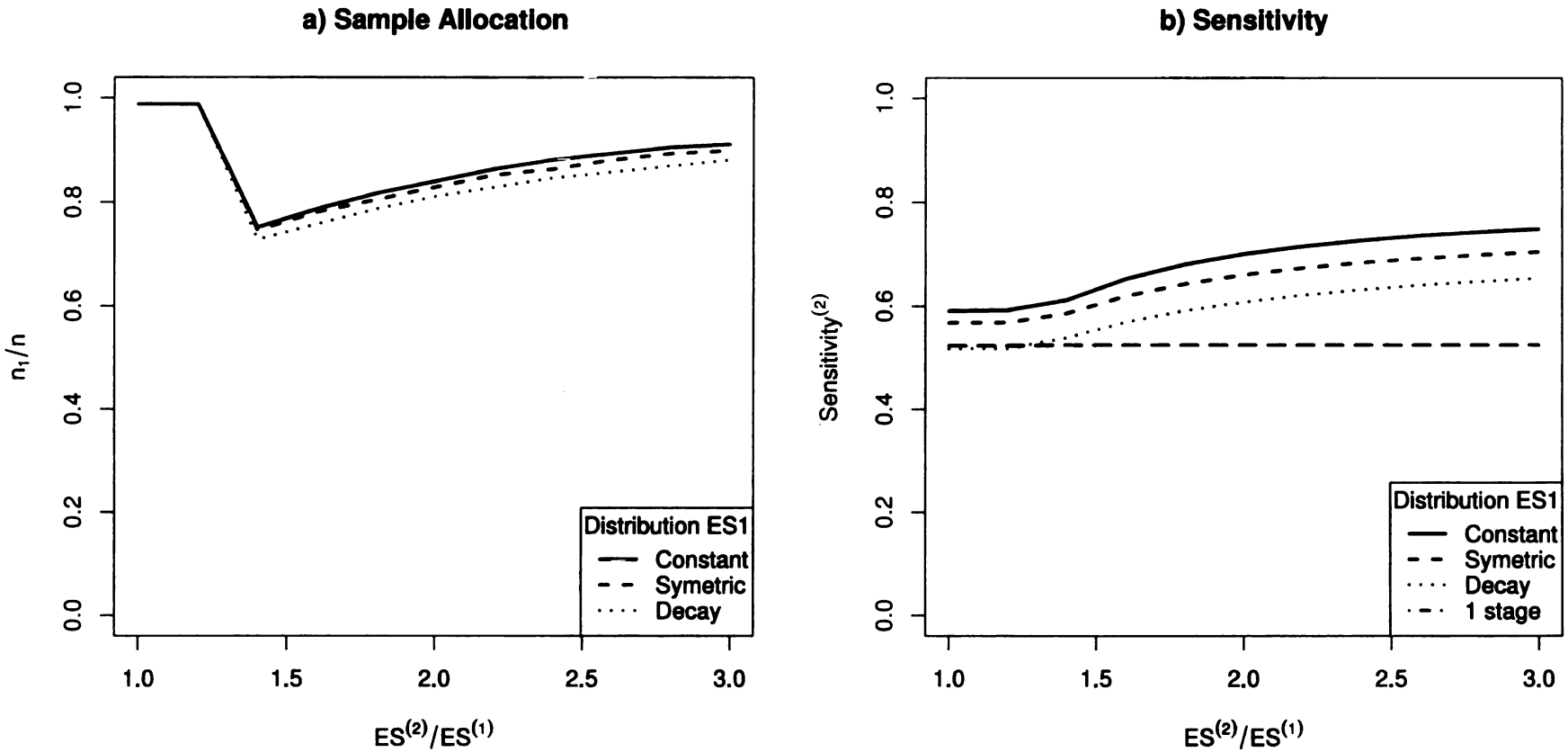


Figure 4.4. Sample size allocation and sensitivity as function of the distribution of the first stage effect size, average $\delta_j^{(1)}$=0.4, n=170 (c,d). π_0=0.99, f=0.05. The single stage sensitivity corresponds to a decaying distribution.

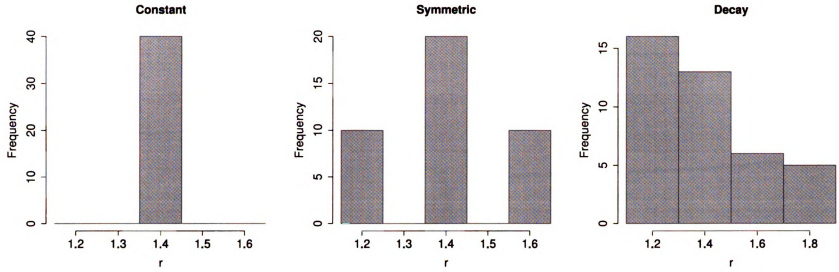


Figure 4.5. Distribution of the ratio of the effect size in the first stage relative to the effect size in the second stage.

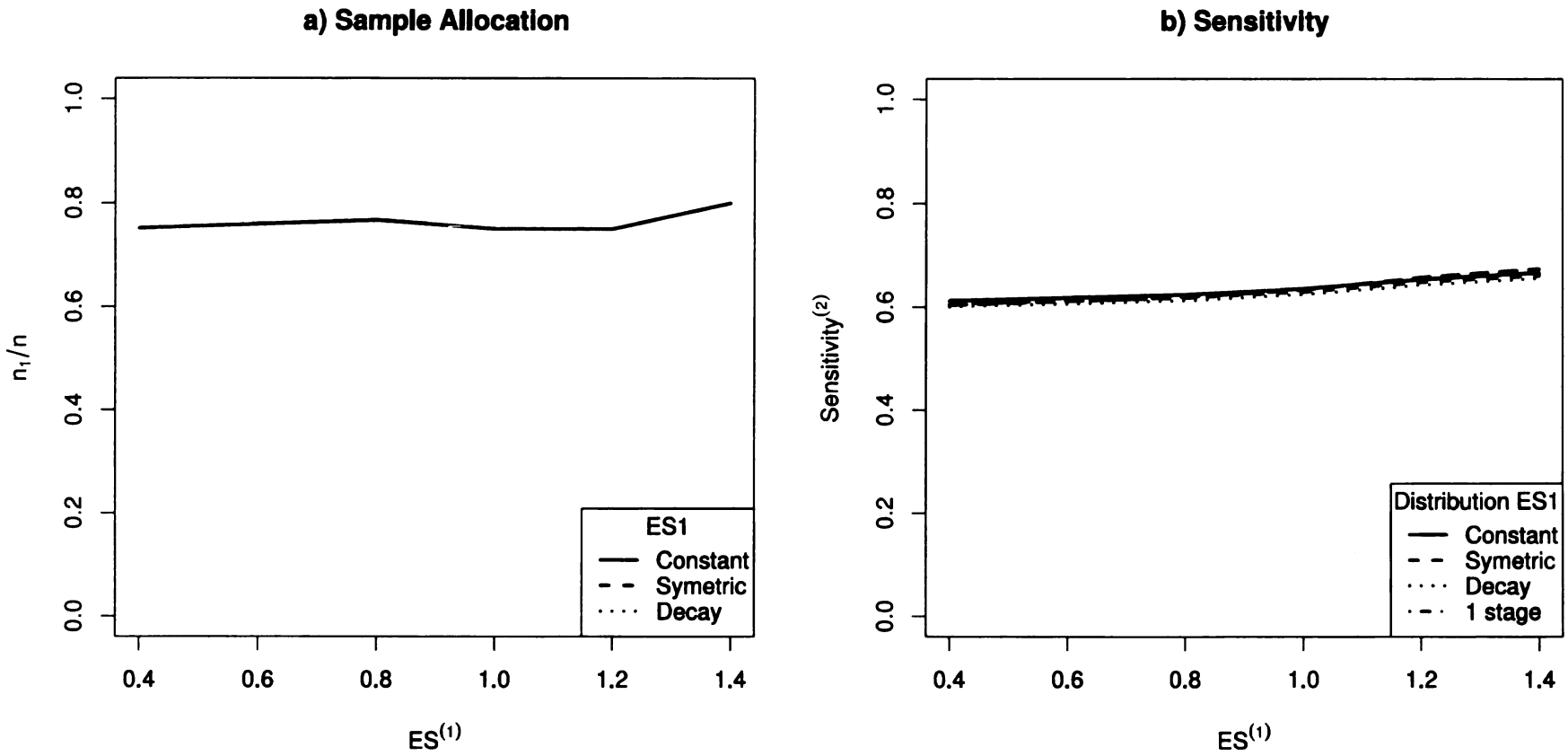


Figure 4.6. Sample size allocation (a) and sensitivity (b) as function of the distribution of the first stage effect size for different effect size ratios (lines). π_0=0.99, f=0.05.

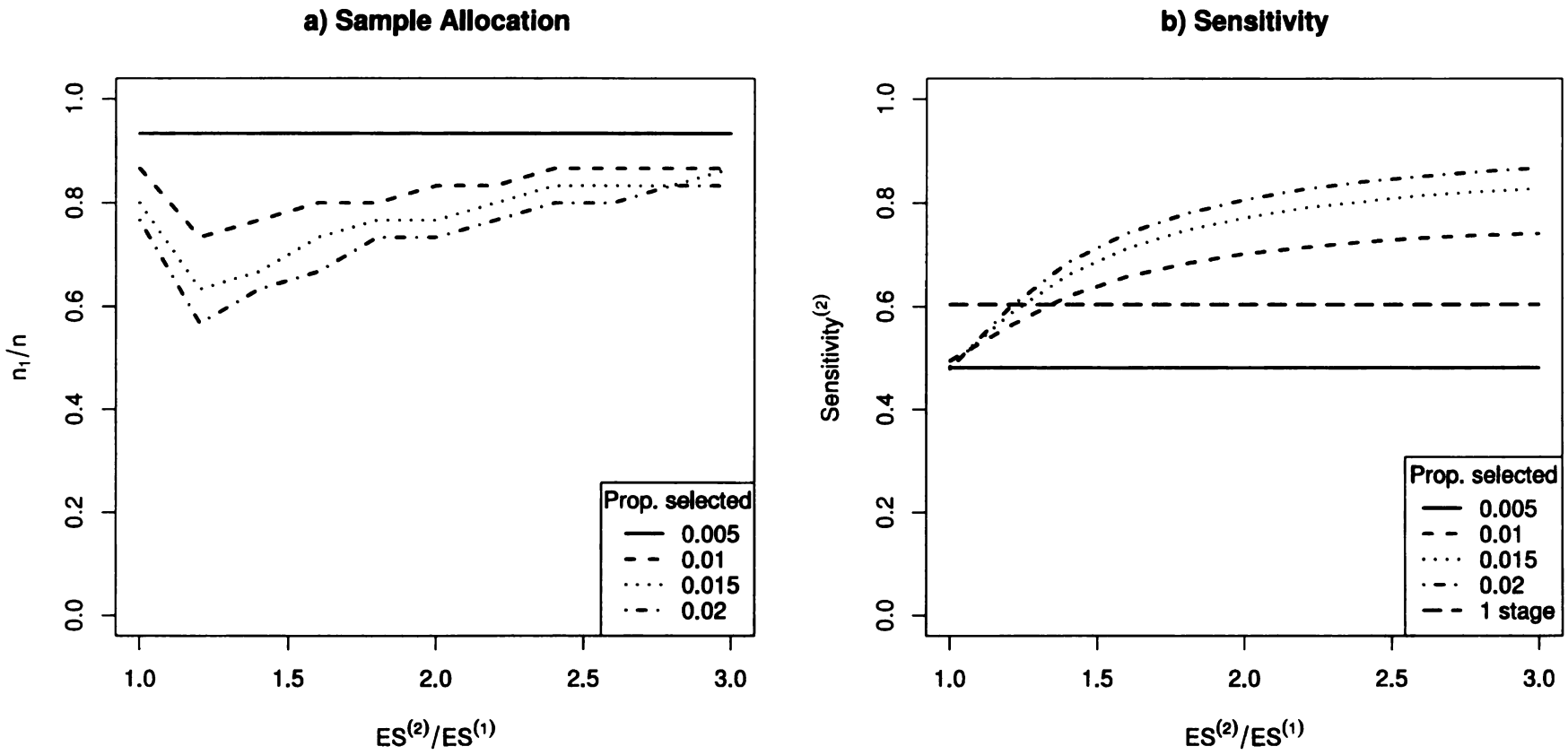


Figure 4.7. Sample allocation and sensitivity of optimal designs as a function of the relative effect size. π_0=0.99, f=0.05, $\delta_j^{(1)}$=1, n=30.

a)

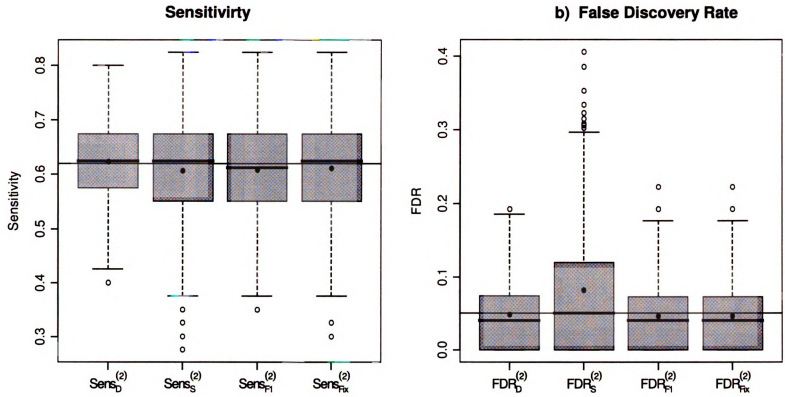


Figure 4.8. Distribution of the sensitivity and false discovery rate obtained from simulations based on different procedures to select the significance cutoff ($\alpha^{(1)}$). The points indicate the mean whereas the line is the expected value obtained from the optimization algorithm.

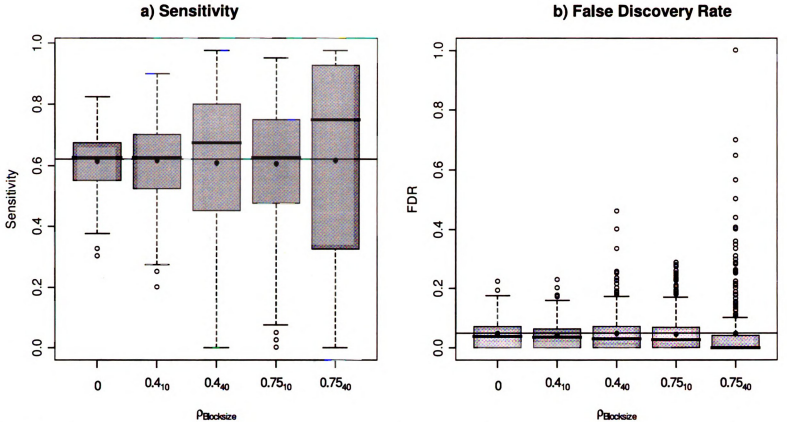


Figure 4.9. Distribution of a) sensitivity and b) false discovery rate obtained from simulations for different correlation structure in the data. The *x*-axis numbers indicate the correlation coefficient whereas the subindex on the x-axis values indicates the block size. The points within each box indicate the mean whereas the line reflects the expected value obtained from the optimization algorithm assuming independence.

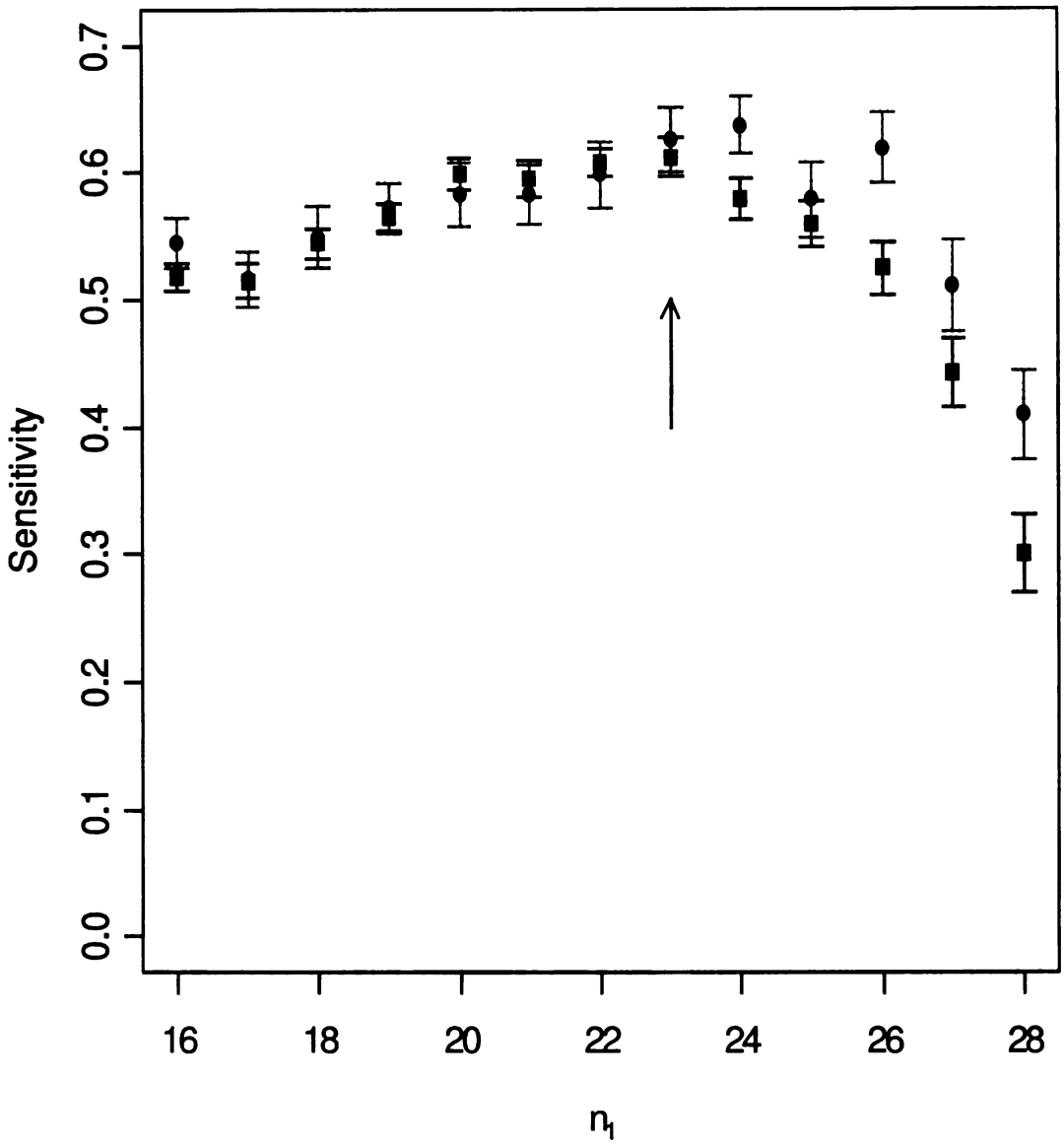


Figure 4.10. Sensitivity of two-step tests for correlated hypotheses as function of the sample size in the first step. N=200 simulation with π_0=0.99, f=0.05, $\delta_j^{(1)}$=1, n=30, $R^{(1)}/m$=0.01. The arrow indicates the optimal sample allocation obtained assuming uncorrelated tests. •: ρ=0.4, block size=40, ■: ρ=0.4, block size=10.

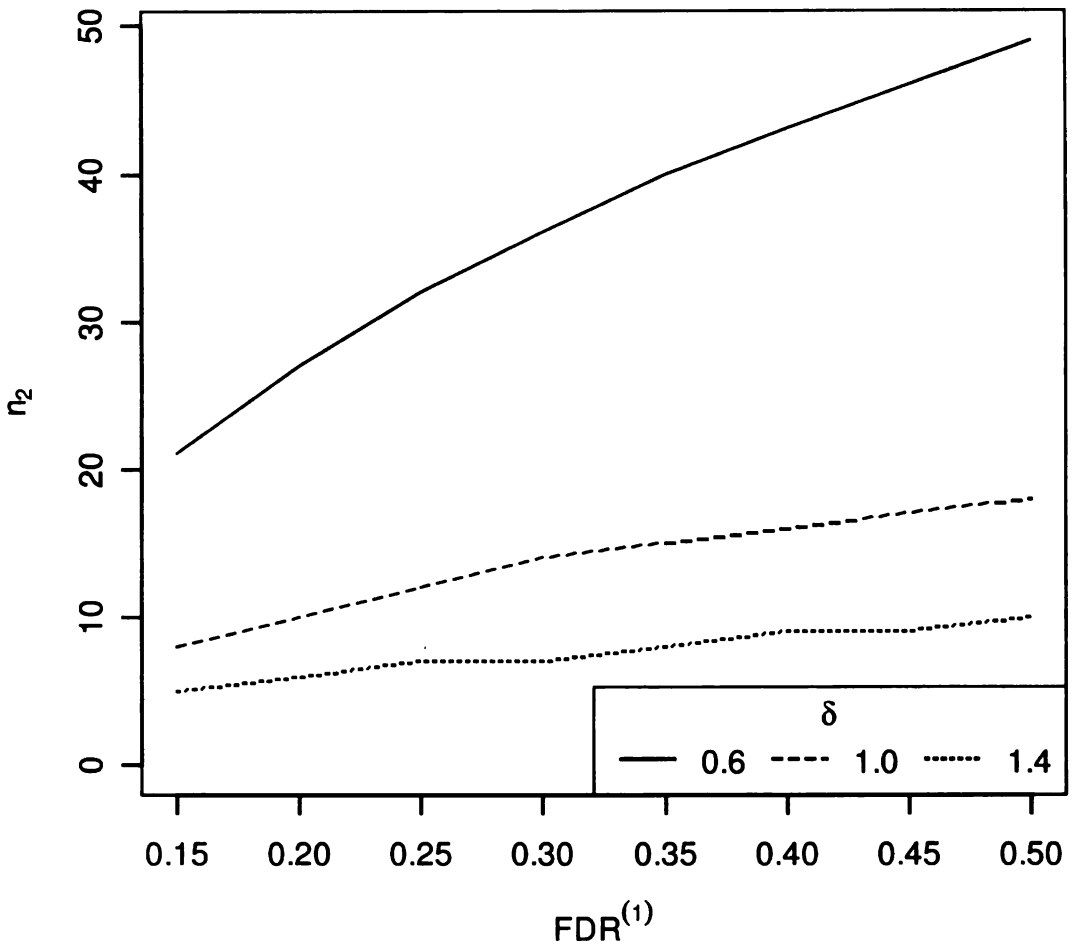


Figure 4.11 Minimum number of replicates per treatment in a validation experiment as function of proportion of null hypotheses ($FDR^{(1)}$) and effect sizes. $R^{(1)}$=25 genes putatively differentially expressed at different levels of FDR , target sensitivity = 0.9, FDR after validation = 0.05.

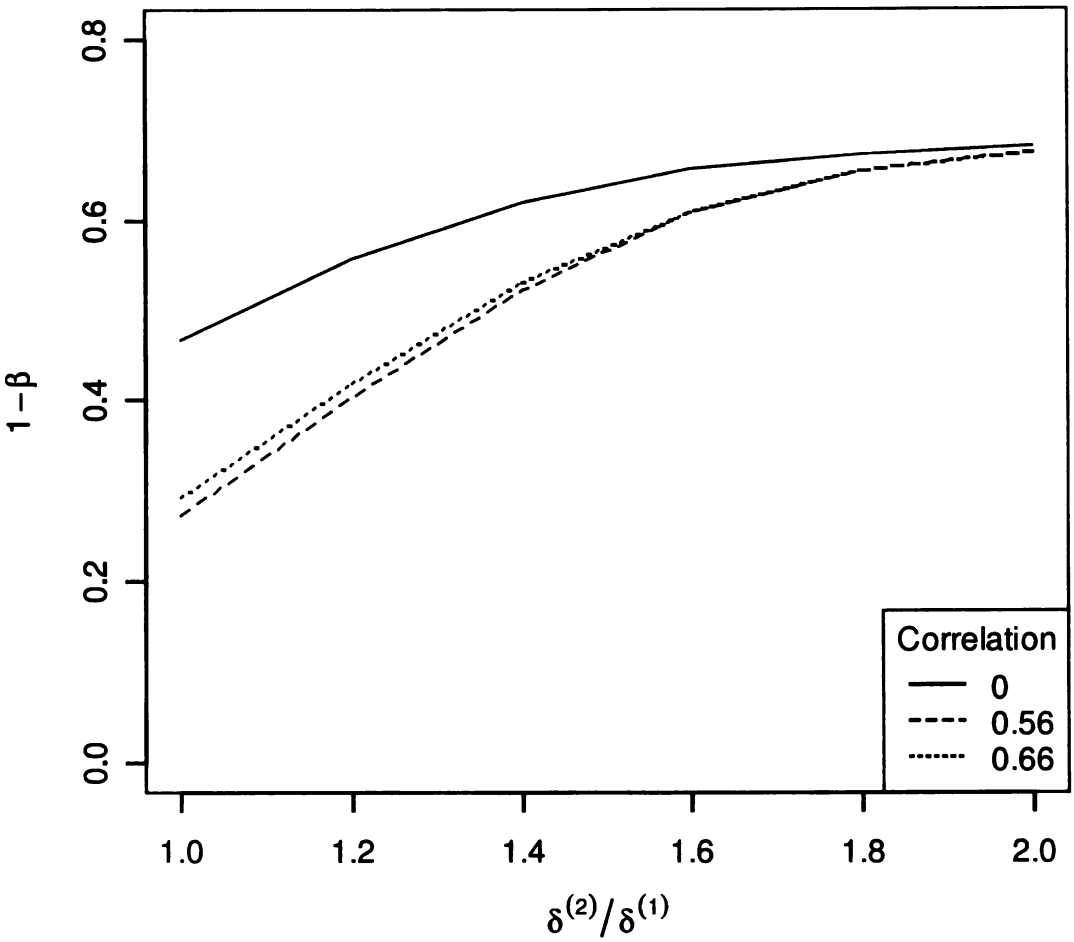


FIGURE 4.12. Power of a single two-test for different correlations (0, 0.56, and 0.66) between the first and second stage data as a function the ratio of second stage to first stage effect sizes. The effect size of the first stage ($\delta^{(1)}$) is 0.6 whereas the number of replicates per treatment is 25 for each test.

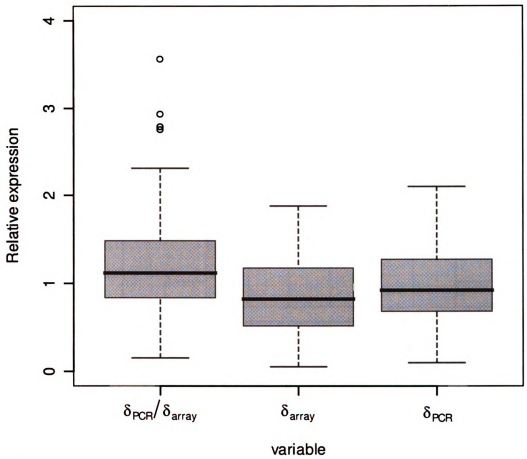


Figure 4.13. Box-Plot of the microarray and qRT-PCR based effect size estimates and their ratios.

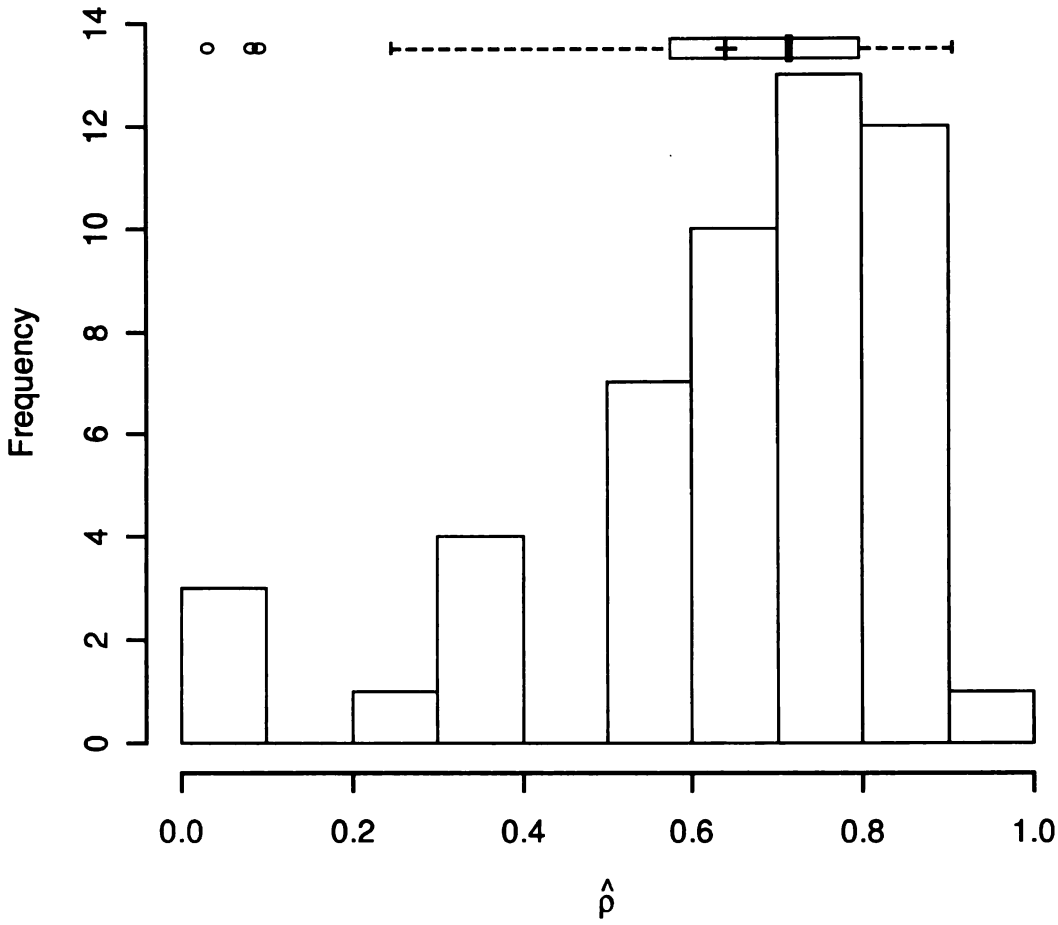


Figure 4.14. Histogram of the estimated correlation coefficient between microarray and qRT-PCR measured expressions.

GENERAL DISCUSSION

Transcription profiling is essential for eludicating gene function and regulation. In particular, microarrays and qRT-PCR are the most commonly used platforms to measure levels of mRNA of target genes. Since the first emergence of microarrays as the high throughput technique of choice for transcription profiling, there has been a consensus that sound experimental designs, appropriate statistical models and methods with validation are crucial to the experimental process (Allison et al., 2006). In this dissertation, I addressed some particular aspects of experimental design, statistical modeling and validation of transcription profiling studies.

1. Objectives revisited and their impact in animal functional genomics.

1) To compare alternative reference designs for statistical efficiency of two color microarray experiments considering multiples sources of variation.

Two color microarrays (i.e. cDNA and oligonucleotide microarrays) are platforms of choice for transcription profiling in animal functional genomics studies. The reference design is one of the most popular designs despite not necessarily being the most statistically efficient. The term reference design is generally used for any different experimental layout where a common reference is hybridized with a different test sample in every array. Reference designs have been previously compared to other designs (Dobbin and Simon, 2002; Kerr, 2003a; Kerr and Churchill, 2001b; Tempelman, 2005;

Yang and Speed, 2002), but an exhaustive comparison among all variants of the reference design has not yet been performed. In Chapter 1, the comparison among designs was made in terms of relative efficiency for the same amount of resources. In contrast to some previous work (Kerr and Churchill, 2001a; Yang and Speed, 2002), we make a clear distinction between technical and biological variability. Additionally, we introduced a new variant of the reference design, that we called Blocked Reference Design (BRD). The BRD had the highest efficiency compared to any of the traditional design alternatives. The BRD is potentially very useful in studies where comparisons with a control group is of interest, yet comparisons between non-control groups are also important. The BRD is as efficient as the replicated reference design for direct comparisons whereas it is as efficient as the common reference design for indirect comparisons (between non-control groups). One criticism made of the common reference design is that each hybridization involves a sample that is of no intrinsic interest to the researcher. This is not the case with the BRD where half of the hybridizations are also used for the control group. Depending on the number of groups, the BDR may be more efficient for some comparisons than other designs such as the loop design. For the comparisons of these designs, we used a model based upon log-ratios of expression as response variables; nevertheless, we took special care to consider all sources of variation to account for hierarchical replication that is present in the BRD.

2) To investigate the ramifications of log-ratio versus log-intensity modeling in two color microarrays using linear mixed effects models

Chapter 2 compared two broad classes of statistical models used for the analysis of microarray data. Kerr (2003b) compared log-intensity and log-ratio models for some simple experimental designs, and we extended their work to more general cases based on using a linear transformation that connect the two models. The focus of our study was in modeling hierarchical sources of variation and the potential loss of information incurred when using log-ratios rather than log-intensities as response variables. For this purpose we used mixed model methods that had already been presented in the context of microarray experiments (Rosa et al., 2005; Wolfinger et al., 2001).

We showed that the analysis of intensities, specifying the array as a random effect, lead to the recovery of inter-block information. In our first chapter, we used a log-ratio model to compare reference designs ignoring the inter-block information. Accounting for such information, however, is more important when comparing reference designs to direct comparison designs where there is no recovery of inter slide information as shown by Kerr (2003a; 2003b). We showed that the recovery of inter-slide information is very important in some other designs such as the split plot design. For that design, the analysis of log-ratios, or the use of a log-intensity model with fixed array effects, leads to the loss of all information for inferring upon the sub-plot factor. Moreover, even if the recovery of interslide information is rather minor as in some cases, we showed that it is often essential to properly account for hierarchical replication that occasionally occurs with two color microarray designs. Formulating the correct model on the log-ratio scale is not always a trivial task; nevertheless, such a model can always be

derived from a log-intensity model by applying a linear transformation. Another important point that was further evident from Chapter 2 is that treating experimental units as fixed effects generally does not lead to proper inferences even though point estimates and treatment differences are the same as those from a log-intensity model. These results are relevant to the applied genomicists, particularly those using statistical analysis software based on log-ratio models, such as LIMMA (Smyth, 2004; Smyth et al., 2005).

3) To develop linear mixed models for the analysis of relative quantification RT-PCR data.

The qRT-PCR technology is currently the method of choice when validating findings from microarray studies. In contrast to microarray data, the development of statistical methodology for the statistical analysis of qRT-PCR data has been overlooked until recently (Yuan et al., 2006). Most of the currently available statistical methods in this area are more useful for simple experimental situations where the main goal consists of pairwise comparisons of experimental groups. Recently, some linear models have been proposed for the analysis of qRT-PCR data (Cook et al., 2004; Fu et al., 2006; Szabo et al., 2004; Yuan et al., 2006). In Chapter 3 we proposed a linear model for the joint analysis of test and control genes that allows one to accommodate an arbitrarily complex design and to test general linear hypotheses in qRT-PCR experiments. In comparison with some of the previous work, we again make a clear distinction between technical and biological replicates. For example, the work by Yuan et. al (2006) or Szabo et. al (2004)

considered only one level of replication. Fu et al. (2006) addressed the issue of technical and biological replicates using a generalized estimation equation model, but they simplified the model by taking the difference of C_T between control and test genes in each technical replicate. Our experience is that this strategy cannot be used in most experimental designs because technical replicates of control and test genes generally cannot be matched. We used a motivating example from an animal genomics study (Poletto et al., 2006a) to demonstrate the implementation of the model and resorted to simulation to evaluate properties of the statistical procedure and estimates derived from it. Finally we applied the analysis model to several datasets (Coussens et al., 2003; Coussens et al., 2004; Szabo et al., 2004) to illustrate its flexibility. Our model is more powerful than the ΔΔCT method (Livak and Schmittgen, 2001) and provides a better control of the Type I error rate than a previously published method (Cook et al., 2004). Using simulation, we demonstrated the importance of including random effects of biological replication within gene in order to correctly draw inferences in the presence of technical replications. The linear mixed model presented in Chapter 3 is not only more powerful than classical analysis methods, but is also more flexible as it allows testing for general linear hypothesis such as interactions and linear trends.

4) To propose a general framework for determining the false discovery rate and sensitivity of gene expression studies when jointly designing microarray screening experiments linked to subsequent and selective qRT-PCR validation.

Optimizing design of validation experiments has not been considered previously in the statistical genomics literature. In Chapter 4, we addressed the issue of sample size calculation and allocation in validation experiments using a two-stage testing approach. Sample size calculations for two stage tests have been developed for association studies (Satagopan et al., 2004; Satagopan et al., 2002; Zehetmayer et al., 2005) but not for transcription profiling experiments. We considered various measures of power and error control concentrating on those cases where several genes are validated in an attempt to discover most differentially expressed genes within an experiment. We demonstrated that more samples are generally needed in the screening stage to ensure high sensitivity for adequate error control as necessary for microarray experiments. Conversely, we showed that a more liberal level of statistical significance could be used in the second stage thereby requiring fewer samples in order to attain a high sensitivity. This result per se is very important because it defies a currently more common practice; i.e. a limited sample size is often used for the microarray study whereas more samples are used in the qRT-PCR stage. Moreover, the traditional levels of significance, 0.05 and 0.01, are typically used in the second stage when a higher cutoff might yield an adequate control of the false discovery rate. A similar finding was reported by Zehetmayer et al. (2005) in two stage designs for association studies. However, we have to acknowledge that the implementation of these guidelines, however, will imply an increase in the cost of the experiments if a high sensitivity is desired. Finally, Chapter 4 addressed the power and false discovery rate of a technical validation study using the same samples in both stages. Even though technical validation has been recognized as having questionable merit

(Allison et al., 2006), it is still used extensively in functional genomics. Our results suggest that if the same samples are used for the validation, the levels of significance should be adjusted downwards accordingly to appropriately control the false discovery rate. Moreover, we provide expressions to numerically calculate these levels based on the first stage significance level and the correlation among measures on gene expression between the two technologies. Accounting for such corrections would help mitigate the problem of double false positives due to correlation among tests.

2. Future research directions

A natural extension of the linear mixed models used in Chapter 2 would be to consider borrowing information across genes. Shrinkage estimators for intensity based mixed models were presented by Cui et al. (2005) while shrinking for log-ratio models was published by Smyth (2004; 2005) but based on the use of fixed effect models or a very restricted class of mixed effects models. The concepts presented by Cui et al. (2005) however, could be readily extended to a mixed model for log-ratios.

The mixed model for qRT-PCR studies assumed a simple correlation structure among genes, as derived from the sampling scheme, but assuming no co-regulation or biological correlation between mRNA level of genes. This assumption, though implausible, is very common in microarrays and qRT-PCR studies; nevertheless, a multivariate model could be used for qRT-PCR if a few genes are assayed for a large number of samples perhaps by generalizing, for example, the model used by Poletto et al (2006b) to incorporate technical replication. It is expected that a multivariate linear

model will use information across genes and could yield important information about the correlation in the expression of co-regulated genes.

Finally, given the very simple experimental setup used in Chapter 4, the sample size allocation and rejection rules should be considered to be preliminary results as applied to potentially more efficient incomplete block designs. In those cases, optimization of two stage tests could be studied using mixed models for qRT-PCR as presented in Chapter 3 for more complex designs. Such an extension, however, is limited by the lack of knowledge regarding the correlation between the two techniques, although this correlation could be estimated using bivariate linear models. Publicly available datasets may provide some information about the relative effect sizes of the two technologies and to some extent of the technical correlation, but at some point, future experiments might be required to fully understand the co-variation for gene expression between microarrays and qRT-PCR technologies.

APPENDIX ONE

SUPPLEMENTARY MATERIAL TO CHAPTER 3

The data used in the first sections of this paper are part of a study that involved the analysis of expression of several genes in the brain of piglets subjected to weaning and isolation stress. The complete presentation of those results is the motivation of an independent publication(Poletto et al., 2006) and a detailed description can be found in that paper. In this section we give more details of the statistical analysis, including variable definition, data structure and program code.

Quantitative Real-Time RT-PCR. Q-RT-PCR was performed as described previously(Gibson et al., 1996; Heid et al., 1996). Briefly, a total of 2 μg of total RNA from each sample was reverse transcribed using oligo $(dT)_{18}$ and SuperScript II Rnase H-reverse transcriptase (Invitrogen Life Technologies Corp., Carlsbad, CA). Reverse transcribed cDNA was quantified using ND-1000 spectrophotometer (NanoDrop Technologies Inc., Rockland, DE). A total of 30 ng of cDNA was employed in each real-time reaction. Forward and reverse primer sequences were designed with Primer Express 2.0 Software (Applied Biosystems, Foster City, CA) and synthesized by Quiagen. The oligonucleotide sequences of the primers are summarized in Table 1. To confirm that primer-dimer products were not influencing final Ct values, control reactions without template but with each set of primers were performed, with the anticipated result that no product was amplified. Q-RT-PCR was performed and analyzed on an ABI Prism 7000 Sequence Detection System (Applied Biosystems). *Sus scrofa* 18S ribosomal RNA was selected as the control gene to be used for normalization purposes based on preliminary test reactions, as expression of this proposed control gene expression did not change

relative to treatments. All reactions were performed using template from individual animals in triplicate. Relative quantification methods were determined based on the primer amplification efficiency tests(Livak and Schmittgen, 2001).

Table 1. Primer sequences and efficiencies.

Gene		Sequence	E^a
18S	Forward	5'-GGCTCATTAAATCAGTTATGGTTCCT-3'	1.99
	Reverse	5'-AGCTCTAGAATTACCACAGTTATCCAAG-3'	
DBI	Forward	5'-GGAAGTTAAGAACCTTAAGACCAAACC-3'	1.96
	Reverse	5'-TCGCTTGTTTGTAGTGGCTGTAG-3'	

[a] amplification efficiency obtained from a relative standard curve (R^2>0.99). The two values are close enough to the ideal efficiency (E=2) such that the raw C_T is a valid proxy for the log-concentration of mRNA.

Response variable for linear model analysis if E≠2. In the presence of estimates of the amplification efficiency (E < 2.0) and C_T value from each amplification curve, the response variable for the analysis will be:

$$y_{gijkr} = Log_2(E_{gijkr}{}^{-CT_{gijkr}}), \quad (1)$$

where the efficiency values may be estimated from the amplification curve (i.e analytical method(Marino et al., 2003)) or from a relative standard curve included in the assay plate(Pfaffl, 2001). If a relative standard curve quantification is selected, the ABI 7000 sequence detection system produces a "quantity" value in the output. Such a value is directly proportional to the efficiency corrected C_T in (1), and the user only has to transform the value by taking logarithm. The back-transformation equivalent to equation (9) of the **Methods** section will be:

$$FC_{(EWI-EWC)} = 2^{diff(EWI-EWC)}. \quad (2)$$

SAS code and data. The SAS program used to fit the linear mixed model and obtain the log-fold change estimates is presented in Figure 1.

Figure 1. SAS Code for implementation of mixed model analysis with model (2).

```
proc mixed data=ratio;
class gene trt sample litter;
model ct=gene|trt /outp=resi;
random sample;
random litter litter*trt/group=gene;

estimate 'ewc-nwc' gene*trt  -1 0 1 0 1 0 -1 0/cl;
estimate 'ewi-nwi' gene*trt  0 -1 0 1 0 1 0 -1/cl;
estimate 'ewi-ewc' gene*trt  1 -1 0 0 -1 1 0 0/cl;
estimate 'nwi-nwc' gene*trt  0 0 1 -1 0 0 -1 1/cl;

estimate 'interaction' gene*trt -1 1 1 -1 1 -1 -1 1;

estimate 'main effect: ew-nw' gene*trt -.5 -.5 .5 .5 .5 .5 -.5 -.5/cl ;
estimate 'main effect: i-c' gene*trt   .5 -.5 .5 -.5 -.5 .5 -.5 .5/cl;
run;
```

The estimate commands provide hypothesis tests, point and interval estimates of the log-fold change.

The input data was prepared from the relative quantification output of the ABI 7000 Sequence detection system. Treatment and litter memberships were added a posteriori. Table 2 presents the complete dataset.

Table 2 | Datafile used by the SAS program to fit the mixed model.

well	Sample	gene	CT	TRT	litter
A4	Fc1	18S	17.02	nwc	3
B4	Fc1	18S	16.46	nwc	3
C4	Fc1	18S	18	nwc	3
A11	Fc11	18S	18.18	nwi	1
B11	Fc11	18S	18.06	nwi	1
C11	Fc11	18S	17.64	nwi	1
A6	Fc12	18S	16.07	ewi	2
B6	Fc12	18S	16.12	ewi	2
C6	Fc12	18S	16.13	ewi	2
A10	Fc15	18S	16.23	ewi	1
B10	Fc15	18S	16.35	ewi	1
C10	Fc15	18S	16.56	ewi	1
B7	Fc16	18S	19.45	nwi	2
C7	Fc16	18S	19.11	nwi	2
B5	Fc17	18S	16.64	ewc	2
C5	Fc17	18S	16.52	ewc	2
A9	Fc18	18S	17.04	ewc	1
B9	Fc18	18S	16.83	ewc	1
C9	Fc18	18S	16.58	ewc	1
A12	Fc19	18S	17.47	nwc	1
B12	Fc19	18S	17.33	nwc	1
C12	Fc19	18S	17.53	nwc	1
A1	Fc3	18S	18.26	ewc	3
B1	Fc3	18S	18.06	ewc	3
C1	Fc3	18S	18.42	ewc	3
A2	Fc4	18S	15.8	ewi	3
B2	Fc4	18S	15.79	ewi	3
C2	Fc4	18S	15.92	ewi	3
A3	Fc6	18S	15.17	nwi	3
B3	Fc6	18S	15.34	nwi	3
C3	Fc6	18S	15.35	nwi	3
A8	Fc9	18S	16.19	nwc	2
B8	Fc9	18S	16.12	nwc	2
C8	Fc9	18S	15.96	nwc	2

well	Sample	gene	CT	TRT	litter
D4	Fc1	DBI	23.33	nwc	3
E4	Fc1	DBI	23.43	nwc	3
F4	Fc1	DBI	23.3	nwc	3
D11	Fc11	DBI	25.31	nwi	1
E11	Fc11	DBI	25.5	nwi	1
F11	Fc11	DBI	25.56	nwi	1
D6	Fc12	DBI	23.86	ewi	2
E6	Fc12	DBI	23.75	ewi	2
F6	Fc12	DBI	24.14	ewi	2
D10	Fc15	DBI	23.85	ewi	1
E10	Fc15	DBI	23.9	ewi	1
F10	Fc15	DBI	23.83	ewi	1
D7	Fc16	DBI	27.19	nwi	2
E7	Fc16	DBI	28.84	nwi	2
F7	Fc16	DBI	28	nwi	2
D5	Fc17	DBI	23.08	ewc	2
E5	Fc17	DBI	23.16	ewc	2
D9	Fc18	DBI	22.31	ewc	1
E9	Fc18	DBI	22.78	ewc	1
F9	Fc18	DBI	23.18	ewc	1
D12	Fc19	DBI	23.93	nwc	1
E12	Fc19	DBI	23.78	nwc	1
F12	Fc19	DBI	24.35	nwc	1
D1	Fc3	DBI	24.83	ewc	3
E1	Fc3	DBI	24.71	ewc	3
F1	Fc3	DBI	24.76	ewc	3
D2	Fc4	DBI	24.57	ewi	3
E2	Fc4	DBI	24.7	ewi	3
D3	Fc6	DBI	24.15	nwi	3
E3	Fc6	DBI	24.22	nwi	3
F3	Fc6	DBI	24.22	nwi	3
D8	Fc9	DBI	23.77	nwc	2
F8	Fc9	DBI	24.01	nwc	2

The dataset is presented in two sets of columns, one for the control gene and one for the target gene, but it was input into SAS as a single dataset with six columns.

Model comparison. Models used in the datasets listed in **Table 4** of the paper.

1. Paratuberculosis infection dataset, Full model:

$$y_{gijkl} = GSI_{gij} + B(S)_{ik} + BI(S)_{ijk} + GB(S)_{gik} + GBI(S)_{gijk} + e_{gijkl},$$

Where: y_{gijkl} is the C_T observed in the l^{th} well corresponding to the g^{th} gene in the k^{th} sample with status i and treatment j. G: Gene, B: Biological sample, S: Status, I: treatment. GSI_{gij} is the mean expression for gene g in the Status level i and treatment $j, B(S)_{ik} \sim N(0,\sigma^2_B)$, $BI(S)_{ijk} \sim N(0,\sigma^2_{BI})$, $GB(S)_{gik} \sim N(0,\sigma^2_{GB_g})$, $GBI(S)_{gijk} \sim N(0,\sigma^2_{GBI_g})$, $e_{gijkl} \sim N(0,\sigma^2_{e_g})$. Reduced models were generated by omitting random effects or assuming homogeneous variances.

2. Paratuberculosis time course, Full model:

$$y_{gikl} = GT_{gi} + B_k + BT_{ik} + GB_{gk} + GBT_{gik} + e_{gikl}$$

Where: $\underline{y_{gjkl}}$ is the C_T observed in the l^{th} well corresponding to the g^{th} gene in the k^{th} sample at time i. G: Gene, B: Biological sample, T: Time. GT_{gi} is the mean expression for gene g in the time i, $B_k \sim N(0,\sigma^2_B)$, $BT_{ik} \sim N(0,\sigma^2_{BT})$, $GB_{gk} \sim N(0,\sigma^2_{GB_g})$, $GBT_{gik} \sim N(0,\sigma^2_{GBI_g})$, $e_{gikl} \sim N(0,\sigma^2_{e_g})$. Reduced models were generated by omitting random effects or assuming homogeneous variances.

3. Stress induced differential expression in piglets. Full model:

See model (5) in Material and methods of the paper.

4. rd/rd mice and Taqman low desity array. Full model:

$$y_{gikl} = GT_{gi} + B(T)_{ik} + GB(T)_{gik} + e_{gikl}$$

Where: y_{gjkl} is the C_T observed in the l^{th} well corresponding to the g^{th} gene in the k^{th} sample from group *i*. *G*: Gene, GT_{gi} is the mean expression for gene *g* in the group *i*. *B*: Biological sample, T: group (rd or wild type in rd/rd dataset and mutated or non-mutated in taqman low density array dataset). $B(T)_{ik} \sim N(0, \sigma^2_{BT})$, $GB(T)_{gik} \sim N(0, \sigma^2_{GBI_g})$, $e_{gikl} \sim N(0, \sigma^2_{e_g})$. Reduced models were generated by omitting random effects or assuming homogeneous variances.

5. Single tissue housekeeping comparison:

$$y_{gk} = G_g + B_k + e_{gk}$$

Where: y_{gjkl} is the C_T observed in the well corresponding to the g^{th} gene in the k^{th} sample. *G*: Gene, *B*: Biological sample. $B_k \sim N(0, \sigma^2_B)$, $e_{gk} \sim N(0, \sigma^2_{e_g})$. Reduced models were generated by omitting the random effect or assuming homogeneous error variance.

Table 3. Effects included in the best-fit models.

Dataset	Fixed effects[a]	Random effects		
		Sample[b]	Sample-gene[c]	Residual[d]
SPLIT	gene * Status* infection	yes	Yes. Gene specific variance	Gene specific variance
TC	Gene*Time	yes	Yes. Gene specific variance	Gene specific variance
PFC	Gene * Weaning * Isolation	yes	Yes. Gene specific variance	Gene specific variance
MRD	Gene*strain	yes	No/Yes[e]	Gene specific variance
TLD	Gene*mutation	yes	Yes. Gene specific variance	Homogeneous variance
SHK	Gene	yes	Gene Specific variance	

[a]specification for fixed effects. [b]random sample effect included (yes) or not (no). [c]random sample by gene interaction included (yes/no) and variance of the effect (Homogeneous or Gene specific). [d]Specification of the residual variance (Homogeneous or Gene specific). [e]the model without sample-gene random effect and the model with a sample-gene interaction (with homogeneous variances) yielded the same values for both selection criteria.

Table 4. Contrasts of interest.

Dataset	Comparisons with ΔCT and LMM[a]		Comparisons with LMM only[b]
	Description	Number	
SPLIT	Simple effect of Infection	3	Interaction infection by status
	Simple effect of status	4	
TC	Every time versus baseline	4	Linear and quadratic trend
PFC	Simple effect of weaning	2	Interaction weaning by isolation
	Simple effect of isolation	2	
Mrd	Wild type versus mutant	1	-
TLD	Mutated versus unmutated	1	-

[a]Contrasts that can be obtained with both methodologies. [b]Comparison with LMM: contrasts that are calculated with the linear model only.

SUPLEMENT REFERENCES

Gibson, U. E., C. A. Heid, and P. M. Williams. 1996. A novel method for real time quantitative rt-pcr. Genome Res 6: 995-1001.

Heid, C. A., J. Stevens, K. J. Livak, and P. M. Williams. 1996. Real time quantitative pcr. Genome Res 6: 986-994.

Livak, K. J., and T. D. Schmittgen. 2001. Analysis of relative gene expression data using real-time quantitative pcr and the 2(-delta delta c(t)) method. Methods 25: 402-408.

Marino, J. H., P. Cook, and K. S. Miller. 2003. Accurate and statistically verified quantification of relative mrna abundances using sybr green i and real-time rt-pcr. J Immunol Methods 283: 291-306.

Pfaffl, M. W. 2001. A new mathematical model for relative quantification in real-time rt-pcr. Nucleic Acids Res 29: e45.

Poletto, R., J. M. Siegford, J. P. Steibel, P. M. Coussens, and A. J. Zanella. 2006. Investigation of changes in global gene expression in the frontal cortex of early-weaned and socially isolated piglets using microarray and quantitative real-time rt-pcr. Brain Res 1068: 7-15.

REFERENCES

REFERENCES

Abruzzo, L. V. et al. 2005. Validation of oligonucleotide microarray data using microfluidic low-density arrays: A new statistical method to normalize real-time rt-pcr data. Biotechniques 38: 785-792.

Alizadeh, A. A. et al. 2000. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403: 503-511.

Allison, D. B., X. Cui, G. P. Page, and M. Sabripour. 2006. Microarray data analysis: From disarray to consolidation and consensus. Nat Rev Genet 7: 55-65.

Andersen, C. L., J. L. Jensen, and T. F. Orntoft. 2004. Normalization of real-time quantitative reverse transcription-pcr data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res 64: 5245-5250.

Belbin, T. J. et al. 2004. Indirect measurements of differential gene expression with cdna microarrays. Biotechniques 36: 310-314.

Ben Ahmed, M., H. Houman, M. Miled, K. Dellagi, and H. Louzir. 2004. Involvement of chemokines and th1 cytokines in the pathogenesis of mucocutaneous lesions of behcet's disease. Arthritis Rheum 50: 2291-2295.

Brunner, A. M., I. A. Yakovlev, and S. H. Strauss. 2004. Validating internal controls for quantitative plant gene expression studies. BMC Plant Biol 4: 14.

Bueno Filho, J. S. S., S. G. Guilmour, and G. J. M. Rosa. 2005. Design of microarray experiments for genetical genomics studies. Submitted.

Bustin, S. A. 2000. Absolute quantification of mrna using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol 25: 169-193.

Churchill, G. A. 2002. Fundamentals of experimental design for cdna microarrays. Nat Genet 32 Suppl: 490-495.

Cook, P., C. Fu, M. Hickey, E. S. Han, and K. S. Miller. 2004. Sas programs for real-time rt-pcr having multiple independent samples. Biotechniques 37: 990-995.

Coussens, P. M., C. J. Colvin, G. J. Rosa, J. Perez Laspiur, and M. D. Elftman. 2003. Evidence for a novel gene expression program in peripheral blood mononuclear cells from mycobacterium avium subsp. Paratuberculosis-infected cattle. Infect Immun 71: 6487-6498.

Coussens, P. M., A. Jeffers, and C. Colvin. 2004. Rapid and transient activation of gene expression in peripheral blood mononuclear cells from johne's disease positive cows exposed to mycobacterium paratuberculosis in vitro. Microb Pathog 36: 93-108.

Cui, X. G., J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 6: 59-75.

Dobbin, K., and R. Simon. 2002. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18: 1438-1445.

Fu, W. J. J., J. B. Hu, T. Spencer, R. Carroll, and G. Y. Wu. 2006. Statistical models in assessing fold change of gene expression in real-time rt-pcr experiments. Computational Biology and Chemistry 30: 21-26.

Gentle, A., F. Anastasopoulos, and N. A. McBrien. 2001. High-resolution semi-quantitative real-time pcr without the use of a standard curve. Biotechniques 31: 502, 504-506, 508.

Gibson, U. E., C. A. Heid, and P. M. Williams. 1996. A novel method for real time quantitative rt-pcr. Genome Res 6: 995-1001.

Giglio, S., P. T. Monis, and C. P. Saint. 2003. Demonstration of preferential binding of sybr green i to specific DNA fragments in real-time multiplex pcr. Nucleic Acids Res 31: e136.

Giulietti, A. et al. 2001. An overview of real-time quantitative pcr: Applications to quantify cytokine gene expression. Methods 25: 386-401.

Glonek, G. F. V., and P. J. Solomon. 2004. Factorial and time course designs for cdna microarray experiments. Biostatistics 5: 89-111.

Gorreta, F., D. Barzaghi, A. J. VanMeter, V. Chandhoke, and L. Del Giacco. 2004. Development of a new reference standard for microarray experiments. Biotechniques 36: 1002-1009.

Hamalainen, H. et al. 2001. Distinct gene expression profiles of human type 1 and type 2 t helper cells. Genome Biol 2: RESEARCH0022.

Heid, C. A., J. Stevens, K. J. Livak, and P. M. Williams. 1996. Real time quantitative pcr. Genome Res 6: 986-994.

Hu, J., F. Zou, and F. A. Wright. 2005. Practical fdr-based sample size calculations in microarray experiments. Bioinformatics 21: 3264-3272.

Jansen, R. C., and J. P. Nap. 2001. Genetical genomics: The added value from segregation. Trends Genet 17: 388-391.

Johnson, M. R., K. Wang, J. B. Smith, M. J. Heslin, and R. B. Diasio. 2000. Quantitation of dihydropyrimidine dehydrogenase expression by real-time reverse transcription polymerase chain reaction. Anal Biochem 278: 175-184.

Jung, S. H. 2005. Sample size for fdr-control in microarray data analysis. Bioinformatics 21: 3097-3104.

Kerr, M. K. 2003a. Design considerations for efficient and effective microarray studies. Biometrics 59: 822-828.

Kerr, M. K. 2003b. Linear models for microarray data analysis: Hidden similarities and differences. J Comput Biol 10: 891-901.

Kerr, M. K., and G. A. Churchill. 2001a. Experimental design for gene expression microarrays. Biostatistics 2: 183-201.

Kerr, M. K., and G. A. Churchill. 2001b. Statistical design and the analysis of gene expression microarray data. Genetical Research 77: 123-128.

Kerr, M. K., M. Martin, and G. A. Churchill. 2000. Analysis of variance for gene expression microarray data. J Comput Biol 7: 819-837.

Kim, H. et al. 2002. Use of rna and genomic DNA references for inferred comparisons in DNA microarray analyses. Biotechniques 33: 924-930.

Kishimoto, Y. et al. 2004. Gene expression relevant to osteoclastogenesis in the synovium and bone marrow of mature rats with collagen-induced arthritis. Rheumatology (Oxford) 43: 1496-1503.

Konig, R., D. Baldessari, N. Pollet, C. Niehrs, and R. Eils. 2004. Reliability of gene expression ratios for cdna microarrays in multiconditional experiments with a reference design. Nucleic Acids Research 32: e29.

Lin, S. J. et al. 2002. Calorie restriction extends saccharomyces cerevisiae lifespan by increasing respiration. Nature 418: 344-348.

Littell, R. C. 1996. Sas system for mixed models. SAS Institute Inc., Cary, N.C.

Liu, W., and D. A. Saint. 2002. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. Anal Biochem 302: 52-59.

Livak, K. J., and T. D. Schmittgen. 2001. Analysis of relative gene expression data using real-time quantitative pcr and the 2(-delta delta c(t)) method. Methods 25: 402-408.

Lonnstedt, I., and T. Speed. 2002. Replicated microarray data. Statistica Sinica 12: 31-46.

Marino, J. H., P. Cook, and K. S. Miller. 2003. Accurate and statistically verified quantification of relative mrna abundances using sybr green i and real-time rt-pcr. J Immunol Methods 283: 291-306.

Martinez, M. J. et al. 2004. Genomic analysis of stationary-phase and exit in saccharomyces cerevisiae: Gene expression and identification of novel essential genes. Mol Biol Cell 15: 5295-5305.

Miller, R. A., A. Galecki, and R. J. Shmookler-Reis. 2001. Interpretation, design, and analysis of gene array expression experiments. J Gerontol A Biol Sci Med Sci 56: B52-57.

Muller, P. Y., H. Janovjak, A. R. Miserez, and Z. Dobbie. 2002. Processing of gene expression data generated by quantitative real-time rt-pcr. Biotechniques 32: 1372-+.

Papp, B., C. Pal, and L. D. Hurst. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424: 194-197.

Pawitan, Y., S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner. 2005. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 21: 3017-3024.

Peirson, S. N., J. N. Butler, and R. G. Foster. 2003. Experimental validation of novel and conventional approaches to quantitative real-time pcr data analysis. Nucleic Acids Res 31: e73.

Pellagatti, A. et al. 2003. Gene expression profiling in polycythemia vera using cdna microarray technology. Cancer Res 63: 3940-3944.

Perou, C. M. et al. 2000. Molecular portraits of human breast tumours. Nature 406: 747-752.

Perreard, L. et al. 2006. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative rt-pcr assay. Breast Cancer Res 8: R23.

Pfaffl, M. W. 2001. A new mathematical model for relative quantification in real-time rt-pcr. Nucleic Acids Res 29: e45.

Pfaffl, M. W., G. W. Horgan, and L. Dempfle. 2002. Relative expression software tool (rest) for group-wise comparison and statistical analysis of relative expression results in real-time pcr. Nucleic Acids Res 30: e36.

Poletto, R., J. M. Siegford, J. P. Steibel, P. M. Coussens, and A. J. Zanella. 2006a. Investigation of changes in global gene expression in the frontal cortex of early-weaned and socially isolated piglets using microarray and quantitative real-time rt-pcr. Brain Res 1068: 7-15.

Poletto, R., J. P. Steibel, J. M. Siegford, and A. J. Zanella. 2006b. Effects of early weaning and social isolation on the expression of glucocorticoid and mineralocorticoid receptor and

11beta-hydroxysteroid dehydrogenase 1 and 2 mrnas in the frontal cortex and hippocampus of piglets. Brain Res 1067: 36-42.

Ramakers, C., J. M. Ruijter, R. H. Deprez, and A. F. Moorman. 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (pcr) data. Neurosci Lett 339: 62-66.

Recinos, A., 3rd et al. 2004. Liver gene expression associated with diet and lesion development in atherosclerosis-prone mice: Induction of components of alternative complement pathway. Physiol Genomics 19: 131-142.

Rhoden, K. J. et al. 2004. Real-time quantitative rt-pcr identifies distinct c-ret, ret/ptc1 and ret/ptc3 expression patterns in papillary thyroid carcinoma. Lab Invest 84: 1557-1570.

Rockett, J. C. 2003. To confirm or not to confirm (microarray data)--that is the question. Drug Discov Today 8: 343.

Rockett, J. C., and G. M. Hellmann. 2004. Confirming microarray data--is it really necessary? Genomics 83: 541-549.

Rosa, G. J. M., J. P. Steibel, and R. J. Tempelman. 2005. Reassessing design and analysis of two-colour microarray experiments using mixed effects models. Comparative and Functional Genomics 6: 123-131.

Satagopan, J. M., and R. C. Elston. 2003. Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25: 149-157.

Satagopan, J. M., E. S. Venkatraman, and C. B. Begg. 2004. Two-stage designs for gene-disease association studies with sample size constraints. Biometrics 60: 589-597.

Schwarz, G. 1978. Estimating dimension of a model. Annals of Statistics 6: 461-464.

Smyth, G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.

Smyth, G. K., J. Michaud, and H. S. Scott. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 21: 2067-2075.

Steibel, J. P., and G. J. M. Rosa. 2005. On reference designs for microarray experiments. Statistical Applications in Genetics and Molecular Biology 4: A36.

Sterrenburg, E., R. Turk, J. M. Boer, G. B. van Ommen, and J. T. den Dunnen. 2002. A common reference for cdna microarray hybridizations. Nucleic Acid Research 30: e116.

Stroup, W. W. 2002. Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. Journal of Agricultural Biological and Environmental Statistics 7: 491-511.

Swillens, S., J. C. Goffard, Y. Marechal, A. de Kerchove d'Exaerde, and H. El Housni. 2004. Instant evaluation of the absolute initial number of cdna copies from a single real-time pcr curve. Nucleic Acids Res 32: e56.

Szabo, A. et al. 2004. Statistical modeling for selecting housekeeper genes. Genome Biol 5: R59.

Tempelman, R. J. 2005. Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet Immunol Immunopathol 105: 175-186.

Tichopad, A., A. Didier, and M. W. Pfaffl. 2004. Inhibition of real-time rt-pcr quantification due to tissue-specific contaminants. Mol Cell Probes 18: 45-50.

Tichopad, A., M. Dilger, G. Schwarz, and M. W. Pfaffl. 2003. Standardized determination of real-time pcr efficiency from a single reaction set-up. Nucleic Acids Res 31: e122.

Tichopad, A., A. Dzidic, and M. W. Pfaffl. 2002. Improving quantitative real-time rt-pcr reproducibility by boosting primer-linked amplification efficiency. Biotechnology Letters 24: 2053-2056.

Timm, N. H. 2004. Estimating effect sizes in exploratory experimental studies when using a linear model. American Statistician 58: 213-217.

Verhoeven, K. J. F., K. L. Simonsen, and L. M. McIntyre. 2005. Implementing false discovery rate control: Increasing your power. Oikos 108: 643-647.

Vinciotti, V. et al. 2005. An experimental evaluation of a loop versus a reference design for two-channel microarrays. Bioinformatics 21: 492-501.

Winer, J., C. K. Jung, I. Shackel, and P. M. Williams. 1999. Development and validation of real-time quantitative reverse transcriptase-polymerase chain reaction for monitoring gene expression in cardiac myocytes in vitro. Anal Biochem 270: 41-49.

Wolfinger, R. D. et al. 2001. Assessing gene significance from cdna microarray expression data via mixed models. J Comput Biol 8: 625-637.

Yang, Y. H., and T. Speed. 2002. Design issues for cdna microarray experiments. Nat Rev Genet 3: 579-588.

Yang, Y. H., and T. P. Speed. 2003. Design and analysis of comparative microarray experiments. In: T. P. Speed (ed.) Statistical analysis of gene expression microarray data. CRC Press, Boca Raton, FL.

Zehetmayer, S., P. Bauer, and M. Posch. 2005. Two-stage designs for experiments with a large number of hypotheses. Bioinformatics 21: 3771-3777.