Analyzing Factorial Designed Microarray Experiments

Denise Scholtens
Harvard School of Pubic Health
Department of Biostatistics
Boston, MA 02115

Alexander Miron
Dana-Farber Cancer Institute
Department of Cancer Biology
Boston, MA 02115

Faisal M. Merchant
Duke University Medical Center
School of Medicine
Durham, NC 27708

Arden Miller
University of Auckland
Department of Statistics
Auckland, New Zealand

Penelope L. Miron
Dana-Farber Cancer Institute
Department of Cancer Biology
Boston, MA 02115

J. Dirk Iglehart
Dana-Farber Cancer Institute
Department of Cancer Biology
Boston, MA 02115

Robert Gentleman
Harvard School of Public Health
Department of Biostatistics
Boston, MA 02115

AUTHOR'S FOOTNOTE

Denise Scholtens is Doctoral Student, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 (dscholte@hsph.harvard.edu). Alexander Miron is Instructor of Surgery, Harvard Medical School, Department of Surgery, Brigham and Women's Hospital, Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115 (Alexander_Miron@dfci.harvard.edu). Faisal M. Merchant is Medical Student, School of Medicine, Duke University Medical Center, Durham, NC 27708 (merch004@mc.duke.edu). Arden Miller is Lecturer, Department of Statistics, University of Auckland, Auckland, New Zealand (miller@stat.auckland.ac.nz). J. Dirk Iglehart is Professor of Surgery, Harvard Medical School, Department of Surgery, Brigham and Women's Hospital, Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115 (jiglehart@partners.org). Penelope L. Miron is Instructor of Surgery, Harvard Medical School, Department of Surgery, Brigham and Women's Hospital, Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115 (Penelope_Miron@dfci.harvard.edu). Robert Gentleman is Associate Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 (rgentlem@hsph.harvard.edu). This work was supported in part by National Cancer Institute Biometry/Epidemiology Training Grant in Biostatistics 5-T32-CA09337 and by National Cancer Institute funded Dana-Farber/Harvard Specialized Program of Research Excellence (SPORE-CA89393) in breast cancer at Dana-Farber Cancer Institute. The authors wish to thank Judi Manola and Patti Bernardo for their contributions to preliminary analyses of this data.

ABSTRACT

High-throughput quantification of gene expression using microarray technology has dramatically changed biological investigation into the roles of genes in normal cell functioning, as well as the mechanisms of disease. We discuss a methodology for framing biological questions in terms of statistical parameters to efficiently and confidently answer questions of interest using microarray data from factorial experiments. Investigators can extract pertinent information from the data about the effects of the factors, their interactions with each other, and the statistical significance of these effects, rather than rely solely on clustering techniques or fold change point estimates in hopes of finding co-expressed genes. By first examining how biological mechanisms are reflected in mRNA transcript abundance, investigators can better plan microarray experiments to answer the most interesting questions.

Keywords: estrogen target, gene expression, gene selection, linear model

1. INTRODUCTION

High-throughput quantification of gene expression using microarray technology has dramatically changed scientific investigation into the roles of genes in normal cell functioning, as well as the mechanisms of disease. Rather than study one gene at a time, investigators can simultaneously measure expression for tens of thousands of genes, and then coordinate these observations for a more complete characterization of cellular activity. The challenge to analyze and interpret large-dimensional microarray output in a variety of experimental settings calls for unprecedented integration of biological understanding and statistical methodology. Global expression profiling has met with much success for sample type classification and gene function hypothesis generation; however, the scientific community lacks biologically interpretable models for the effects of treatments on specific genes within the dynamic cellular network for complex, multifactor experimental designs.

One of the most important applications of microarray technology is gene expression profiling in cancer patients. In some settings, known cancer subtypes direct supervised machine learning analyses for the characterization of these subtypes at the genetic level. For other collections of samples, such as those from symptomatically similar patients who experience vastly different, but unexplained clinical outcomes, unsupervised learning techniques work toward the discovery of new subtypes of disease. Whether supervised or unsupervised, machine learning reasonably accommodates microarray data from patient samples by identifying broad gene expression similarities within groups of samples, but allowing for natural genetic variability between patients (Alizadeh et al., 2000; Alon et al., 1999; Bhattacharjee et al., 2001; Golub et al., 1999; Perou et al., 1999).

Another noted use of microarrays is the grouping of functionally related genes in time-course experiments. In these studies, investigators longitudinally track gene expression levels in cells from organisms such as yeast after exposure to some type of extracellular stress or during a basic process like cell division. Several studies confirm that genes with similar temporal expression patterns often perform similar functions. Well-characterized genes can then inform researchers about the function of their less understood co-clustered counterparts. As in the tumor subtype problem, machine learning techniques such as hierarchical clustering and self-organizing maps can help resolve time-course data, with the direction of the analysis in some sense reversed. Patient sample studies seek similar groups of patients; time-course experiments generally seek groups of co-expressed genes (DeRisi et al., 1996; Eisen et al., 1998; Gasch et al., 2000; Spellman et al., 1998).

The landmark papers in computational biology focus on gene expression profile similarity for the identification of related patients or genes. Somewhat surprisingly, little attention has been devoted to a third type of microarray experiment: the application of treatments in combination to model organisms such as cell lines. Cell lines are genetically identical, immortalized cells which can be grown in culture and manipulated in a tightly controlled manner. Even though their highly altered state precludes immediate applicability of experimental results in vivo, cell lines do allow for in depth examination of the mechanisms involved in cellular response and regulation. Frequently, these studies are carefully designed factorial experiments meant to pinpoint the perturbation of genetic networks by various combinations of factors. Practicality and interpretability often drive investigators to select genes of interest according to multiple pairwise fold change values without fully exploiting the use of replicates or modeling to assess statistical significance. Biologically interpretable and statistically reasonable models for gene expression changes are necessary to explore the behavior of genes in response to treatment and to guide the design of experiments so that the questions of interest are indeed answerable.

In this paper, we demonstrate how classic statistical linear modeling with thoughtful experimental design and biological interpretation of the parameters offers a natural paradigm for analysis. Interpreting the linear model parameters to reflect the interactivity of genes and their protein products in the cells' response to treatment affords comprehensive identification of genes that respond to treatment in different ways. Such versatility comes at a price, however, specifically in terms of the assumptions required to tie known biology to the observable data. This study underscores the paramount importance of carefully coordinating known cellular behavior with statistical modeling and experimental design in order to make confident, biologically relevant inference using microarray data from factorial experiments. This paper also serves as an introduction to the basic biology that is important for analyzing microarray data from a factorial experiment on cell lines designed to investigate the details of molecular interactions.

Linear model experimental design principles from the classic agricultural studies conducted by Fisher (1951) have been used to investigate two-color cDNA microarray data in a series of papers by Kerr et al. (2000), Kerr and Churchill (2001), and Kerr et al. (2002). Many of the terms in their linear model are devoted to normalization of the fluorescence intensities, and the design element of their problem emphasizes dye-color and array assignment for evaluating the effects of one "variety" or biological factor. Dudoit et al. (2002) extend similar design considerations to include replicate arrays, and propose a permutation algorithm for multiple comparisons adjustment in a one-treatment setting. Mixed models for microarray data are considered by Wolfinger et al. (2001), also to assess differential expression due to treatment conditions treated as one multi-level factor. While these methods do apply linear models to microarray data, their experimental design concerns are quite different. These authors aim to improve confidence in differential levels of expression on cDNA arrays when only one treatment is of biological interest by controlling for experimental artifacts. We use Fisher's design principles to focus on complex multifactor experimental designs and the interpretation of the linear model parameters when investigating the mechanisms that drive cellular function. For multifactor experiments done on two-color cDNA arrays, the two approaches could be combined. In this paper, we discuss data from an experiment using Affymetrix chips, and so we do not need to include the results from the other authors.

It is well known that, in general, genes do not behave independently of each other and some parametric models that allow for correlation between genes have been proposed. For example, Ibrahim et al. (2002) describe a Bayesian model for detecting differentially expressed genes in cancer versus normal tissues. In this model, the subgroup-specific correlation between genes depends on the number of patients in the normal and tumor subgroups, as well as two variance parameters in the hierarchical prior structure. The authors suggest that the model may be biologically plausible if the genes under investigation are close to each other on the same chromosome, or if the genes have similar chemical and biological behaviors.

The difficulty with parametric models that require a specified covariance structure for the genes is that they do not and, given the current state of scientific knowledge, cannot accurately reflect the biological systems that drive gene co-regulation, particularly when highly specific, consistent effects of treatments on genes within controlled cellular contexts are of interest. For example, the protein products of genes usually function in complexes (Hartwell et al., 1999; Gavin et al., 2002; Ho et al., 2002), and evidence suggests that the proteins of co-expressing genes tend to interact more frequently than genes with different expression profiles (Ge et al., 2001). Other regulatory mechanisms, for example the binding of transcription factors to specific promoter sites, drive the expression of many genes in a coordinated fashion . At this point, insight into protein complex composition and the transcription factors governing gene regulation is insufficient to accurately model joint behavior even for small sets of known co-regulated genes, let alone the thousands of genes that are measured using microarrays.

While correct multivariate analyses are certainly preferable to univariate analyses, we proceed with a gene-at-a-time technique to show how candidate genes demonstrating biologically interpretable expression profiles can provide a starting point from which to understand and possibly construct the regulatory systems in the cell. Zhao et al. (2001) also find gene-at-a-time analyses for parametric modeling to be useful in time-course investigations. Once the genes that result from our selection criteria are studied in detail, and more knowledge exists about the correct mechanisms of joint behavior, the univariate linear model can be easily be extended with the appropriate covariance structure to jointly model gene expression. Since the response of these genes under certain treatments conditions is well-summarized by the linear model, the selected genes would also serve as excellent candidates for the multivariate association determination model of Kim et al. (2000), or the network investigations of Wagner (2001) and Ideker et al. (2001).

In Section 2 of this paper, we summarize two basic cellular processes, transcription and translation, which inform the appropriate analysis and design of microarray factorial experiments. In Section 3, we describe the objectives of a specific experiment on cells from the MCF-7 breast cancer cell line (Soule et al., 1973) and why the factors were chosen to address these objectives. In Section 4, we discuss how mRNA abundance reflects the cellular mechanisms under investigation and state the biological assumptions necessary for modeling the data. In Section 5, we propose a linear model for the data from the MCF-7 experiment and explain how the parameters capture the mRNA abundance patterns noted in Section 4. Section 6 addresses outlier detection for this particular data set. Section 7 outlines an algorithm for data analysis according to the statistical model proposed in Section 5. Section 8 presents the biological results for this analysis, and compares them to results from a similar investigation. Section 9 proposes extensions of this methodology to answer other questions in the study. Section 10 addresses some limitations and advantages of the analysis and design of factorial experiment microarray data using linear models.

2. TRANSCRIPTION AND TRANSLATION

The nucleus of almost every cell in an organism contains a copy of DNA for that organism. DNA consists of four nucleotides, adenosine (A), cytosine (C), thymine (T), and guanine (G), and provides the pattern for manufacturing the functional components of a cell, namely *proteins*. The entire DNA sequence, called the *genome*, is divided into smaller regions, called *genes*, which code for specific proteins based on their A, C, T, G sequence. Cells behave differently according to their gene expression pattern; even though cells may have the same genes, different activity levels of these genes direct the protein composition, and therefore the function, of individual cells.

Genes code for proteins through intermediary nucleotides called *mRNA*. In a process called *transcription*, the nucleotide uracil (U) substitutes for T, and then (A,U) and (C,G) base-pairing directs the formation of a gene's complementary mRNA sequence. In a subsequent process called *translation*, the mRNA strands serve as templates for the formation of proteins. The correspondence of DNA to mRNA to protein explains the relevance of gene expression in characterizing functional differences in cells. In general, the amount of mRNA transcribed by a gene directs the amount of resulting protein. Microarrays measure abundance of mRNA transcripts to quantify gene expression since they closely relate to protein abundance, but are easier to work with in the lab (Lockhart et al., 1996).

DNA is a stable molecule that remains largely unchanged throughout the life of a cell. mRNA and proteins, on the other hand, are degradable molecules that may exist only in different types of cells, at different times in a cell's life, or in response to external stimuli. Certain proteins called *transcription factors* control the dynamic expression of genes seen in varying amounts of mRNA and protein. Treatments in cell line microarray experiments often target specific transcription factors to accomplish desired changes in the cells. Transcription factors bind to a gene's *promoter* sequence that is typically upstream from the actual coding region of the gene and thereby control mRNA production. Promoter sequences vary among genes and allow for specificity of a transcription factor for its target genes. Transcription factors may increase or decrease the expression of a gene, they may act in conjunction with other transcription factors, they may bind to similar promoter sequences in several genes, and they may compete with other transcription factors for binding to the same promoter. Furthermore, transcription factors are themselves proteins which often require regulation by other genes. In such a complex, interacting system, changing the activity of one transcription factor may change the expression of its primary (or direct) gene targets, but a number of secondary (or indirect) effects may also occur.

The living, dynamic system into which treatments enter make factorial experiments for measuring gene expression in cell lines somewhat different from classical factorial experiments discussed in statistical literature. While parameter estimation is procedurally the same, the interpretation of the parameters must account for multiple directions and magnitudes of the effects, possible interactions of the treatments, and perhaps most importantly, the interactions of the genes, their respective protein products, and other cellular components under the treatment conditions. Simplifying assumptions must be made in order to draw biologically relevant conclusions from the data, but as in any modeling context, reasonable assumptions often allow for reasonable conclusions.

3. THE EXPERIMENT

In this section, we describe in some detail the biology for our particular experiment of interest. The intention in this section is to demonstrate the design considerations and assumptions that must be made when tying the biological details of a factorial microarray experiment to a statistical linear model. Early collaborations between biologists and statisticians that frame molecular questions in terms of statistical parameters can afford more realistic expectations for the types of biological observations and conclusions allowed by a given experimental design.

The data discussed in this paper comes from two experiments on cells from an estrogen receptor positive (ER+) human breast cancer cell line (MCF-7). There were three factors in the experiments, each at two different levels (P=present or A=absent): cyclohexamide (CX), estrogen (ES, specifically 17β-estradiol), and treatment Z. (Z is an unnamed compound that is still under investigation.) The experiments were identical except for the length of time of treatment exposure before data collection (10 and 48 hours). For each time point, sixteen independent samples of the MCF-7 cells were exposed in replicate to one of the eight possible combinations of the three factors, for a total of thirty-two samples. By treating time as a fourth factor with two levels, the experiments can be jointly analyzed as one experiment with a 2^4 factorial design. mRNA transcript abundance was measured using Affymetrix® HGU95Av2 microarrays. Model-based expression indices were computed in dChip using the PM-only model after normalization by the Invariant Set Method (Li and Wong, 2001a,b). All other statistical analyses were done in R (Ihaka and Gentleman, 1996) using packages from Bioconductor (http://www.bioconductor.org).

The main objective of this study was to identify ES target genes, and to further differentiate between primary and secondary targets (more explicit definitions will be made in Section 3). The second objective of this study was to determine which of the primary ES targets also respond to Z. In this paper, we will focus on the first objective regarding the identification of primary and secondary ES targets at a single time point (48 hours). Although this is a 2^4 factorial experiment, we begin as though there are only two factors, CX and ES, at two levels, P and A. The discussion of this as a 2^2 factorial experiment simplifies the considerations in using the experimental design to answer specific biological questions. Straightforward extensions incorporate the other factors.

ES target identification is possible in MCF-7 cells since they are ER+. In breast cancer cells, ER+ refers to the presence of estrogen receptor (ER) in the cell. When ES enters the cellular environment, ER binds to ES, and then acts as a transcription factor for specific genes. The best understood mechanism of transcriptional control by ER involves the binding of an ER subtype called ERα to estrogen response elements (EREs) within the promoter sequence of its target genes (Katzenellenbogen, 1996). ERα, as well as the less understood estrogen receptor β (ERβ), also modulate gene transcription through other promoters, including AP-1 sites (Paech et al., 1997). Since our experimental design does not allow us to observe individual ES-stimulated pathways, in this paper, we make no attempt to distinguish between the mechanisms through which the addition of ES accomplishes changes in gene expression. We simply note the observed expression levels and look to identify any changes that might be attributable to ES. Since the MCF-7 cells are ER+, the introduction of ES should at the very least stimulate transcriptional regulation of the genes with EREs targeted by ERα.

CX is a universal translational inhibitor; that is, it prevents the translation of mRNA into protein. In cells treated with CX, we can observe the effects of ES and Z, independent of the production of new protein. Although CX is of tremendous value in differentiating between primary and secondary ES targets as explained in Section 4, it does complicate matters and is in fact the source of many assumptions made in the interpretation of the data.

4. PRIMARY AND SECONDARY ESTROGEN TARGETS

A *primary ES target* is a gene that is affected at the mRNA transcript level when ES is present by a mechanism that does not require the protein products of other ES-regulated genes, e.g. a gene with an ERE in its promoter that is directly recognized by ES-bound ERα. A *secondary ES target* is a gene that is affected at the transcript level downstream of an initial ES-gene interaction; that is, a gene whose expression is induced or suppressed by the protein product of another ES target gene's mRNA. The use of CX in this experiment facilitates the distinction between ES target types.

The biological activity of CX does require several assumptions for interpretation of the data, mainly related to protein degradation. Protein half-lives are quite variable, ranging from a few minutes to several days. Once a protein degrades, if CX is present, new protein cannot be translated to take its place. First, we assume that ER itself, and any other coregulating proteins that facilitate the activity of ER, do not degrade so quickly that they are absent in the cell before ES can affect the genes. Second, we assume that some proteins that exist in the cell at the time of CX addition do indeed degrade over the course of the experiment, and that this allows for effects of ES that might not otherwise be observed. We further assume that the proteins responsible for carrying out the degradation of other proteins do not themselves degrade to depletion during the experiment. Finally, we assume that many genes are differentially expressed in the presence of CX simply in response to its drastic effects on the cell, and therefore take appropriate measures in our analysis to account for these effects.

Figures 1-4 propose stylized, hypothetical mechanisms by which ES and CX might affect mRNA transcript levels of primary and secondary ES targets. These schemes are highly simplified; in reality, many more cellular components contribute to the process through which ES and CX affect the genes. Although these figures show joint expression of primary and secondary targets, the goal of this paper is not to decide which primary and secondary targets act in the same pathway (however, this is a topic for future research). Rather, these scenarios demonstrate the mRNA transcript abundance patterns that distinguish between primary and secondary targets as we might observe them. In all four figures, gene A (represented by the solid rectangle labeled A) is the primary ES target and gene B (represented by the solid rectangle labeled B) is the secondary ES target. mRNA is represented by rectangles with subscripts matching the corresponding genes; recall that microarrays measure mRNA transcript abundance so these levels are our observed data. Transcription factors are represented by ovals with plug-in shapes to match the indents in the gene-specific promoters. TF_C is a transcription factor that exists in the cell at initial exposure to treatment, but degrades fairly quickly, and so by the time of data collection is absent in the presence of CX. Transcription is represented by a single arrow with the number of '+' or '-' signs representing the strength of expression induction or suppression. CX is a lightning bolt and ER is a tulip shape. ES is a pie-shape that fits into ER. ER will only bind to the promoter of primary target gene A in the presence of ES. Table 1 summarizes the observed mRNA levels in Figures 1-4.

Consider Figure 1. Genes A and B are both inactive in the untreated cell. The addition of CX alone does not change the (in)activity of these genes since a transcription factor does not regulate their expression. Exposure to ES alone enables ER to induce expression of gene A, whose protein product TF_A then induces expression of gene B. When both CX and ES are present, gene A produces $mRNA_A$ as expected, but translational inhibition by CX prevents expression of gene B.

In Figure 2, ES-bound ER and TF_C compete for binding to the promoter for gene A. When there is no treatment, TF_C induces expression of gene A, and in turn, TF_A induces expression of gene B. In the presence of CX alone, TF_C is no longer available, so production of mRNA transcripts for both genes A and B stops. When ES is added alone, there is no change in the $mRNA_A$ transcript level since TF_C already causes gene A to express as much $mRNA_A$ as possible. When ES and CX are added together, TF_C is absent, but gene A is still expressed in response to the presence of ES. $mRNA_B$ is not present, however, due to the lack of TF_A.

In Figure 3, gene A is transcribed in response to TF_C, and produces TF_A which suppresses gene B. In the presence of CX alone, neither TF_C nor TF_A are present, allowing only moderate expression of gene A and high expression of gene B. When ES is added alone, the effect of TF_C is dampened, causing lower expression of gene A. There is then less TF_A, so only moderate abundance of $mRNA_B$. When both CX and ES are present, ES completely suppresses gene A, and gene B expresses at a high level in the absence of TF_A.

Finally, consider Figure 4. TF_C induces transcription of gene A, and TF_A induces transcription of gene B. In the presence of CX, both genes are not expressed due to the lack of the transcription factors that stimulate their expression. The addition of ES causes gene A to be expressed more than when TF_C is present alone, causing higher levels of both $mRNA_A$ and TF_A, and, in turn, higher levels of $mRNA_B$. When both CX and ES are present, gene A is still expressed due to the presence of ES, but again at a moderate level since TF_C is now absent. Gene B is no longer expressed due to the absence of TF_A.

Figures 1-4 represent a few highly simplified mechanisms by which ES might alter the expression of its gene targets. There are certainly many more possibilities, but these four examples demonstrate two patterns that distinguish between primary and secondary targets of ES. (1) For *primary targets*, ES will change the gene's mRNA transcript level in some way, whether or not CX is present. (2) For *secondary targets*, the mRNA transcript level that is observed when both CX and ES are present should be the same as the transcript level when CX is present alone. ES may have an effect when introduced alone, but CX should prevent any alteration in the expression of a secondary target gene caused by new protein synthesis.

Note that the use of CX as a factor in this experiment allows for convenient interpretations of the observable data in regards to primary and secondary transcriptional regulation by ES. Since microarrays measure mRNA transcript abundance, CX was particularly useful given the nature of the technology and the observable data for the genes of interest. This is an example of the type of conclusions that are possible by designing experiments with carefully selected factors under reasonable assumptions about the behavior of those factors. Other factors would provide similarly valuable information for other questions of interest. In Section 9, we consider extensions of the modeling framework to other questions in this particular study.

Given the variety of ways in which mRNA transcript levels of both primary and secondary targets change in the presence of ES, the challenge in this experimental setting is to develop efficient, comprehensive gene selection criteria that encompass the different combinations of effect size and direction. The statistical model presented in Section 5 aims to do just that.

5. STATISTICAL MODEL

In this section, we discuss the interpretation of the parameters in the linear model for the factorial designed estrogen target experiment. As stated in the Introduction, well-informed multivariate models would certainly be desirable for assessing the activity of genes in concert with other genes when exposed to various levels of treatment, but the current state of biological knowledge does not afford such sophisticated and realistic characterization of genetic interactivity. We adopt a gene-at-a-time univariate analysis of the treatment effects since the primary goal of the investigation is high-throughput identification of primary and secondary estrogen targets. Once the common biological regulators of the estrogen targets are better understood, then appropriate joint models can be formulated.

Consider the data for a single time point and exposure to the two levels (P and A) of ES and CX for a total of eight samples in a 2^2 factorial design. Suppose $\log y_{gi}$, the log of the observed expression level y_{gi} of gene g ($g = 1, ..., n$) in chip i ($i = 1, .., 8$), can be modeled as in (1) where $\epsilon_{gi} \sim N(0, \sigma^2)$ and x_{CXi} and x_{ESi} are indicator variables for the presence of CX and ES in sample i respectively.

$$\log(y_{gi}) = \mu_g + \beta_{gCX} x_{CXi} + \beta_{gES} x_{ESi} + \beta_{gCX:ES} x_{CXi} x_{ESi} + \epsilon_{gi} \tag{1}$$

The β_g coefficients represent gene-specific main effects for CX and ES and their two-way interaction; in what follows, the g subscripts are dropped, but the estimated parameters are still assumed to be gene-specific. Estimation of the coefficients using least squares is straightforward, and under the assumption of Normality, p-values for these estimates easily follow. A log transformation helps satisfy Normality in our data set, particularly when examining both time points. Model checking will be addressed in Section 6. While the actual model fitting and parameter estimation applies straightforward linear model theory, the interpretation of the parameters in this specific biological framework is the challenge.

For the purpose of ES target identification, the estimates of μ and β_{CX} are not particularly informative. μ describes the mRNA level in the cell without the addition of the factors, and β_{CX} measures the effect on the gene of preventing new protein synthesis. The parameters β_{ES} and $\beta_{\mathrm{CX:ES}}$ describe the effects of ES on the gene, and so direct the selection of ES targets. Specifically, we can test the null hypothesis H_{0ESt} (2) using nested linear models, and if we reject H_{0ESt}, conclude that the gene is affected by ES at the transcriptional level and is therefore an ES target.

$$H_{0ESt} : \beta_{ES} = \beta_{CX:ES} = 0 \tag{2}$$

Since this hypothesis test is performed individually on thousands of genes, multiple comparisons is an obvious issue. We control for the expected proportion of falsely rejected hypotheses using the false discovery rate (FDR) method of Benjamini and Hochberg (1995). In our case, a rejected hypothesis indicates an ES target, so we can interpret the FDR as the proportion of falsely identified ES targets. The application of FDR also acts as a natural filtering tool to reduce the number of genes under investigation to a manageable size.

Multiple testing procedures for microarray data continue to be debated. The FDR method tends to result in more genes for which the test of hypothesis is rejected compared to other standard multiple testing procedures; biologists often prefer a longer list of credible genes for further investigation. It is also comparable to the empirical Bayes microarray analysis procedure for one two-level factor proposed by Efron et al. (2001) and is suggested by Dudoit et al. (2002) to be appropriate for microarray data. Leave-one-out cross validation, which has been used in some microarray analyses, is inappropriate in this study due to the structure of the experimental design. Permutation tests that reasonably accomodate the structure in this data are of interest and are currently under investigation.

Once ES targets are identified, the next step is to differentiate between primary and secondary ES targets. Consider the mRNA transcript abundance pattern for primary target gene A in Figure 1. When ES is added alone, we observe an increase in mRNA_A, indicating that $\beta_{ES} > 0$. When ES is added in the presence of CX, we see the same increase in mRNA_A, so $\beta_{CX:ES} = 0$. $\beta_{ES} > 0$ and $\beta_{CX:ES} = 0$ correspond to Condition 1 in Table 2. There is a similar correspondence between the mRNA_A levels in Figures 2-4 and Conditions 2-4. Note that Conditions 5-8 in Table 2 are simply directional opposites of Conditions 1-4. These 8 combinations for the values of β_{ES} and $\beta_{CX:ES}$ summarize all of the different ways in which we can observe expression alteration by ES in presence of CX as we should for primary targets. In fact, these 8 conditions can be summarized as $\beta_{ES} + \beta_{CX:ES} \neq 0$. This can be seen by considering all (in)equality possibilities of the parameters β_{ES} and $\beta_{CX:ES}$ in relation to 0 (<,>,=).

Now consider the mRNA transcript abundance pattern for secondary target gene B in Figure 1. When ES is introduced alone, we observe an increase in mRNA_B, indicating that $\beta_{ES} > 0$. When ES is added in the presence of CX, the amount of mRNA_B is the same as when CX was added alone, indicating that $\beta_{CX:ES} < 0$ and in fact $\beta_{ES} + \beta_{CX:ES} = 0$. In all Figures 1-4, $\beta_{CX:ES}$ in some sense "reverses" β_{ES} so that the observed mRNA_B levels are the same in the presence of CX, regardless of the addition of ES. This is consistent with the biology of secondary targets in that ES should not have an effect if translation cannot occur.

The fact that $\beta_{ES} + \beta_{CX:ES} \neq 0$ for primary targets and $\beta_{ES} + \beta_{CX:ES} = 0$ for secondary targets provides concise criteria for differentiating between the two. Specifically, for the ES targets that are selected according to H_{0ESt} (2), we can test the linear contrast

$$H_{0pt} : \beta_{ES} + \beta_{CX:ES} = 0, \tag{3}$$

and if we reject H_{0pt} (3), conclude that the gene is a primary ES target. Otherwise, it is a secondary ES target.

Selecting ES targets and differentiating between primary and secondary ES targets is completely reliant upon the data we can observe in this experiment. Note the secondary ES target gene B in Figure 2. For this gene, $\beta_{CX} < 0$, but $\beta_{ES} = \beta_{CX:ES} = 0$. We would fail to select this gene as an ES target at all, let alone identify it as a secondary ES target. This is a limitation of this particular experimental design in that the downstream effect of ES for this gene cannot be observed. There may well be similar situations in this experiment; other studies with different experimental designs might prove useful in improving upon the list of ES target genes.

An added feature of the linear contrast selection technique in (3) for primary targets is its relationship to fold change criteria that is often used to select induced or suppressed genes. Consider the ratio of the mRNA level for a gene for cells treated with ES and CX to those treated with CX alone. For log-transformed data, the fold change null hypothesis H_{0fc} for differential expression would be

$$H_{0fc} : \frac{\exp(\mu + \beta_{CX} + \beta_{ES} + \beta_{CX:ES})}{\exp(\mu + \beta_{CX})} = \exp(\beta_{ES} + \beta_{CX:ES}) = 1. \quad (4)$$

A gene for which H_{0fc} (4) is rejected might be considered a primary ES target since any change in expression in the presence of both ES and CX, when compared to CX alone, signifies expression alteration directly due to ES. Algebraically, the equation in H_{0fc} (4) can be reduced to the contrast criteria proposed in H_{0pt} (3). More restrictive criteria for fold change magnitude can be imposed by changing H_{0fc} (4) to reflect the fold change induction or suppression of interest.

6. OUTLIER DETECTION

The replicate structure of the experiment under consideration allows for a specialized algorithm for outlier detection. Residuals from the regression model are generally used for model checking in the linear model framework; however, the constraints imposed on the residuals by the factorial design render conventional methods for outlier detection questionable. Normality and independence are better satisfied by taking advantage of the replicate observations in this data set.

For each individual gene, consider the squared differences between r sets of replicate observations, say $d_1^2, ..., d_r^2$, where $d_i^2 = (\log(y_{i1}) - \log(y_{i2}))^2$ for all $i = 1, ..., r$. If $\log(y_{i1})$ and $\log(y_{i2})$ are Normally distributed with variance $\sigma_i^2/2$ for all $i = 1, ..., r$, then the d_i^2/σ_i^2 are χ_1^2 random variables. Under a null hypothesis of constant variance $\sigma_i^2/2 = \sigma^2/2$ for all $i = 1, ..., r$,

$$f = \frac{d_{(r)}^2}{\frac{1}{r-1}\sum_{i=1}^{r-1} d_{(i)}^2} \quad (5)$$

is $F_{1,r-1}$ where $d_{(i)}^2$ is the ith order statistic ($i = 1, ..., r$). An adjusted p-value to test for suspected outliers can be calculated by $r * P\{F(1, r-1) > f\}$, the adjustment taken since f compares the maximum squared difference of replicate observations to the mean of the rest. When $f \geq r - 1$, this adjusted p-value is exact, otherwise it is an upper bound for the true p-value.

Since this statistic evaluates differences between observations, it only identifies which pair the outlier observation may come from, not the outlier observation itself. For some genes, such as those in Figure 5a, the outlier observation is easily identified. In other cases, such as the observations in Figure 5b and 5c, pairs with large differences between the replicates are identified, but it is inappropriate to label one observation as an outlier rather than the other. For this reason, once pairs with significantly large differences are identified, a second criteria is applied to exclude questionable cases from being labeled as single outliers. In particular, if only one of the tagged replicates falls outside the range of (med_e-4*mad_e,med_e+4*mad_e) where med_e is the median of the expression values for the gene and mad_e is the median absolute deviation of the expression values for the gene, the observation is designated the single outlier.

In our analysis, single outliers were discarded and parameter estimates computed based on the remaining observations. Other outlier detection methods, such as those discussed in Carey et al. (1997), could also be employed.

7. SELECTION ALGORITHM

Our algorithm for primary and secondary ES target selection is outlined below. The algorithm can be suitably adjusted for other experiments.

1. Average the replicate observations and exclude any genes with a maximum average less than 100. (For Affymetrix data, PM-only expression indices that are less than 100 are taken to indicate genes that are not expressed.) Remove all Affymetrix control sequences.

2. Apply any necessary transformations to satisfy Normality, and then test for single outliers as in Section 5. If outliers are detected, remove them from the data set.

3. Fit the linear model proposed in Equation (1) for each gene.

4. Test $H_{0ESt} : \beta_{ES} = \beta_{CX:ES} = 0$ (2) for each gene.

5. Reject H_{0ESt} (2) for the genes with the lowest resultant p-values using an FDR of 0.01. Call these genes *ES targets*.

6. For the ES targets, test $H_{0pt} : \beta_{ES} + \beta_{CX:ES} = 0$ (3). Call all ES target genes with p-values < 0.01 for the test of H_{0pt} (3) *primary ES targets*. Call the remaining ES target genes *secondary ES targets*.

8. RESULTS

The results presented here are from an analysis of the sixteen observations in the 48 hour experiment. We adjusted step 3 in the algorithm in Section 7 and fit a linear model for the log-transformed expression values with main effects for CX, ES, and Z, and all possible two-way interactions. In principle, the smaller 2^2 factorial experiment can be analyzed, but in practice there are substantial benefits in terms of power to including more observations.

For our application, Step 1 in the gene selection algorithm reduced the number of genes under investigation from 12625 to 12352. Step 2 detected 49 single outliers. Step 5 selected 305 ES target genes. Step 6 identified 130 of these as primary ES targets, and 175 as secondary ES targets. The specific primary and secondary target genes are reported at http://www.bioconductor.org/Docs/Papers/2002/FactorialDesign.

The observations for three genes that we identified as primary targets are recorded in Figure 6. The contrast test p-value for H_{0pt} (3) is included in these plots (contrast p), as is the nominal p-value for the β_{ES} parameter estimate (ES p). In these plots, observations on the left (right) of the dotted line are for samples that were unexposed (exposed) to CX. All samples exposed to Z are white circles, and can be discounted for now. Gray (black) circles are for samples unexposed (exposed) to ES. $TFF1$ (trefoil factor 1, $pS2$) and $ERBB2$ ($HER2/NEU$) respectively exhibit increased and decreased expression in the presence of ES, regardless of the presence of CX. mRNA transcript abundance for $CYP1B1$ (cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile)) increases in the presence of ES only when CX is also present. These three plots demonstrate the variety of direction and magnitude of expression differences that our selection criteria detect. The comparison of the observed expression levels in the presence of CX to those in the presence of CX and ES suggest that ES affects transcription of these genes without requiring intermediate proteins thus satisfying the definition of primary ES target in Section 3.

Figure 7 demonstrates the mRNA levels for three genes we identified as secondary targets. The expression of $BRCA1$ (breast cancer 1, early onset) is increased in the presence of ES alone, but when ES is added in the presence of CX, the expression level remains the same as when CX is added alone. ES triggers a decrease in the expression of $GADD45A$ (growth arrest and DNA-damage-inducible alpha) and JUN (v-jun sarcoma virus 17 oncogene homolog (avian)), but this same decrease does not transpire in the presence of CX. All three plots demonstrate patterns of secondary ES targets in that ES alters their expression level, but in the presence of CX, expression levels do not change. We ascribe this to the inhibition of translation, though there could be other reasons.

8.1 Comparison to Other Target Selection Criteria

The use of CX as one of the experimental treatments in this study motivates a comparison of the genes we identify as ES targets with those reported by Soulez and Parker (2001). They examined ZR75-1 breast cancer cells treated with CX and one of three antiestrogens (tamoxifen, raloxifen, or faslodex) to ZR75-1 cells treated with CX and 17β-estradiol. Potential ES targets were selected by examining fold change ratios under antiestrogen and ES conditions. Since Soulez and Parker's experiments were done in the presence of CX, the ES targets they identify are presumably primary ES targets. Although the experimental designs are not identical, and the different cell lines could account for discordant findings, this recently published gene list does present a natural point of comparison for our results.

Of the 53 primary ES target genes reported in Soulez and Parker (2001), we identified Affymetrix probes for 49 of the genes. We applied Steps 2-6 of Section 7 to these 49 genes without adjusting for FDR and found that 14 of the genes were identified as primary targets, 6 were secondary targets, and 29 were not targets at all. $UGCG$ (UDP-glucose ceramide glucosyltransferase, Figure 8a) is a gene from Soulez and Parker's list that we identified as a secondary target; note the expression induction by ES in the absence of CX and lack thereof in the presence of CX. This is similar to the other 5 genes that were similarly identified. The 29 genes from Soulez and Parker's list that we do not identify as targets at all generally exhibited quite variable expression values, one example is $BAG1$ (BCL2-associated athanogene) in Figure 8b.

Since we did not involve antiestrogen treatment in our study, it is difficult to apply Soulez and Parker's exact selection criteria to our data; however, we did apply two different forms of fold change estimation and examined the 49 genes with the highest magnitude of fold change in either direction. (We chose 49 for comparison with the 49 Affymetrix probe sets we were able to identify out of Soulez and Parker's list.) The first fold change method (FC1) compared the average of the expression levels of the two samples exposed to both ES and CX to the average of the expression levels of the two samples exposed to just CX:

$$FC1 = \frac{\text{mean(two samples exposed to CX and ES)}}{\text{mean(two samples exposed to CX)}}. \tag{6}$$

The second fold change method (FC2) compared the expression estimates from the linear model for the CX and ES vs. CX condition:

$$FC2 = \frac{\exp(\hat{\mu} + \hat{\beta}_{CX} + \hat{\beta}_{ES} + \hat{\beta}_{CX:ES})}{\exp(\hat{\mu} + \hat{\beta}_{CX})} = \exp(\hat{\beta}_{ES} + \hat{\beta}_{CX:ES}). \tag{7}$$

Of the top 49 genes resulting from FC1 (6), 26 were included in our primary target list. Of the top 49 genes resulting from FC2 (7), 25 were included in our primary target list. Although the expression patterns for the remaining genes selected by these criteria were reasonably reflective of primary ES target genes, the observations were more variable. The apparent altered expression in the presence of ES cannot be convincingly attributed to ES rather than random variation. Given the fact that there are only two replicates for each experimental condition, small variation within the pair is required for a believable distinction as a primary target of ES.

8.2 Comparison with Genes Reported in Literature

There are several primary ES target genes described in the literature, and in general, our methodology selects these genes as primary targets. $TFF1$ (trefoil factor 1, $pS2$) (Masiakowski et al., 1982), $CCND1$ (cyclin D1, $PRAD1$) (Planas-Silva and Weinberg, 1997), $CTSD$ (cathepsin D) (Westley and May, 1987), and $CYP1B1$ (cytochrome P450, subfamily 1, (dioxin-inducible) polypeptide 1 (glaucoma 3, primary infantile)) (Hayes et al., 1996) are all well-characterized, long-standing primary ES targets. $ERBB2$ ($HER2/NEU$) is a breast cancer diagnostic marker whose expression is inhibited by direct transcriptional repression by ES (De Bortoli et al., 1996). An experiment involving CX demonstrated that $GREB1$ (*KIAA0575* gene product) is directly controlled by ES-bound ER (Ghosh et al., 2000). Similarly, CX does not prevent an increase in mRNA levels for $TRIM16$ ($EBBP$, tripartite motif-containing 16) (Liu et al., 1998) due to ES-bound ER. We identify all of these documented primary ES targets.

Some reported primary ES targets are not contained in our list, for example $NRIP1$ (nuclear receptor interacting protein 1) (Cavailles et al., 1995). $NRIP1$ was not identified due to the multiple comparisons adjustment using FDR. When the contrast test H_{0pt} (3) is applied to the observations for $NRIP1$, the p-value is less than 0.01. This omission demonstrates the familiar difficulty with multiple comparisons adjustment and p-value cutoffs in that some positive results are not identified. *LIV-1* (*LIV-1* protein, estrogen regulated) is another direct ES target (el Tanini and Green, 1996) that we did not identify. In our data set, the observations for $LIV-1$ were too variable for distinction as an ES target.

Our secondary target list is also confirmed by recent literature and suggests that there is great potential for network construction using the ES target genes we report. $BRCA1$ is a known secondary ES target (Marks et al., 1997) that is, among other things, a transcription factor for many genes including $GADD45A$ (growth arrest and DNA-damage-inducible, alpha) (Fan et al., 2002). $BARD1$ ($BRCA1$ associated RING domain 1) and $RBBP8$ (retinoblastoma binding protein

8, $CtIP$) co-regulate transcriptional control by $BRCA1$ (Irminger-Finger and Leung, 2002; Sum et al., 2002). All four of these genes are reported on our secondary target list.

JUN and FOS are known to participate with ERα in transcriptional control at AP1 promoter sites (Teyssier et al., 2001). $DDIT3$ (DNA-damage-inducible transcript 3, $GADD153$) is known to be transcriptionally regulated, at least in part, by AP1 transcription factors (Qiao et al., 2002). All three of these genes are reported on our secondary target list.

9. EXTENSIONS

The methodology discussed for the two factors CX and ES can easily be extended to questions requiring information from the other factors. Consider a linear model including all four main effects for CX, ES, Z, and TIME and possible interactions for $\log y_{gi}$, the log of observed expression level y_{gi} of gene g ($g = 1, ..., n$) in chip i ($i = 1, .., 32$), with error term $\epsilon_{gi} \sim N(0, \sigma^2)$.

Suppose we want to find genes that behave like primary ES targets at the transcriptional level at 10 and/or 48 hours. This requires differential expression of genes in the presence of CX compared to the presence of CX and ES at one or both time points, and can be tested using the simultaneous set of contrasts in (8).

$$H_0 : \begin{cases} \beta_{ES} + \beta_{CX:ES} = 0 \\ (\beta_{ES} + \beta_{ES:TIME}) + (\beta_{CX:ES} + \beta_{CX:ES:TIME}) = 0 \end{cases} \tag{8}$$

Recall the second objective in this experiment regarding the modification of the effect of ES by Z. For simplicity, consider only one time point. The genes of interest should exhibit differential expression in the presence of Z and ES when compared to ES alone. If the null hypothesis $H_0 : \beta_Z + \beta_{ES:Z} = 0$ is rejected, then there is either a change in the expression of the gene directly due to a main effect of Z or the combination of the main effect and the ES:Z interaction. In either case, the expression of these genes in the presence of ES is altered by Z.

Now consider genes that, in the presence of ES, are regulated by Z through transcriptional control rather than translational control. Since CX is a translational inhibitor, the expression levels in the presence of CX and ES should be different from those when CX, ES, and Z are all present. In this case, rejecting the null hypothesis $H_0 : (\beta_Z + \beta_{CX:Z}) + (\beta_{ES:Z} + \beta_{CX:ES:Z}) = 0$ would identify genes that are affected by RAP in the presence of ES through posited direct transcriptional control. The behavior of Z in the presence of ES and CX at both time points could be investigated by coordinating all four factors in the experiment.

10. DISCUSSION

This paper demonstrates the careful coordination of experimental design and biological function of the factors in a microarray study for the identification of primary and secondary ES targets. No less important was the coordinated efforts of biologists and statisticians in making sense of the resultant data. Incorporating universal translational inhibition by CX and the transcription factor behavior of ER upon binding to ES into the interpretation of the model enabled the identification of genes of interest using standard statistical linear modeling. Multiple comparisons adjustment according to FDR provided a natural data reduction step and helped quantify the proportion of falsely identified genes. Statistical significance according to the p-values for the tests of H_{0ESt} (2) and H_{0pt} (3) provide a measure of confidence in the genes that were selected as ES targets.

The use of CX in this experiment facilitated convenient interpretations for the parameters in the statistical model under the assumption of universal translational inhibition in the presence of CX. Since microarrays are designed to measure mRNA transcript abundance, and given our interest in transcriptional regulation by ES, this was a particularly useful factor for addressing the main objective of this study. Factors with such dramatic effects in the cell are not necessary in these types of experiments. Section 8 in this paper addressed the use of $\beta_Z + \beta_{ES:Z}$ to investigate alteration of expression in the presence of ES due to Z. Factors with smaller effects can provide just as interpretable results depending on the research question of interest.

The principles behind this specific example are applicable to any carefully designed microarray study. Investigators can extend their analyses beyond gene filtering according to fold change, or expression profile similarity according to clustering techniques. Instead, they can efficiently select genes according to biologically interpretable parameters with a quantifiable measure of confidence.

Often times, the goal of a microarray experiment is to identify genes that might compose a functional pathway or network in the cell with the aim to explore these genes individually using gene-at-a-time experiments. Wagner (2001) discusses the reconstruction of genetic networks using various gene perturbations including point mutations, gene deletions, over-expression, or inhibition of translation, and then proposes a graph-theoretic framework for inferring direct effects of these manipulations. Gene selection according to statistical parameters from carefully designed microarray studies, such as the one described here, would provide much better candidates for these types of experiments than large-scale clustering or filtering analyses. The known interactions between several of the genes on our primary and secondary list underscore the great promise for future research in constructing genetic networks using these types of experiments that may inform the more desirable multivariate models of gene behavior.

The limitations in this methodology are the same as those in any setting involving multiple comparisons or linear models. We used the Benjamini and Hochberg (1995) false discovery rate for multiple comparisons adjustment since this tends to give longer gene lists and therefore more options for further investigation. Regarding the linear model, as in any factorial experiment, the higher-order interaction terms are somewhat difficult to interpret individually. Under appropriate assumptions about the biology, the mechanisms that we suspect drive the observations of mRNA transcript abundance do somewhat alleviate this problem. We must also make the familiar Normality assumption for computing statistical significance for all of the parameters in the linear model.

Despite the limitations, the use of experimental design in the analysis of microarray data can be used to efficiently and confidently address specific questions about cellular function that are reflected in mRNA transcript abundance. Once the connection between the observed data and the statistical parameters is understood, there is tremendous flexibility in selecting genes that demonstrate biological behaviors of interest. The design of microarray experiments with this type of analysis in mind will greatly enhance the characterization of gene behavior.

REFERENCES

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., others, Powell, J., , Yang, L., Marti, G., Moore, T., Hudson, J., Lu, J., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenberger, J., Armitage, R., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, D., and Staudt, L. (2000), "Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling," *Nature*, 403, 503–511.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999), "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Sciences*, 96, 6745–6750.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society B*, 57, 289–300.

Bhattacharjee, A., Richards, W., Stuanton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E., Lander, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D., and Meyerson, M. (2001), "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Sub-classes," *Proceedings of the National Academy of Sciences*, 98, 13790–13795.

Carey, V., Walters, E., Wagner, C., and Rosner, B. (1997), "Resistant and Test-Based Outlier Rejection: Effects on Gaussian One- and Two- Sample Inference," *Technometrics*, 39, 320–330.

Cavailles, V., Dauvois, S., L'Horset, F., Lopez, G., Hoare, S., Kushner, P., and Parker, M. (1995), "Nuclear Receptor RIP140 Modulates Transcriptional Activation by the Estrogen Receptor," *The EMBO Journal*, 14, 3741–3751.

De Bortoli, M., Maggiora, P., Cappello, D., Antoniotti, S., Saviozzi, S., Sapei, M., and Dati, C. (1996), "Hormonal Control of Growth Factor Receptor Expression," *Annals of the New York Academy of Sciences*, 784, 336–348.

DeRisi, J., Penland, L., Brown, P., Bittner, M., Meltzer, P., Ray, M., Chen, Y., Su, Y., and Trent, J. (1996), "Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer," *Nature Genetics*, 14, 457–460.

Dudoit, S., Yang, Y., Callow, M., and Speed, T. (2002), "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments," *Statistica Sinica*, 12, 111–139.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95, 14863–14868.

el Tanini, M. and Green, C. (1996), "Insulin/IGF-1 Modulation of the Expression of Two Estrogen Induced Genes in MCF-7 Cells," *Molecular and Cellular Endocrinology*, 121, 29–35.

Fan, W., Jin, S., Tong, T., Zhao, H., Fan, F., Antonire, M., Rajasekeran, B., Wu, M., and Zhan, Q. (2002), "BRCA1 regulates GADD45 through its Interactions with the OCT-1 and CAAT Motifs," *Journal of Biological Chemistry*, 277, 8061–8067.

Fisher, R. (1951), *The Design of Experiments*, Edinburgh: Oliver and Boyd Ltd., 6th ed.

Gasch, A., Spellman, P., Kao, M., Carmel-Ha, O., Eisen, M., Storz, G., Botstein, D., and Brown, D. (2000), "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes," *Molecular Biology of the Cell*, 11, 4241–4257.

Gavin, A., Bösche, M., Krause, R., Grandi, P., et al. (2002), "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes," *Nature*, 415, 141–147.

Ge, H., Liu, Z., Church, G., and Vidal, M. (2001), "Correlation between Transcriptome and Interactome Mapping Data from *Saccharomyces cerevisiae*," *Nature Genetics*, 29, 482–486.

Ghosh, M., Thompson, D., and Weigel, R. (2000), "PDZK1 and GREB1 are Estrogen-Regulated Genes Expressed in Hormone-Responsive Breast Cancer," *Cancer Research*, 60, 6367–6375.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caliguri, M., Bloomfield, C., and Lander, E. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Hartwell, L., Hopfield, J., Lebler, S., and Murray, A. (1999), "From Molecular to Modular Cell Biology," *Nature*, 402, C47–C52.

Hayes, C., Spink, D., Spink, B., Cao, J., Walker, N., and Sutter, T. (1996), "17β-Estradiol Hydroxylation Catalyzed by Human Cytochrome P450 1B1," *Proceedings of the National Academy of Sciences*, 93, 9776–9781.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G., et al. (2002), "Systematic Identification of Protein Complexes in *Saccharomyces cerevisiae* by Mass Spectrometry," *Nature*, 415, 180–183.

Ibrahim, J., Chen, M., and Gray, R. (2002), "Bayesian Models for Gene Expression with DNA Microarray Data," *Journal of the American Statistical Association*, 97, 88–99.

Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R., and Hood, L. (2001), "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network," *Science*, 292, 929–934.

Ihaka, R. and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Irminger-Finger, I. and Leung, W. (2002), "BRCA1-Dependent and Independent Functions of BARD1," *The International Journal of Biochemistry and Cell Biology*, 34, 582–587.

Katzenellenbogen, B. (1996), "Estrogen Receptors: Bioactivities and Interactions with Cell Signaling Pathways," *Biology of Reproduction*, 54, 287–293.

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N., and Churchill, G. (2002), "Statistical Analysis of a Gene Expression Microarray Experiment with Replication," *Statistica Sinica*, 12, 203–217.

Kerr, M. and Churchill, G. (2001), "Experimental Design for Gene Expression Microarrays," *Biostatistics*, 2, 183–201.

Kerr, M., Martin, M., and Churchill, G. (2000), "Analysis of Variance for Gene Expression Microarray Data," *Journal of Computational Biology*, 7, 819–837.

Kim, S., Dougherty, E., Chen, Y., Krishnamoorthy, S., Meltzer, P., Trent, J., and Bittner, M. (2000), "Multivariate Measurement of Gene Expression Relationships," *Genomics*, 67, 201–209.

Li, C. and Wong, W. (2001a), "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection," *Proceedings of the National Academy of Sciences*, 98, 31–36.

— (2001b), "Model-Based Analysis of Oligonucleotide Arrays (II): Model Validation, Design Issues and Standard Error Application," *Genome Biology*, 2, research0032.1–0032.11.

Liu, H., Golder-Novoselsky, E., Seto, M., Webster, L., McClary, J., and Zajchowski, D. (1998), "The Novel Estrogen-Responsive B-Box Protein (EBBP) Gene is Tamoxifen-Regulated in Cells Expressing an Estrogen Receptor DNA-Binding Domain Mutant," *Molecular Endocrinology*, 12, 1733–1748.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayahi, M., Horton, H., and Brown, E. (1996), "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays," *Nature Biotechnology*, 14, 1675–1680.

Marks, J., Huper, G., Vaughn, J., Davis, P., Norris, J., McDonnell, D., Wiseman, R., Futreal, P., and Iglehart, J. (1997), "BRCA1 Expression is not Directly Responsive to Estrogen," *Oncogene*, 14, 115–121.

Masiakowski, P., Breathnach, R., Bloch, J., Gannon, F., Krust, A., and Chambon, P. (1982), "Cloning of cDNA Sequeces of Hormone-Regulated Genes from the MCF-7 Human Breast Cancer Cell Line," *Nucliec Acids Research*, 10, 7895–7903.

Paech, K., Webb, P., Kuiper, G., Nilsson, S., Gustafson, J., Kushner, P., and Scanlan, T. (1997), "Differential Ligand Activation of Estrogen Receptors ERα and ERβ at AP1 Sites," *Science*, 277, 1508–1510.

Perou, C., Jeffrey, S., van de Rijn, M., Rees, A., Eisen, M., Ross, D., Pergamenschikov, A., Williams, C., Zhu, S., Lee, J., Lashkari, D., Shalon, D., Brown, P., and Botstein, D. (1999), "Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancers," *Proceedings of the National Academy of Sciences*, 96, 9212–9217.

Planas-Silva, M. and Weinberg, R. (1997), "Estrogen-Dependent Cyclin E-Cdk2 Activation through p21 Redistribution," *Molecular and Cellular Biology*, 17, 4059–4069.

Qiao, D., Im, E., Qi, W., and Martinez, J. (2002), "Activator Protein-1 and CCAAT/Enhancer-Binding Protein Mediated GADD153 Expression is Involved in Deoxycholic Acid-Induced Apoptosis," *Biochimica et Biophysica Acta*, 1583, 108–116.

Soule, H., Vazquez, J., Long, A., Albert, S., and Brennan, M. (1973), "A Human Cell Line from A Pleural Effusion Derived from a Breast Carcinoma," *Journal of the National Cancer Institute*, 51, 1409–1416.

Soulez, M. and Parker, M. (2001), "Identification of Novel Oestrogen Receptor Target Genes in Human ZR75-1 Breast Cancer Cells by Expression Profiling," *Journal of Molecular Endocrinology*, 27, 259–274.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., and Botstein, D. (1998), "Comprehensive Identification of Cell-Cycle Regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization," *Molecular Cell Biology*, 9, 3273–3297.

Sum, E., Peng, B., Yu, X., Chen, J., Byrne, J., Lindeman, G., and Visvader, J. (2002), "The LIM Domain Protein LMO4 Interacts with the Cofactor CtIP and the Tumor Suppressor BRCA1 and Inhibits BRCA1 Activity," *Journal of Biological Chemistry*, 277, 7849–7856.

Teyssier, C., Belguise, K., Galtier, F., and Chalbos, D. (2001), "Characterization of the Physical

Interaction between Estrogen Receptor α and JUN Proteins," *Journal of Biological Chemistry*, 276, 36361–36369.

Wagner, A. (2001), "How to Reconstruct a Large Genetic Network from n Gene Perturbations in Fewer then n^2 Easy Steps," *Bioinformatics*, 17, 1183–1197.

Westley, B. and May, F. (1987), "Oestrogen Regulates Cathepsin D mRNA Levels in Oestrogen Responsive Human Breast Cancer Cells," *Nucleic Acids Research*, 15, 3773–3786.

Wolfinger, R., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. (2001), "Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models," *Journal of Computational Biology*, 8, 625–637.

Zhao, L., Prentice, R., and Breeden, L. (2001), "Statistical modeling of large microarray data sets to identify stimulus-response profiles," *Proceedings of the National Academy of Sciences*, 98, 5631–5636.

Table 1: Observed mRNA Transcript Abundance in Figures 1-4

Figure 1			Figure 2		
Treatment	$mRNA_A$	$mRNA_B$	Treatment	$mRNA_A$	$mRNA_B$
None	Low	Low	None	High	High
CX	Low	Low	CX	Low	Low
ES	High	High	ES	High	High
CX,ES	High	Low	CX,ES	High	Low
Figure 3			**Figure 4**		
Treatment	$mRNA_A$	$mRNA_B$	Treatment	$mRNA_A$	$mRNA_B$
None	High	Low	None	Moderate	Moderate
CX	Moderate	High	CX	Low	Low
ES	Moderate	Moderate	ES	High	High
CX,ES	Low	High	CX,ES	Moderate	Low

Table 2: Parameter Combinations for Primary ES Targets

Condition no.	β_{ES}	$\beta_{CX:ES}$	Condition no.	β_{ES}	$\beta_{CX:ES}$
1	> 0	$= 0$	5	< 0	$= 0$
2	$= 0$	> 0	6	$= 0$	< 0
3	< 0	< 0	7	> 0	> 0
4[a]	> 0	< 0	8[a]	< 0	> 0

[a] $|\beta_{ES}| \neq |\beta_{CX:ES}|$

FIGURE CAPTIONS

Figure 1. **Possible Transcription Mechanism: Condition 1.** Primary ES target gene A and secondary ES target gene B are not active in the untreated samples, nor are they active in the presence of CX alone since proteins do not regulate their expression. When ES is added alone, gene A produces $\mathrm{mRNA}_{\mathrm{A}}$. The resulting transcription factor TF_{A} induces the transcription of gene B. When CX and ES are both present, ES-bound ER induces the transcription of gene A, but CX prevents translation of $\mathrm{mRNA}_{\mathrm{A}}$ into TF_{A}, and so gene B is not transcribed.

Figure 2. **Possible Transcription Mechanism: Condition 2.** TF_{C} induces transcription of primary ES target gene A, and in turn TF_{A} causes transcription of secondary ES target gene B. When CX is added alone, the absence of TF_{C} prevents the expression of both genes A and B. When ES is added alone, no difference is observed in $\mathrm{mRNA}_{\mathrm{A}}$ or $\mathrm{mRNA}_{\mathrm{B}}$ since TF_{C} competes with ES-bound ER for binding to gene A's promoter. When CX and ES are both present, ES-bound ER now induces the transcription of gene A, but CX prevents the formation of TF_{A} and gene B is not transcribed.

Figure 3. **Possible Transcription Mechanism: Condition 3.** P_{C} induces transcription of primary ES target gene A, and its protein product TF_{A} suppresses transcription of secondary ES target gene B. CX inhibits the production of TF_{C} and TF_{A}, causing moderate abundance of $\mathrm{mRNA}_{\mathrm{A}}$ and high abundance of $\mathrm{mRNA}_{\mathrm{B}}$. ES added alone suppresses transcription of gene A, and gene B is subsequently transcribed at a moderate level. When CX and ES are added together, gene A is completely suppressed, and gene B is highly active.

Figure 4. **Possible Transcription Mechanism: Condition 4**. Genes A and B are moderately expressed in the presence of TF_{C} and TF_{A}. CX prevents the production of both transcription factors, so neither gene is transcribed. ES alone enhances transcription of gene A, and in turn gene B is highly expressed. CX and ES added together results in moderate transcription of gene A, but the lack of new protein production due to CX prevents transcription of gene B.

Figure 5. **Examples of Primary Targets.** $TFF1$ (trefoil factor 1 (breast cancer, estrogen-inducible sequence expressed in)), $ERBB2$ ($HER2/NEU$), and $CYP1B1$ (cytochrome P450, subfamily I (dioxin-inducible), polypeptide 1 (glaucoma 3, primary infantile)) were identified as primary targets in our analysis. In all three examples, ES affects transcript levels in the presence of CX, indicating that these genes are primary ES targets.

Figure 6. **Examples of Secondary Targets.** $BRCA1$ (breast cancer 1, early onset), $GADD45A$ (growth arrest and DNA-damage-inducible, alpha), and JUN (v-jun sarcoma virus 17 oncogene homolog (avian)) were identified as secondary ES targets in our analysis. ES affects their transcript level only when CX is absent, indicating that the effect of ES is likely translationally mediated.

Figure 7. **Examples of Outliers. a)** For this gene, the single outlier is obvious. **b)** Both observations in the replicate pair with maximum difference lie *outside* the range of the other observations so we cannot label one as a single outlier. **c)** Both observations in the replicate pair with maximum difference lie *inside* the range of the other observations so we cannot label one as a single outlier.

Figure 8. **Examples from Soulez and Parker. a)** $UGCG$ was identified by Soulez and Parker (2001) as a primary ES target. We identified $UGCG$ as a secondary ES target. **b)** $BAG1$ was identified by Soulez and Parker (2001) as a primary ES target. In our data, the observations for $BAG1$ were too variable for distinction as an ES target.

Figure 1: Possible Transcription Mechanism: Condition 1

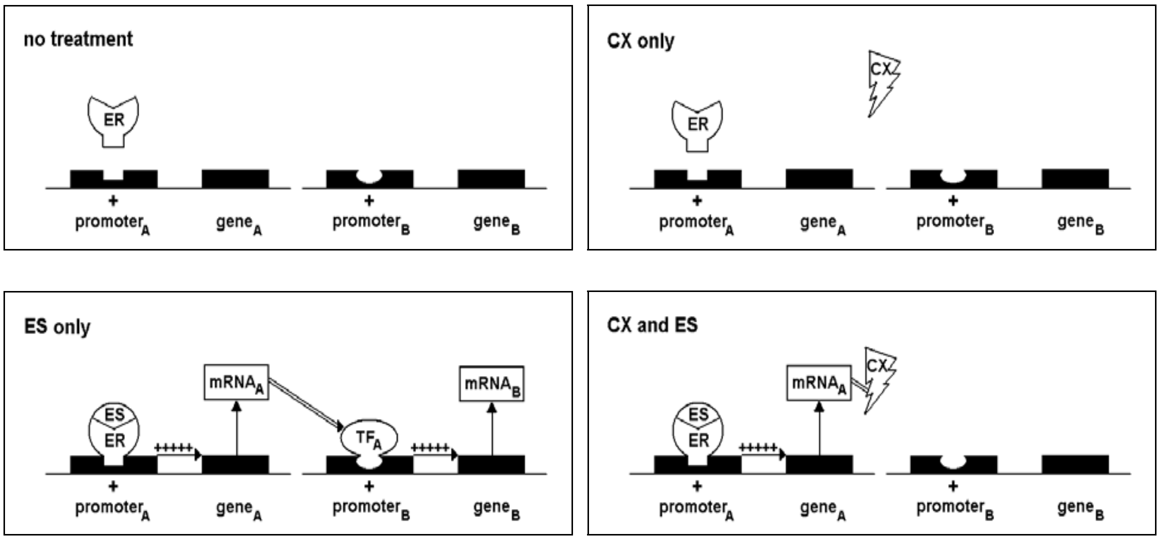

Figure 2: Possible Transcription Mechanism: Condition 2

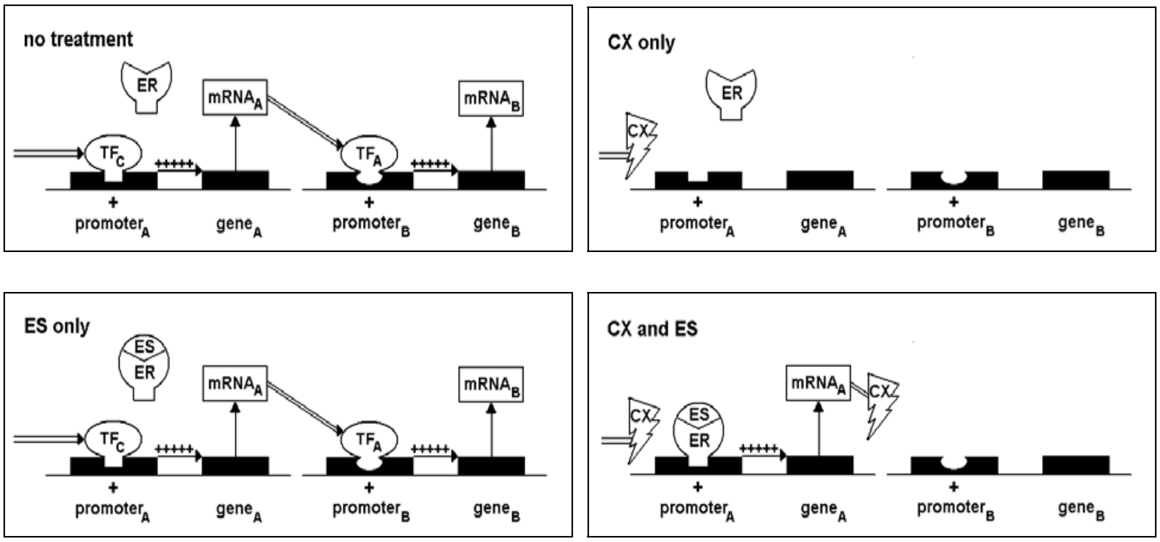

Figure 3: Possible Transcription Mechanism: Condition 3

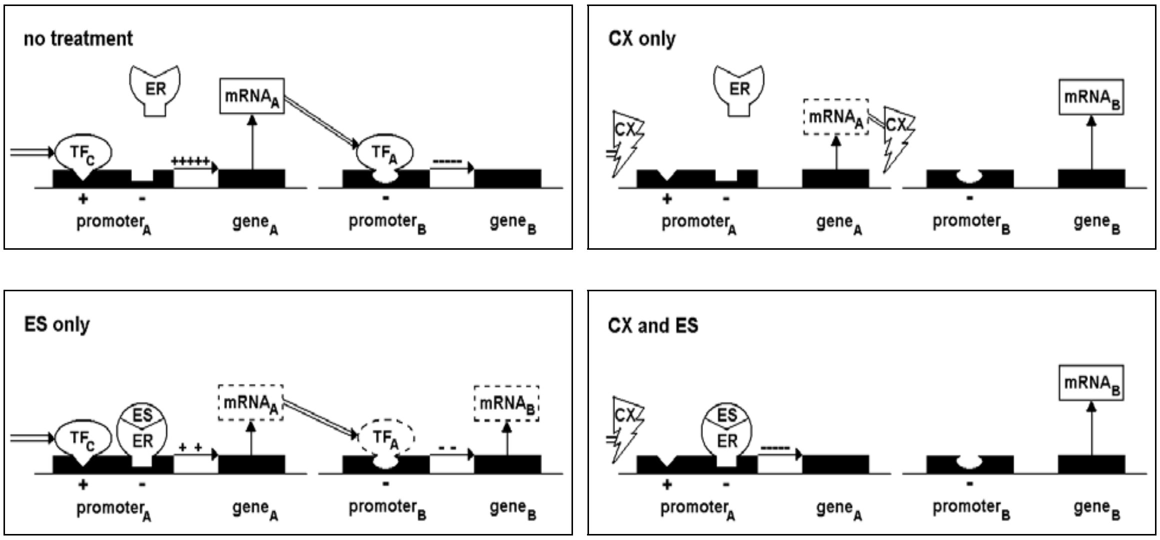

Figure 4: Possible Transcription Mechanism: Condition 4

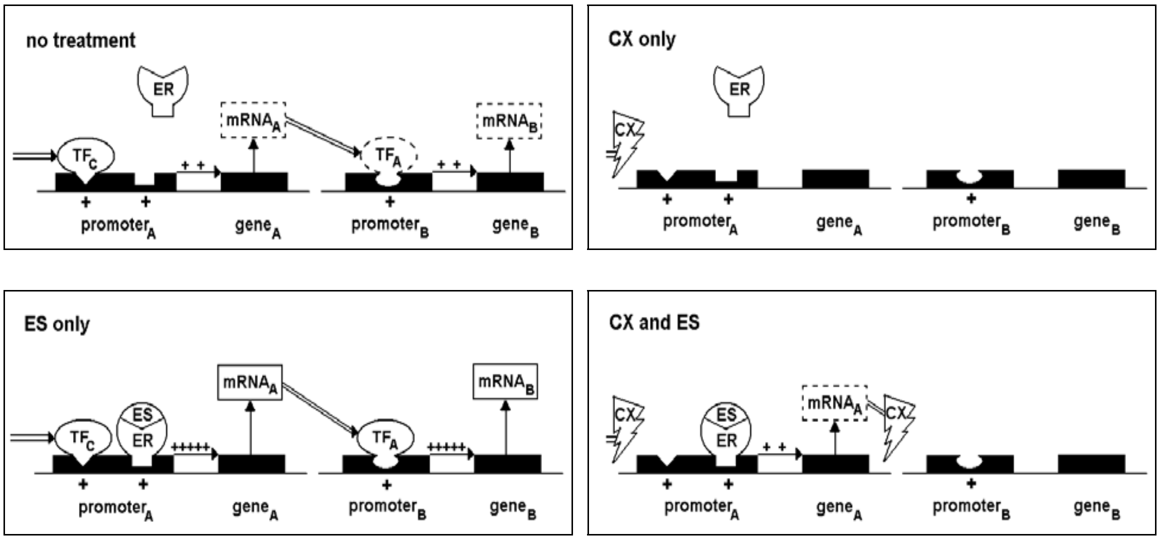

Figure 5: Examples of Outliers

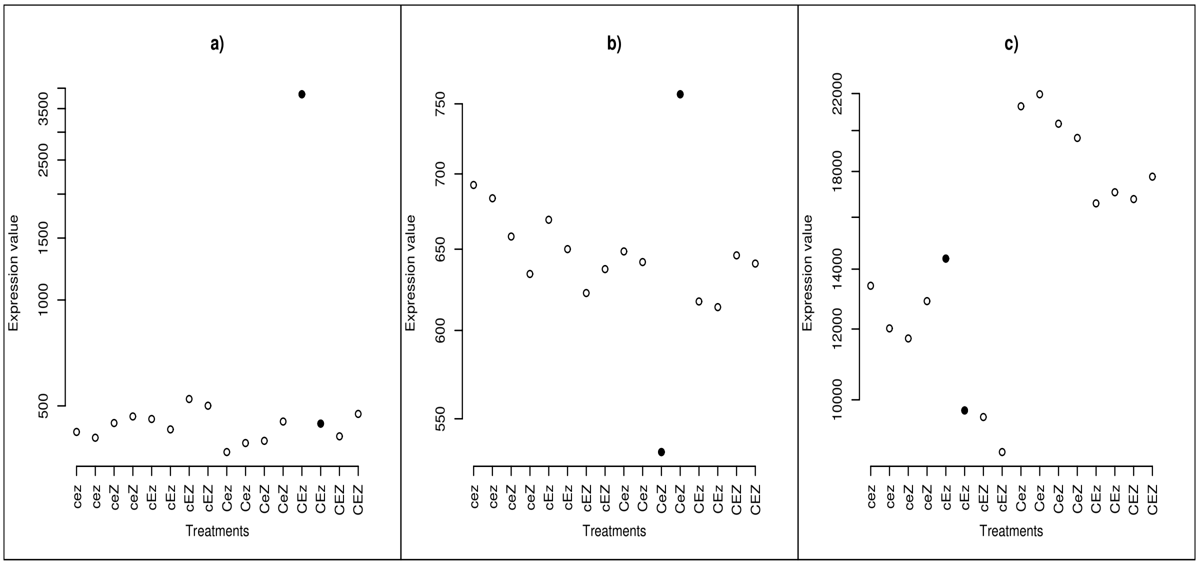

Figure 6: Examples of Primary Targets

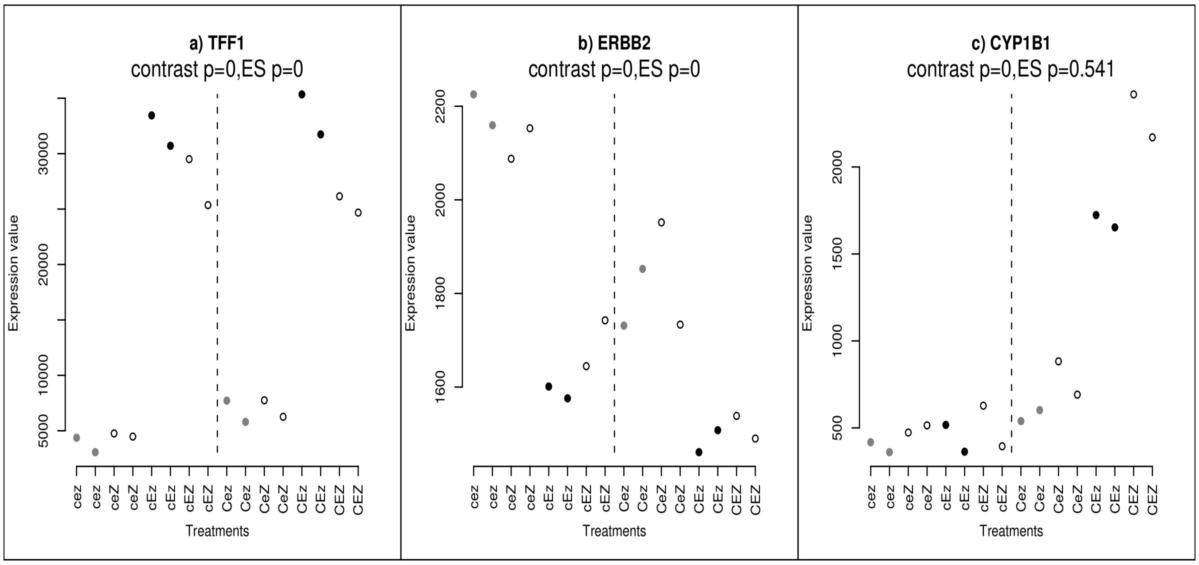

Figure 7: Examples of Secondary Targets

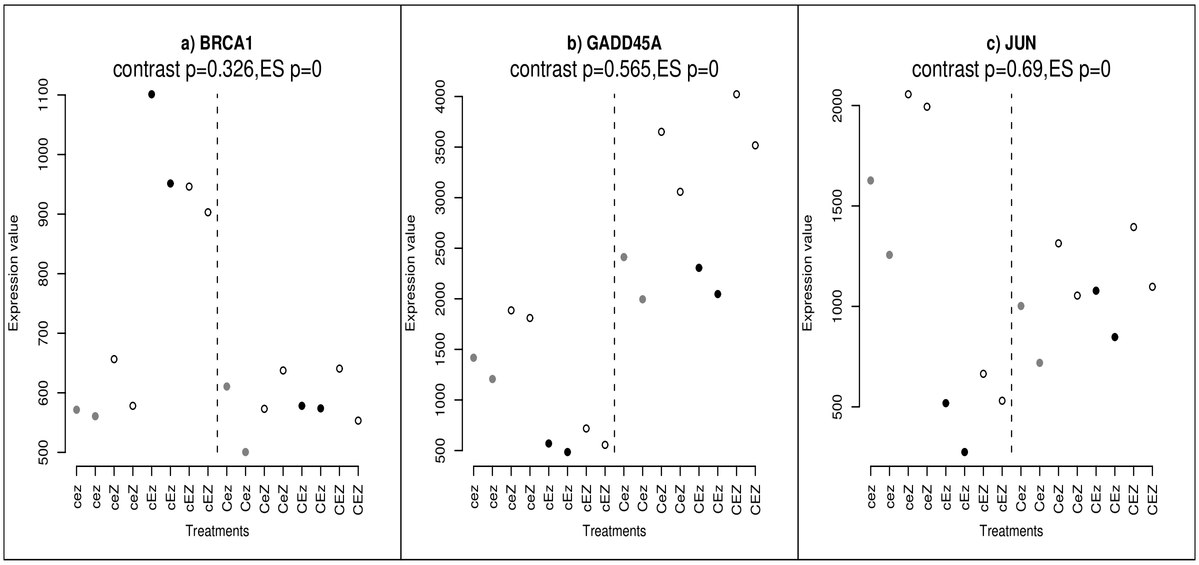

Figure 8: Examples from Soulez and Parker