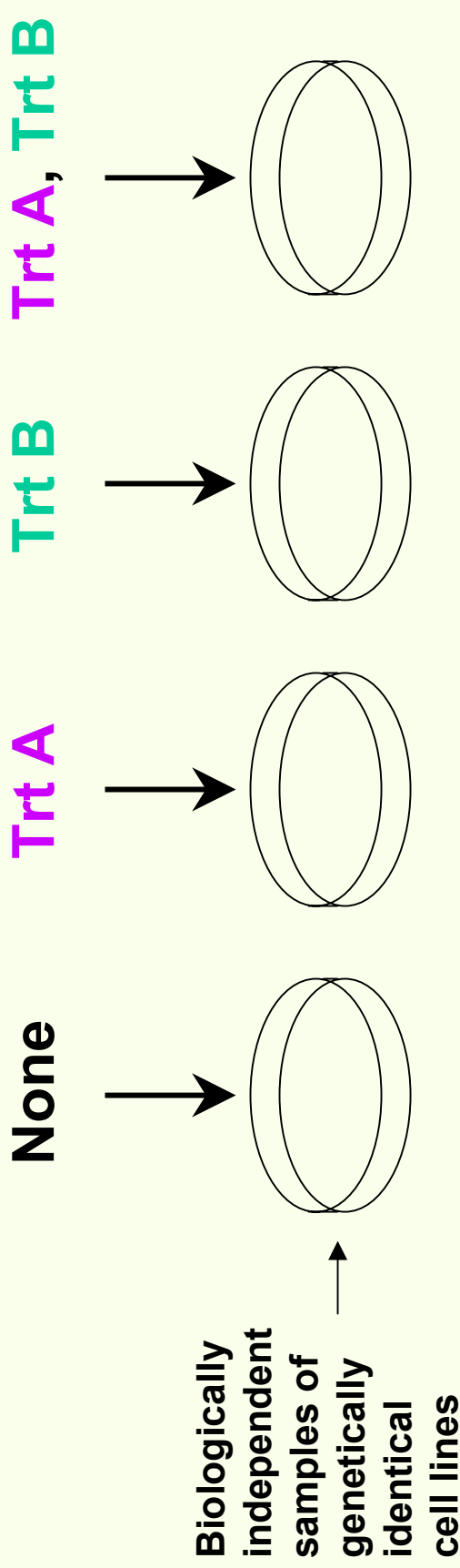


Analyzing Factorial Designed Microarray Experiments

Denise Scholtens,
Alexander Miron, Faisal Merchant,
Arden Miller, Penelope Miron,
J. Dirk Iglehart, Robert Gentleman

Factorial Designs for Microarray Experiments



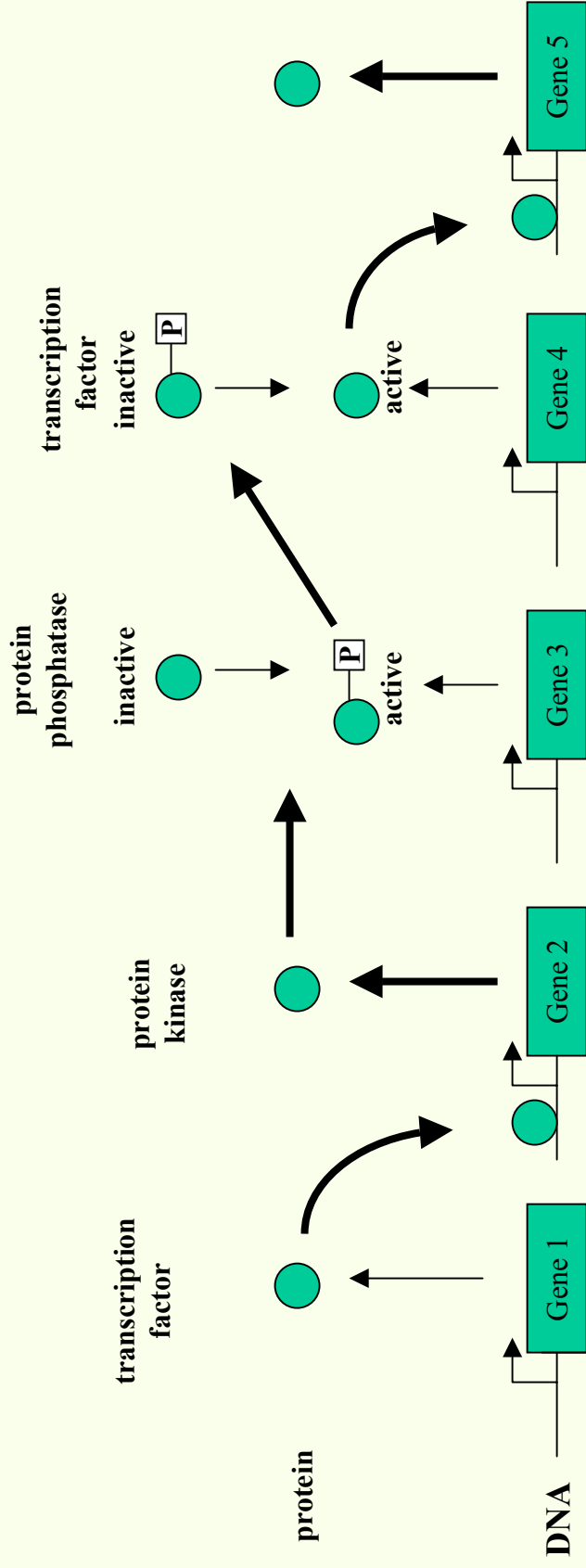
Questions: Can we pinpoint perturbations of genetic networks by these treatments? What biological interpretation can we give to the genes that we find?

Unique Challenge in Analyzing and Interpreting Factorial Experiment Microarray Data:

Treatments are applied in a living, dynamic cell.

mRNA abundance affected by:

transcription factors, protein complexes, methylation, phosphorylation,...



example of gene interactions,
adapted from Wagner (2001)

Investigative Concerns

- **Biologist:**
 - Which treatments will elucidate the biological mechanism under investigation?
- **Statistician:**
 - Existing machine learning techniques for microarray data analysis find broad expression pattern similarities, but we are interested in identifying specific perturbations of the network.
 - A classic linear model is appropriate for the experimental design, but the notion of “baseline” has changed.
 - Multivariate model?
 - Multiple testing?
- **Both:**
 - High-throughput gene selection
 - Interpretation of results

Estrogen Target Experiment

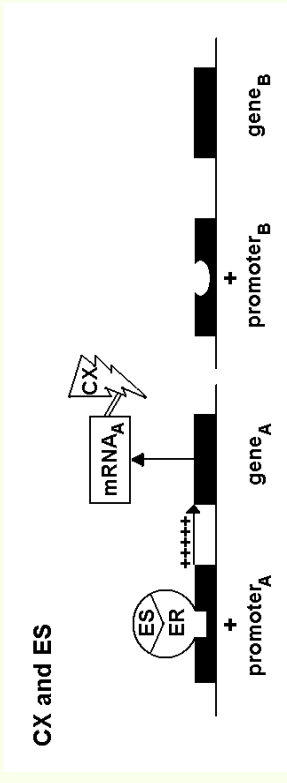
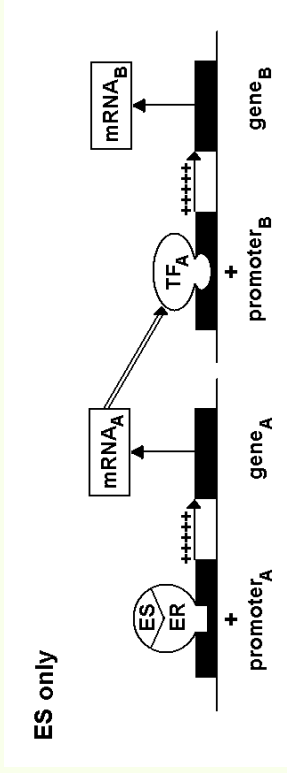
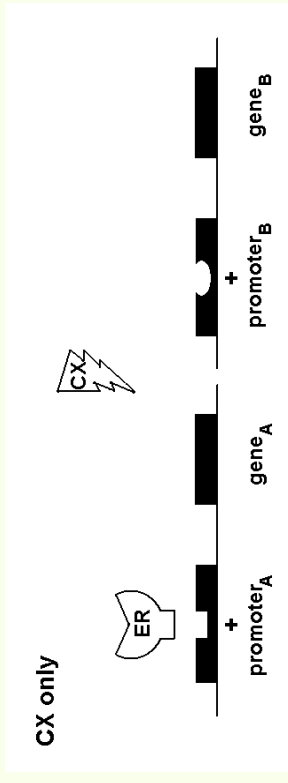
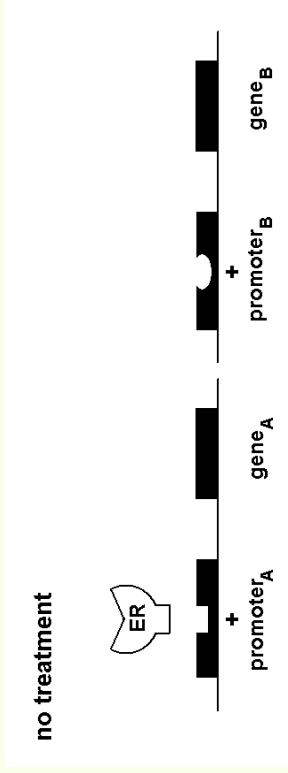
Experimental question:

- Which genes are **targets** of estrogen (ES), and can we differentiate between **primary** and **secondary** targets?
- **Target:** gene affected at the mRNA level in the presence of ES
- **Primary target:** gene that is directly activated or repressed at the mRNA transcript level by a mechanism that does not require the protein products of other ES-regulated genes,
 - e.g. A gene with an estrogen receptor element (ERE) in its promoter that is directly activated by ES-bound estrogen receptor α (ER).
- **Secondary target:** gene that is affected at the transcript level downstream of an initial ES-gene interaction;
 - e.g. A gene whose expression is induced or repressed by the protein product of another ES target gene

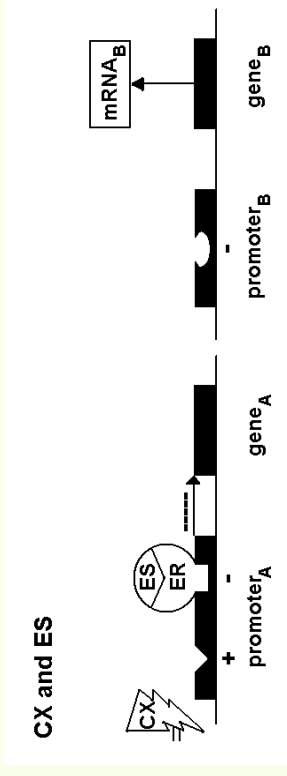
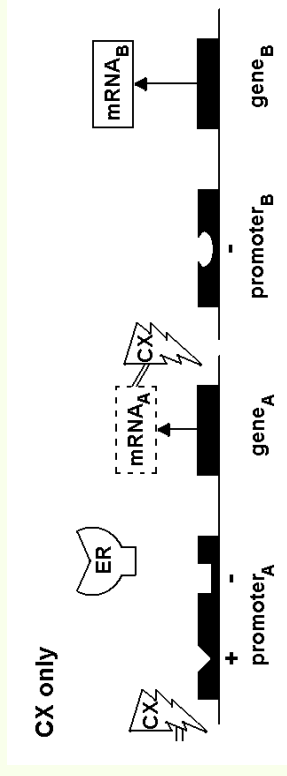
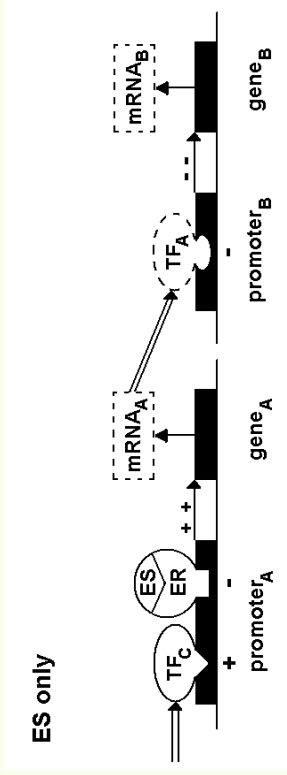
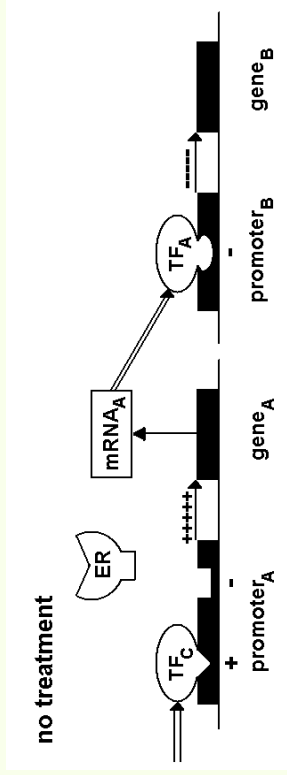
Experimental Design

- MCF-7 cells: ER+ breast cancer cell line
- Biologically independent replicates of each treatment condition in a 2x2 factorial experiment, giving 8 total samples.
- Factor 1: **estrogen (ES)**
 - Upon binding to ES, ER acts as a transcription factor for certain genes
- Factor 2: **cyclohexamide (CX)**
 - Universal translation inhibitor, i.e., mRNA can be transcribed, but it is not translated into protein
- *Note: The actual experiment was a 2⁴ factorial design with replicates for each experimental condition, but we will consider a simpler design for illustrative purposes.*

Changes in mRNA: Scenario 1



Changes in mRNA: Scenario 2



Linear Model

$$\log(Y_i) = \mu + \beta_{CX} X_{CX,i} + \beta_{ES} X_{ES,i} + \beta_{CX:ES} X_{CX,i} X_{ES,i}$$

$i = 1, \dots, 8$

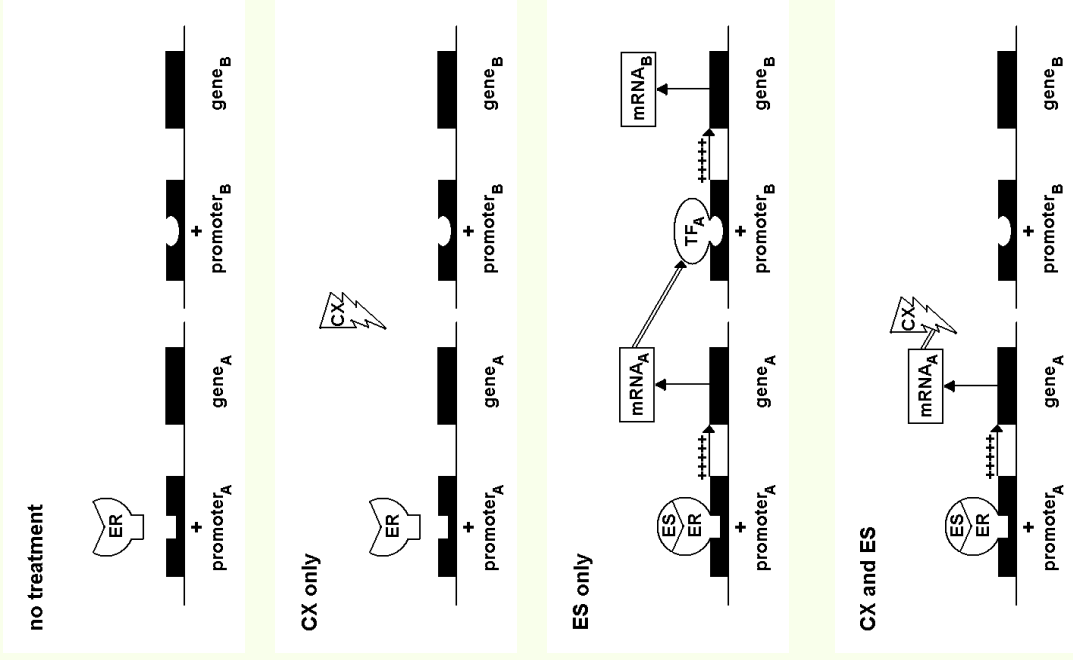
$$X_{CX,i} = \begin{cases} 1 & \text{if CX is present in } i^{\text{th}} \text{ sample} \\ 0 & \text{if CX is absent in } i^{\text{th}} \text{ sample} \end{cases}$$

$$X_{ES,i} = \begin{cases} 1 & \text{if ES is present in } i^{\text{th}} \text{ sample} \\ 0 & \text{if ES is absent in } i^{\text{th}} \text{ sample} \end{cases}$$

Linear Model

- Here we use univariate linear models since we do not yet fully understand the biological components that drive joint regulation.
- The estrogen targets that we identify using these models would be candidates for gene-at-a-time network perturbation models.

Linear Model Parameters: Scenario 1

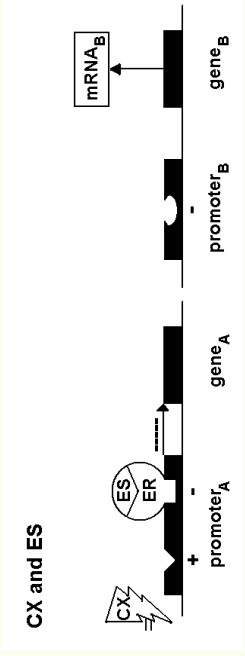
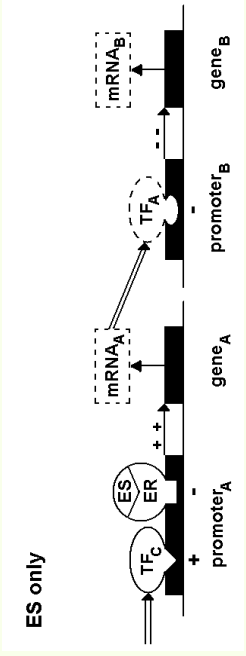
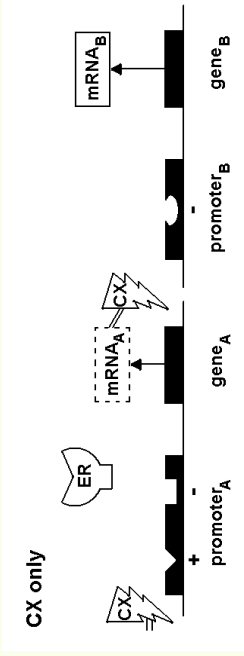
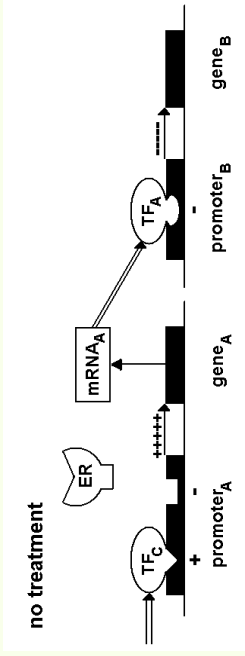


	mRNA _A	mRNA _B
β_{CX}	$=0$	$=0$
β_{ES}	>0	>0
$\beta_{CX:ES}$	$=0$	<0

Note for mRNA_A:
 $\beta_{ES} + \beta_{CX:ES} \neq 0$

Note for mRNA_B:
 $\beta_{ES} + \beta_{CX:ES} = 0$

Linear Model Parameters: Scenario 2



	mRNA _A	mRNA _B
β_{CX}	<0	>0
β_{ES}	<0	>0
$\beta_{CX:ES}$	$=0$	<0

Note for mRNA_A:
 $\beta_{ES} + \beta_{CX:ES} \neq 0$

Note for mRNA_B:
 $\beta_{ES} + \beta_{CX:ES} = 0$

ES Target Identification

- In this experiment, we can identify a gene as an ES target if

$$\beta_{ES} \neq 0 \text{ or } \beta_{CX:ES} \neq 0.$$

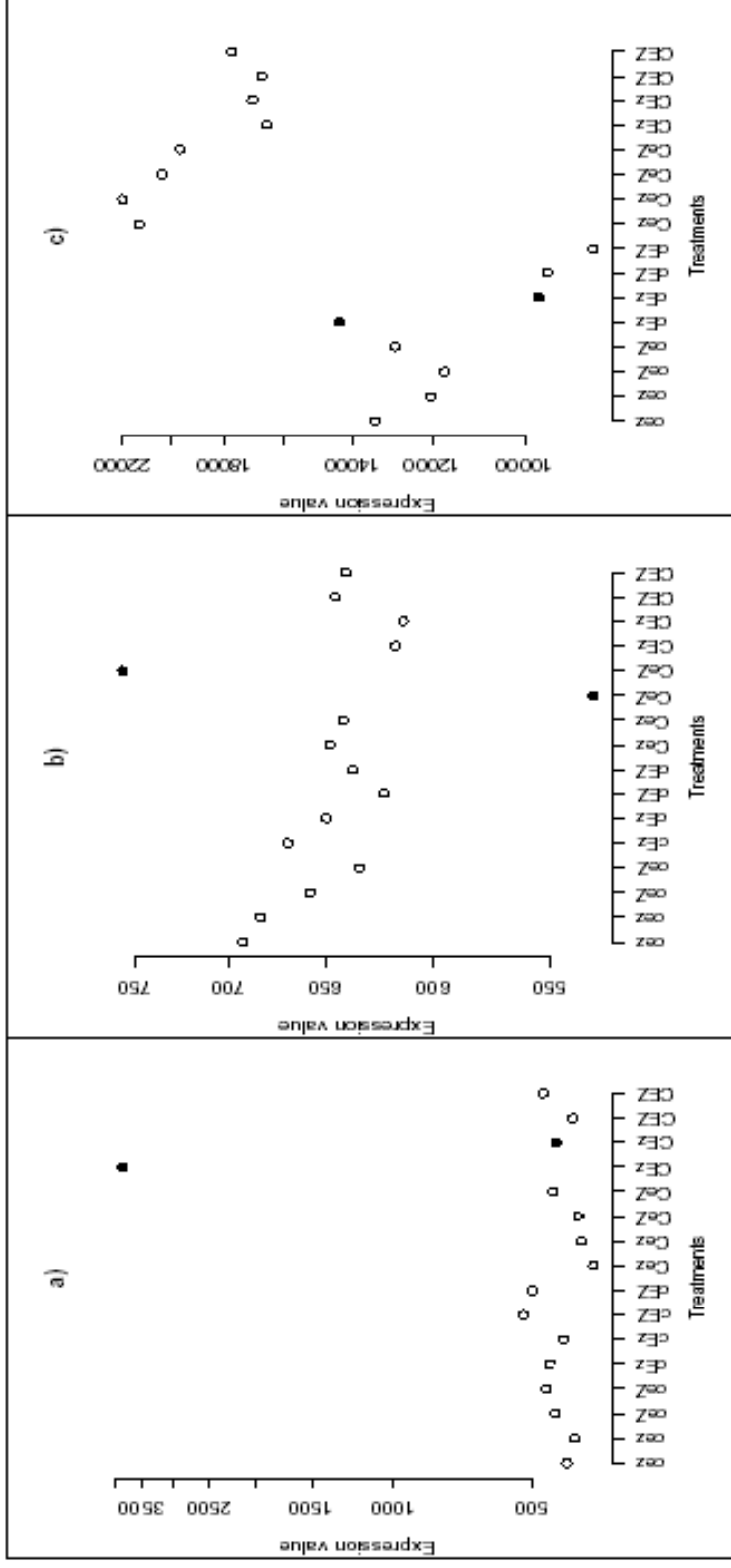
- If a gene is a **primary** ES target, then

$$\beta_{ES} + \beta_{CX:ES} \neq 0.$$

- If a gene is a **secondary** ES target, then

$$\beta_{ES} + \beta_{CX:ES} = 0.$$

Outliers



We use the replicate structure of the experimental design to locate single outliers in the data set. The algorithm is based on differences between the replicate expression values that are larger than expected.

Outlier Detection

Consider $d_i^2 = (\log(y_{i1}) - \log(y_{i2}))^2, i = 1, \dots, 4$.

If $\log(y_{i1}), \log(y_{i2}) \sim N(\mu_i, \sigma_i^2/2)$, then under

$$H_0 : \sigma_i^2 = \sigma^2, i = 1, \dots, 4,$$

$$f = \frac{d_{(r)}^2}{\frac{1}{r-1} \sum_{i=1}^{r-1} d_{(i)}^2} \sim F(1, r-1)$$

where $d_{(i)}^2$ is the i^{th} order statistic.

Then calculate an adjusted p - value by

$$4 * P\{F(1, r-1) > f\},$$

which is exact if $f \geq r-1$, and an upper bound otherwise.

Outlier Detection

- This method only identifies pairs with large differences, not the single outlier itself.
- Once pairs are identified, we designate single outliers if one of the tagged replicates falls outside the range

$$(\text{med}_e - 4 * \text{mad}_e, \text{med}_e + 4 * \text{mad}_e),$$

med_e = median of expression values,

mad_e = median absolute deviation
of expression values.

Multiple Testing

- Since we will form linear models and apply a test of contrast to tens of thousands of genes, we need to adjust for multiple testing.
- We use the False Discovery Rate (FDR) method of Benjamini and Hochberg (1995).
 - Tends to give longer lists of genes.
 - Since a rejected hypothesis indicates an ES target in our application, we can interpret the **FDR** as the *proportion of falsely identified ES targets*.

Gene Selection Algorithm

1. Average the replicate observations and exclude any genes with a maximum average less than 100 (using the PM-only model for gene expression in dChip). Remove all Affymetrix control sequences
2. Apply any necessary transformations to satisfy Normality, then test for single outliers. If outliers are identified, remove them from the data set.
3. Fit the linear model

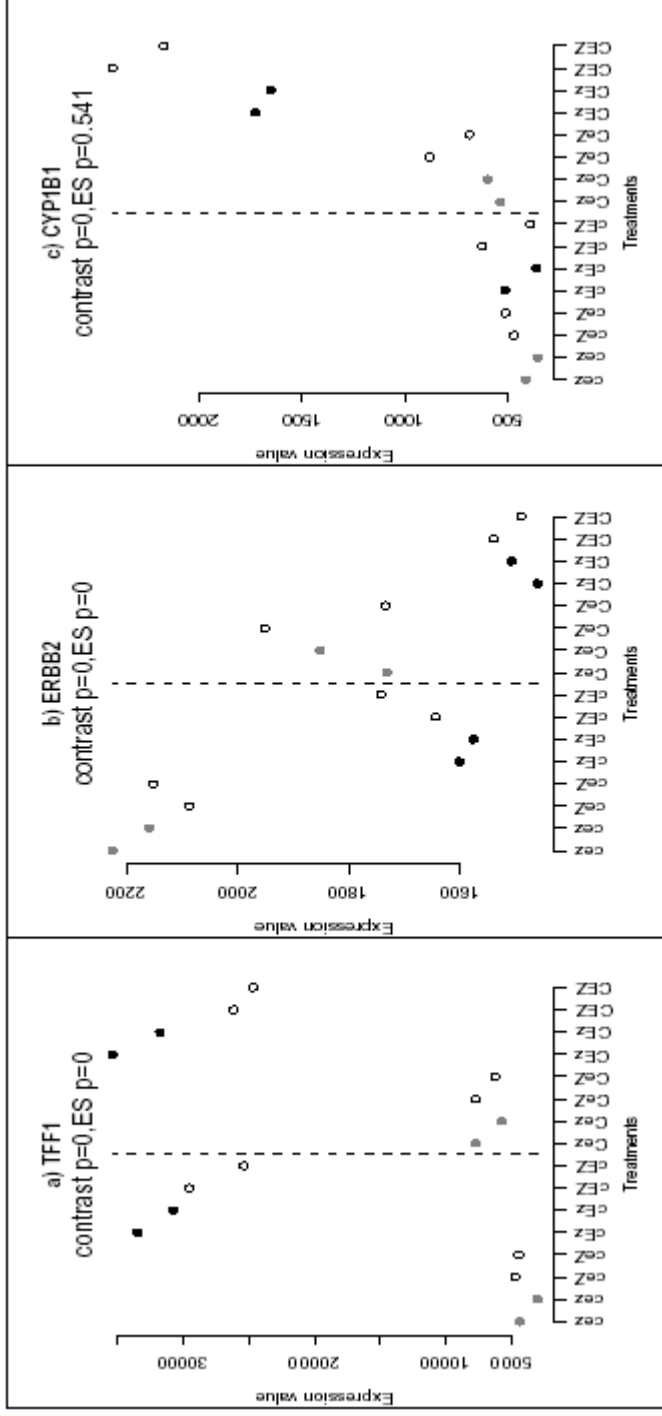
$$\log(Y_i) = \mu + \beta_{CX}X_{CX,i} + \beta_{ES}X_{ES,i} + \beta_{CX:ES}X_{CX,i}X_{ES,i}$$

for each gene.

Gene Selection Algorithm

4. Test $H_{0ES_t} : \beta_{ES} = \beta_{CX:ES} = 0$ for each gene.
5. Reject H_{0ES_t} 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 $H_{0pt} : \beta_{ES} + \beta_{CX:ES} = 0$. Call ES target genes with p -values < 0.01 for the test of H_{0pt} **primary ES targets**. Call the remaining ES target genes **secondary ES targets**.

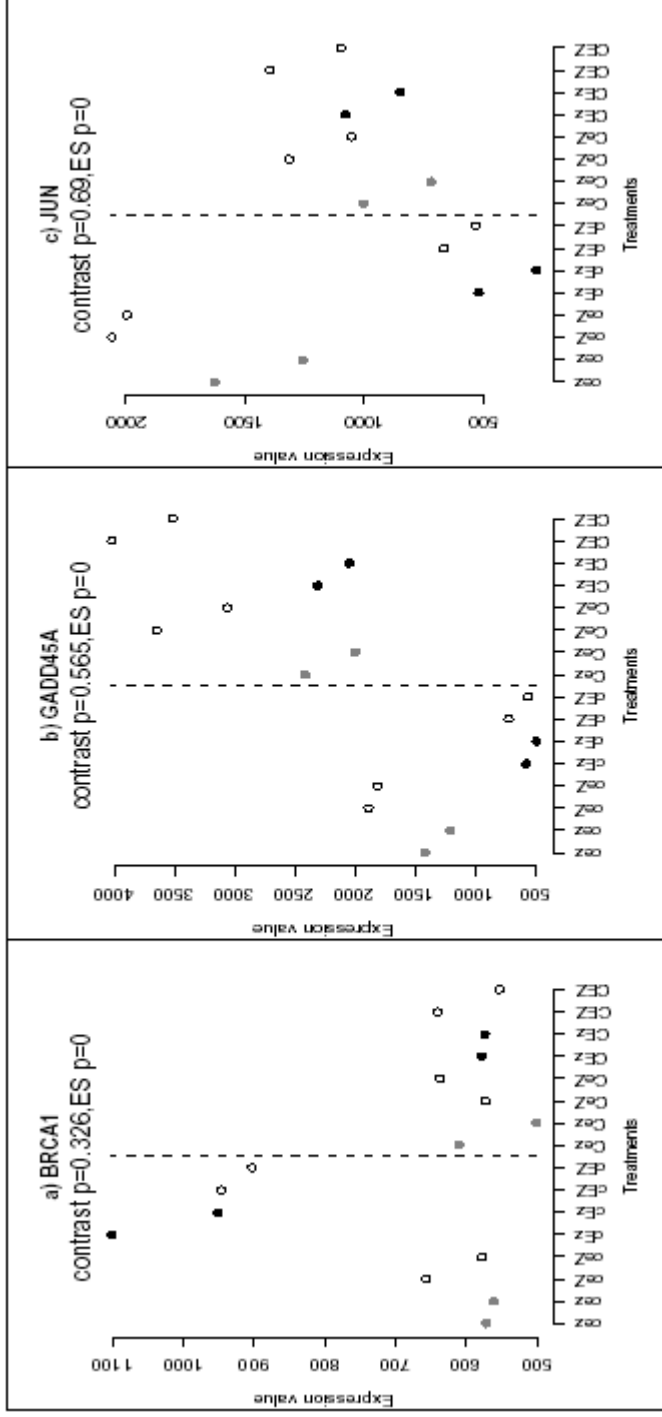
Primary Target Examples



$$\beta_{ES} \neq 0 \text{ or } \beta_{CX:ES} \neq 0$$

$$\beta_{ES} + \beta_{CX:ES} \neq 0 \text{ (i.e. } CEZ \neq CEZ)$$

Secondary Target Examples



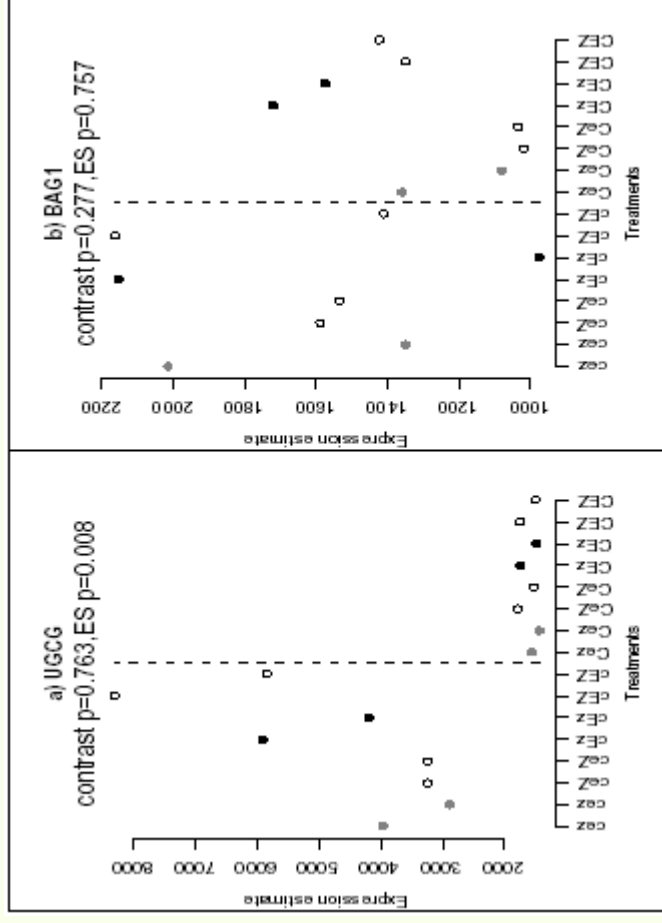
$$\beta_{ES} \neq 0 \text{ or } \beta_{CX:ES} \neq 0$$

$$\beta_{ES} + \beta_{CX:ES} = 0 \quad (\text{i.e. } \text{Cez} = \text{CEZ})$$

Comparison with Soulez and Parker (2001)

- Treated ZR75-1 breast cancer cells with CX and ES or CX and one of three antiestrogens (tamoxifen, raloxifen, faslodex).
- Selected primary estrogen targets by examining fold change ratios under antiestrogen/CX and ES/CX conditions.

Comparison with Soulez and Parker (2001)



UCGG and BAG1 were identified as primary targets by Soulez and Parker. In our analysis, UGGG is a secondary target (note $CEz \approx CEz$) and BAG1 is not a target at all (note the variability in expression levels).

Primary and Secondary ES Target Genes in the Literature that We also Identified

- **Primary Targets**
 - *TFF1*, trefoil factor 1, ps2
 - *CCND1*, cyclin D1, PRAD1
 - *CTSD*, cathepsin D
 - *CYP1B1*, cytochrome P450, subfamily 1
 - *ERBB2*, HER2/NEU
 - *GREB1*, KIAA0575
 - *TRIM16*, EBBP, tripartite motif containing 16
- **Secondary Targets**
 - *BRCA1*
 - *GADD45A*, growth arrest and DNA-damage-inducible alpha
 - *BARD1*, *BRCA1* associated RING domain 1
 - *RBBP8*, retinoblastoma binding protein 8, CtIP
 - *JUN* (AP1 site)
 - *FOS* (AP1 site)
 - *DDIT3* (AP1 site)

Conclusions

- For gene selection using data from factorial designed microarray studies, linear models offer natural paradigm for analysis so long as careful consideration is given to the interpretation of the model parameters.
- The use of CX in this experiment is one example of a treatment that allows for the identification of primary and secondary ES targets.

Conclusions

- For experiments with more treatments of interest, fractional factorial designs may be applicable.
- The candidate genes that are selected using linear models would serve as good candidates for network reconstruction algorithms, e.g. Wagner (2001).

Other Thoughts

- While the tests of contrasts seem somewhat obvious now, it was challenging to translate the biologists' request to "identify ES targets" into linear model parameters.
- Targets we identify can help elucidate meaningful joint regulation, question will be which joint model is appropriate for co-regulated genes.