Estrogen 2×2 Factorial Design

Denise Scholtens, Robert Gentleman

Experimental Data

In this vignette, we will demonstrate how to use linear models and the package *factDesign* to analyze data from a factorial designed microarray experiment. In this package, an `exprSet` called `estrogen` contains gene expression levels for 500 genes from Affymetrix HGU95av2 chips for eight samples from a breast cancer cell line. The expression estimates were calculated using the PM-only model in dChip after normalization by the Invariant Set Method (Li and Wong, 2001).

```
> library(Biobase)

Welcome to Bioconductor
         Vignettes contain introductory material.  To view,
         simply type: openVignette()
         For details on reading vignettes, see
         the openVignette help page.

> library(annotate)
> library(affy)
> library(mva)
> library(factDesign)
```

The investigators in this experiment were interested in the effect of estrogen on the genes in ER+ breast cancer cells over time. After serum starvation of all eight samples, they exposed four samples to estrogen, and then measured mRNA transcript abundance after 10 hours for two samples and 48 hours for the other two. They left the remaining four samples untreated, and measured mRNA transcript abundance at 10 hours for two samples, and 48 hours for the other two. Since there are two factors in this experiment (*estrogen* and *time*), each at two levels (*present* or *absent*,*10 hours* or *48 hours*), this experiment is said to have a 2×2 factorial design. Table 1 shows the correspondence of the sample names in `estrogen` with the experimental conditions.

```
> data(estrogen)
> estrogen

Expression Set (exprSet) with
        500 genes
```

Table 1: Experimental Conditions for `.cel` Files

time	estrogen	
	absent	present
10 hours	et1	Et1
	et2	Et2
48 hours	eT1	ET1
	eT2	ET2

```
        8 samples
                 phenoData object with 2 variables and 8 cases
         varLabels
                ES: absence/presence of estrogen
                TIME: 10/48 hours
```

```
> pData(estrogen)
```

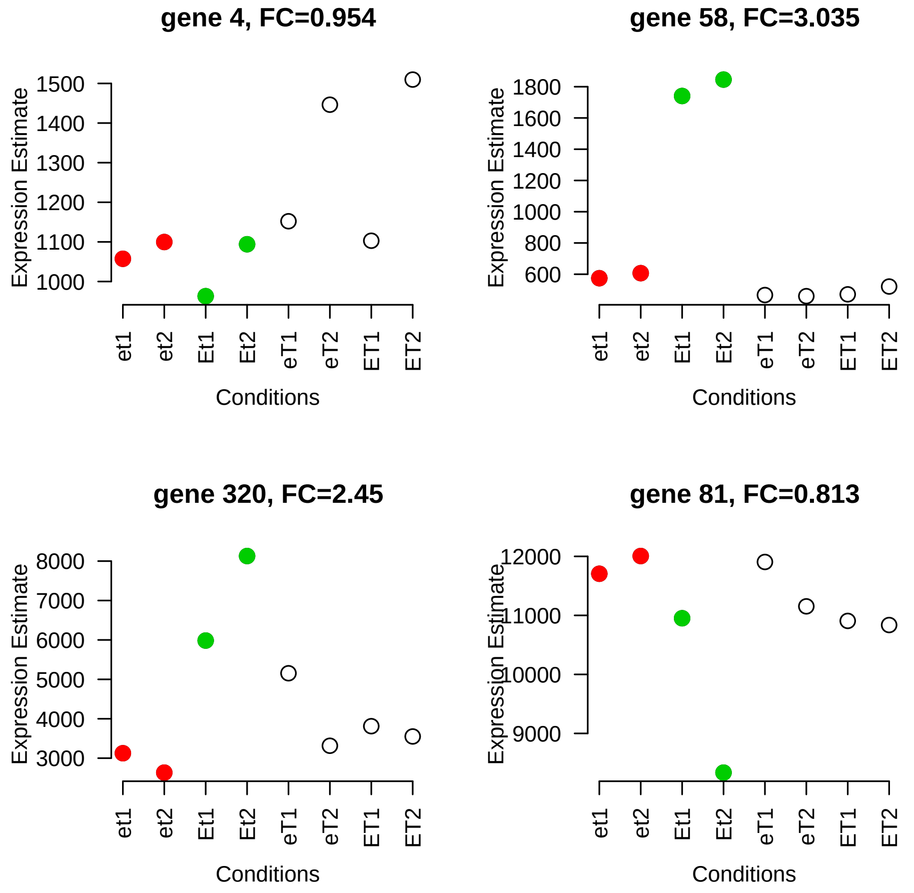
```
    ES TIME
et1  A  10h
et2  A  10h
Et1  P  10h
Et2  P  10h
eT1  A  48h
eT2  A  48h
ET1  P  48h
ET2  P  48h
```

Analysis Using Fold Change Criteria

A simple method for finding estrogen-affected genes would be to form a ratio of the mean expression levels of the estrogen-treated samples to the mean of the expression levels for the untreated samples. Suppose we consider only the 10-hour time point, calculate fold change (FC) values for the estrogen-treated vs. untreated samples, and select genes for which we observe FC>2. In the plots below, absence/presence of estrogen is represented by `e/E` and on the horizontal axis. The proposed FC criteria would compare the mean of the green dots to the mean of the red dots.

If we used a FC> 2 criteria to identify ES-affected genes in the `estrogen` data set, we would successfully eliminate genes like gene 4 and select genes like gene 58; however, we might include many false positive results. Gene 320 has a FC value greater than 2, but the variability of the expression estimates causes some concern. Note that for gene 81, the FC value is quite low due to the presence of a single outlier.

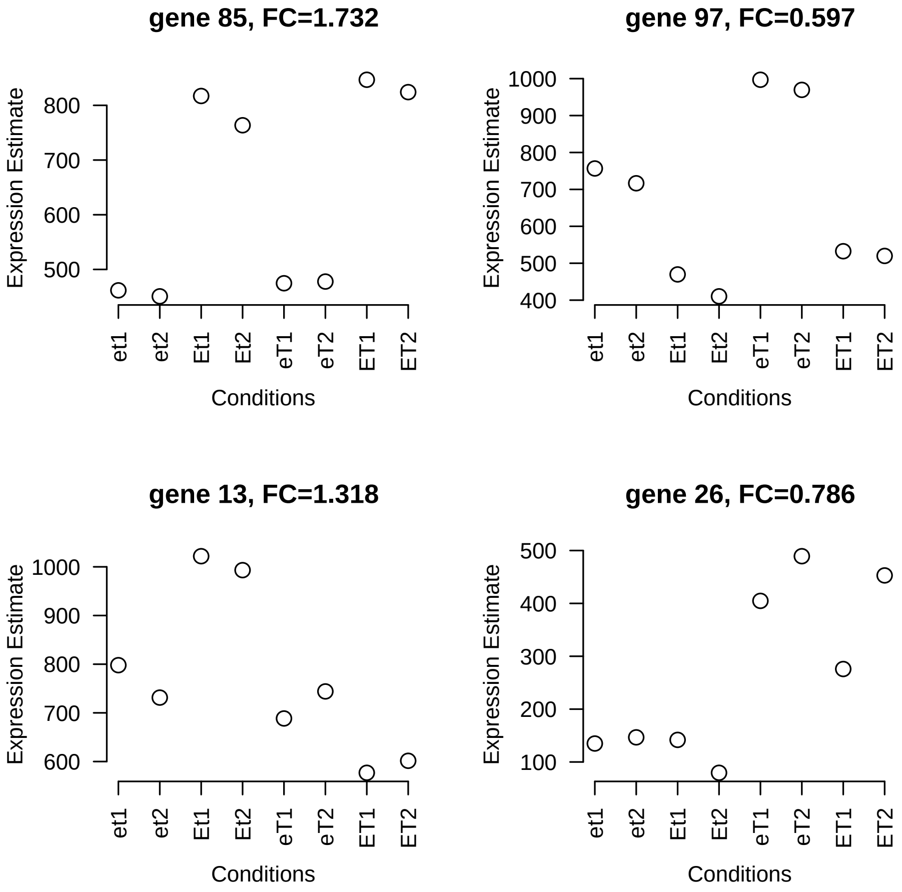
```
> par(mfrow = c(2, 2))
> par(las = 2)
> for (i in c(4, 58, 320, 81)) {
+     index <- i
+     expvals <- exprs(estrogen)[index, ]
+     plot(expvals, axes = F, cex = 1.5, xlab = "Conditions", ylab = "Expression Estimate")
+     points(1:2, expvals[1:2], pch = 16, cex = 1.5, col = 2)
+     points(3:4, expvals[3:4], pch = 16, cex = 1.5, col = 3)
+     axis(1, at = 1:8, labels = rownames(pData(estrogen)))
+     axis(2)
+     FC <- round(mean(expvals[3:4])/mean(expvals[1:2]), 3)
+     title(paste("gene ", index, ", FC=", FC, sep = ""))
+ }
```



We would like to find genes with consistent expression estimates between replicate samples that are either up- or down-regulated by estrogen, for example genes 85 and 97. We would also like to find genes like gene 13 for which the magnitude of the effect of estrogen changes over time. Furthermore, we would like to exclude genes like gene 26 that demonstrate change primarily over time, and not necessarily due to estrogen.

Selecting genes according to fold change estimates alone does not take advantage of the measure of variability in gene expression offered by the replicate sampels. Furthermore, we cannot attach statistical significance (i.e., a p-value) to the fold change estimates computed in this manner. It is also difficult to quantify any change in estrogen effect over time. Classical statistical linear modeling with thoughtful biological interpretation of the parameters offers a natural paradigm for the analysis of factorial designed microarray experiments.

```
> par(mfrow = c(2, 2))
> par(las = 2)
> for (i in c(85, 97, 13, 26)) {
+     index <- i
+     expvals <- exprs(estrogen)[index, ]
+     plot(expvals, axes = F, cex = 1.5, xlab = "Conditions", ylab = "Expression Estimate")
+     axis(1, at = 1:8, labels = rownames(pData(estrogen)))
+     axis(2)
+     FC <- round(mean(expvals[3:4])/mean(expvals[1:2]), 3)
+     title(paste("gene ", i, ", FC=", FC, sep = ""))
+ }
```



gene 85, FC=1.732

gene 97, FC=0.597

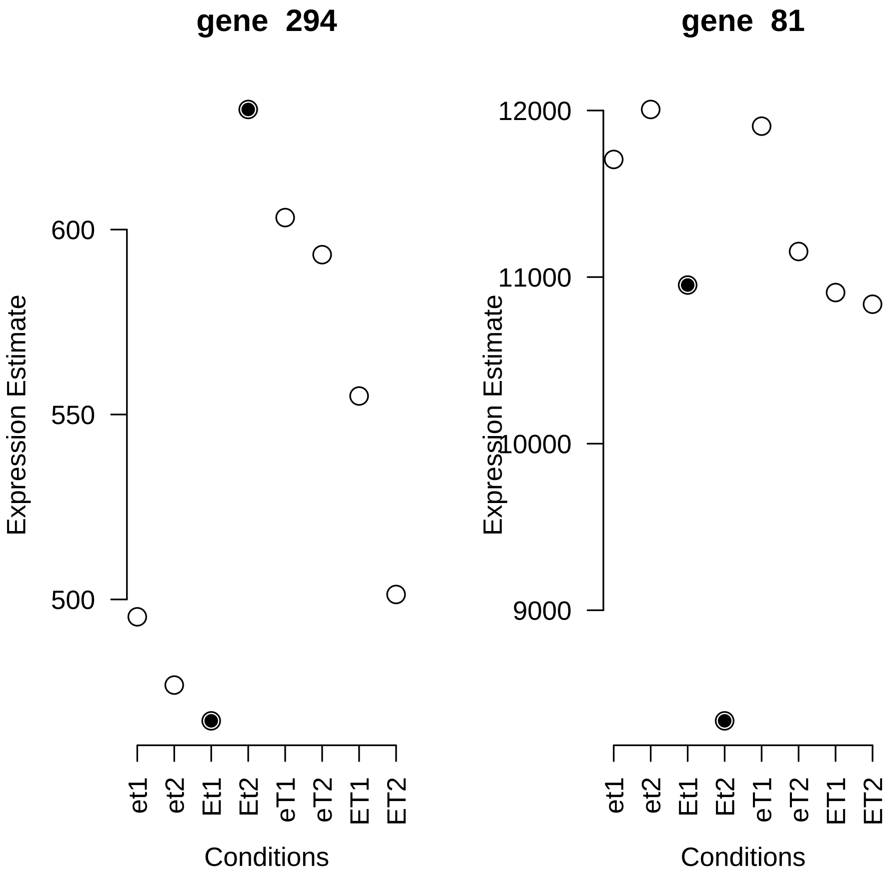
gene 13, FC=1.318

gene 26, FC=0.786

Removing Outliers

Before defining the linear model for this particular experiment, we want to remove observations that might be single outliers in the data set. The test we used is based on the differences between replicates and is appropriate for small factorial experimental designs. First, we identify replicate pairs with differences that are significantly larger than expected, and then we can apply a median absolute deviation filter to make sure one of the observations is indeed the single outlier. For example, gene 294 has a replicate pair with a large difference, but we wouldn't want to label either observation as the single outlier. Gene 81 has one observation that indeed appears to be a single outlier.

```
> op1 <- outlierPair(exprs(estrogen)[294, ], INDEX = pData(estrogen),
+     p = 0.05)
> print(op1)

$test
[1] FALSE

$pval
[1] 0.0626216

$whichPair
[1] 3 4

> madOutPair(exprs(estrogen)[25, ], op1[[3]])

[1] "NA"

> op2 <- outlierPair(exprs(estrogen)[81, ], INDEX = pData(estrogen),
+     p = 0.05)
> print(op2)

$test
[1] TRUE

$pval
[1] 0.04561594

$whichPair
[1] 3 4

> madOutPair(exprs(estrogen)[81, ], op2[[3]])

[1] 4

> par(mfrow = c(1, 2))
> par(las = 2)
> for (i in c(294, 81)) {
+     index <- i
+     expvals <- exprs(estrogen)[index, ]
+     plot(expvals, axes = F, cex = 1.5, xlab = "Conditions", ylab = "Expression Estimate")
+     points(3:4, expvals[3:4], pch = 16)
+     axis(1, at = 1:8, labels = rownames(pData(estrogen)))
+     axis(2)
+     title(paste("gene ", i))
+ }
```



Describing the Linear Model

The 2×2 factorial design of this experiment allows us to use a statistical linear model to measure the effects of estrogen and time on gene expression. In equation (1), $y_{full,ij}$ is the observed expression level for gene i in sample j ($j = 1, ..., 8$). $x_{ESj} = 1$ if estrogen is present and 0 otherwise; $x_{TIMEj} = 1$ if gene expression was measured at 48 hours and 0 otherwise. μ_i is the expression level of untreated gene i at 10 hours. β_{ESi} and β_{TIMEi} represent the effects of estrogen and time on the expression level of gene i, respectively. $\beta_{ES:TIMEi}$ is called an interaction term for gene i; this allows us to quantify any change in estrogen effect over time for probes like `1700_at`. ϵ_{ij} represents random error for gene i and sample j, and is assumed to be independent for each gene and sample, and normally distributed with mean 0 and variance σ_i^2. The biologically independent replicates of the experimental conditions in this study allow us to estimate σ_i^2.

$$y_{full,ij} = \mu_i + \beta_{ESi} x_{ESj} + \beta_{TIMEi} x_{TIMEj} + \beta_{ES:TIMEi} x_{ESj} x_{TIMEj} + \epsilon_{ij} \tag{1}$$

To proceed with the analysis, we estimate the β parameters for every gene using least squares, and call the estimates $\hat{\beta}_{ESi}$, $\hat{\beta}_{TIMEi}$, and $\hat{\beta}_{ES:TIMEi}$. For gene i, the samples that were not treated with estrogen and were measured at 10 hours will have estimated expression values of $\hat{\mu}_i$. The estrogen-treated, 10-hour samples will have estimates $\hat{\mu}_i + \hat{\beta}_{ESi}$. The untreated, 48-hour samples will have estimates $\hat{\mu}_i + \hat{\beta}_{TIMEi}$. The estrogen-treated, 48-hour samples will have estimates $\hat{\mu}_i + \hat{\beta}_{ESi} + \hat{\beta}_{TIMEi} + \hat{\beta}_{ES:TIMEi}$.

We will also form a reduced model with only an effect for time (2), and use this to decide if a model including estrogen is appropriate for the gene of interest.

$$y_{time,ij} = \mu_i + \beta_{TIMEi} x_{TIMEj} + \epsilon_i \tag{2}$$

```
> lm.full <- function(y) lm(y ~ ES + TIME + ES * TIME)
> lm.time <- function(y) lm(y ~ TIME)
> lm.f <- esApply(estrogen, 1, lm.full)
> lm.t <- esApply(estrogen, 1, lm.time)
> summary(lm.f[[1]])

Call:
lm(formula = y ~ ES + TIME + ES * TIME)

Residuals:
      1       2       3       4       5       6       7       8
 -3.246   3.246  41.774 -41.774  13.413 -13.413  17.822 -17.822

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   790.83      23.73  33.321 4.84e-06 ***
ESP           -38.03      33.56  -1.133   0.3205
TIME48h       -74.22      33.56  -2.211   0.0915 .
ESP:TIME48h    99.76      47.47   2.102   0.1035
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.56 on 4 degrees of freedom
Multiple R-Squared: 0.5884,         Adjusted R-squared: 0.2797
F-statistic: 1.906 on 3 and 4 DF,  p-value: 0.27

> summary(lm.t[[1]])
```

```
Call:
lm(formula = y ~ TIME)

Residuals:
   Min     1Q Median     3Q    Max
-60.79 -24.16  14.41  22.39  48.69

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   771.81      20.17  38.268 2.13e-08 ***
TIME48h       -24.34      28.52  -0.853    0.426
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.34 on 6 degrees of freedom
Multiple R-Squared: 0.1082,       Adjusted R-squared: -0.04039
F-statistic: 0.7282 on 1 and 6 DF,  p-value: 0.4262
```

Selecting Genes of Interest using the Linear Model

We are only interested in genes which are affected by estrogen. One way to select such genes is to compare the full linear model (`lm.f`) to the linear model consisting of only a term for time (`lm.t`) using an ANOVA F-test. If the full model `lm.f` fits better than the reduced model `lm.t`, then we know the gene must be affected by estrogen.

```
> Fpvals <- rep(0, length(lm.f))
> for (i in 1:length(lm.f)) {
+     Fpvals[i] <- anova(lm.t[[i]], lm.f[[i]])$P[2]
+ }
```

Since we have so many genes to consider, multiple comparisons is an obvious problem. We can adjust the p-values resulting from the ANOVA F-test using a the False Discovery Method for dependent hypothesis tests of Benjamini and Yekutieli (2001). If we select genes that have a p-value < 0.10 after the adjustment, then we know the genes are significantly affected by estrogen with a false positive rate of 0.10.

```
> library(multtest)
> F.res <- mt.rawp2adjp(Fpvals, "BY")
> F.adjps <- F.res$adjp[order(F.res$index), ]
> numgenes.Fsub <- sum(F.adjps[, 2] < 0.1)
> Fsub <- which(F.adjps[, 2] < 0.1)
> estrogen.Fsub <- estrogen[Fsub]
> lm.f.Fsub <- lm.f[Fsub]
> estrogen.Fsub
```

```
Expression Set (exprSet) with
        31 genes
        8 samples
                 phenoData object with 2 variables and 8 cases
         varLabels
                ES: absence/presence of estrogen
                TIME: 10/48 hours
```

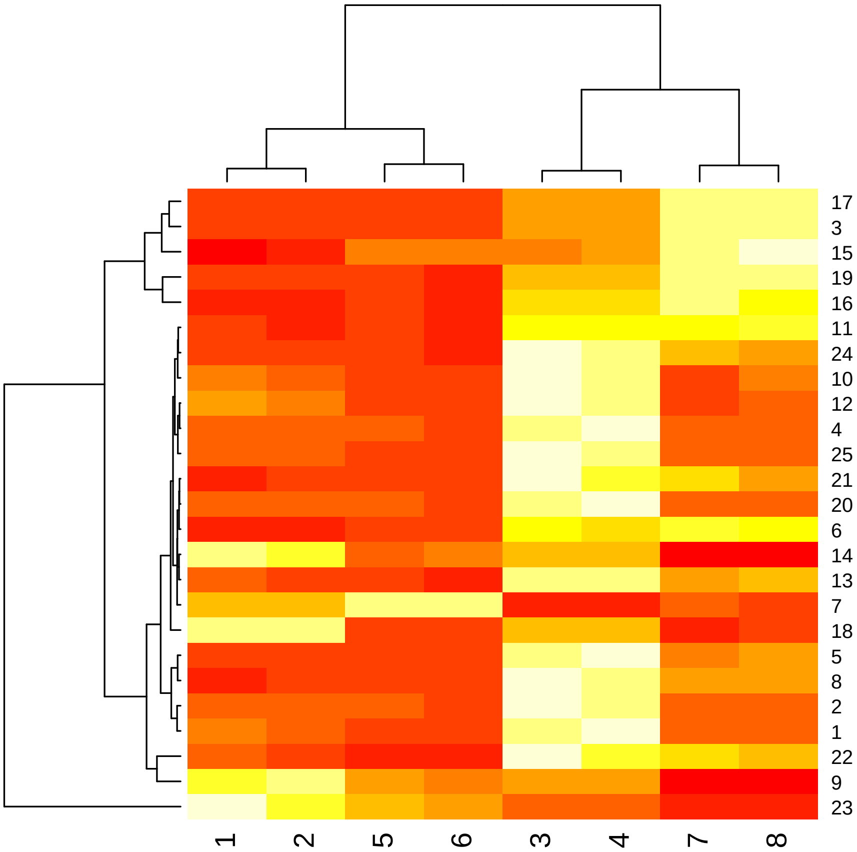
Suppose we want to identify genes that are affected by estrogen at 10 hours. In our linear model, this corresponds to testing a null hypothesis $H_{0ES} : \beta_{ES} = 0$, and if the hypothesis rejected, concluding that the gene has a main estrogen effect.

```
> betaNames <- names(lm.f[[1]][["coefficients"]])
> lambda <- par2lambda(betaNames, c("ESP"), c(1))
> mainES <- function(x) contrastTest(x, lambda)[[1]]
> mainESgenes <- sapply(lm.f.Fsub, FUN = mainES)
> sum(mainESgenes == "REJECT")
```

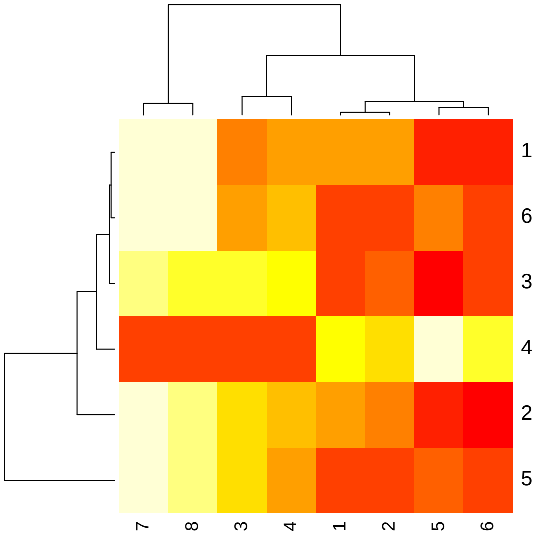
```
[1] 25
```

Heatmaps can be a useful way to visualize genes that are selected according to a certain criteria. In the first heatmap that follows, we see genes for which the null hypothesis H_{0ES} was rejected at a 0.01 significance level. In the second heatmap, we see the genes for which the main estrogen effect was not statistically significant; it appears that estrogen affected these genes only after 48 hours.

```
> heatmap(exprs(estrogen.Fsub)[mainESgenes == "REJECT", ], Colv = 1:8)
```



```
> heatmap(exprs(estrogen.Fsub)[mainESgenes == "FAIL TO REJECT",
+     ], Colv = 1:8)
```



Selecting genes according to p-value can produce some possibly misleading results. For example, the 17th gene with a main ES effect had a p-value for β_{ES} less than 0.01, but the estimate of fold change at 10 hours is only 1.42. While this small effect is statistically significant, it may not be biologically interesting.

```
> lambdaNum <- par2lambda(betaNames, list(c("(Intercept)", "ESP")),
+     list(c(1, 1)))
> lambdaDenom <- par2lambda(betaNames, list(c("(Intercept)")),
+     list(c(1)))
> FCval <- findFC(lm.f.Fsub[[17]], lambdaNum, lambdaDenom)
> print(FCval)
```

```
         [,1]
[1,] 1.422268
```

Now suppose we want to find genes that are affected by estrogen after 48 hours. We want to compare the gene expression levels of the untreated samples that were measured at 48 hours with the estrogen-treated samples at 48 hours. In terms of our linear model, for each gene, we want to test the null hypothesis $H_{0ES,TIME}$ in (3).

$$H_{0ES,TIME} : \mu + \beta_{TIME} = \mu + \beta_{ES} + \beta_{TIME} + \beta_{ES:TIME} \tag{3}$$

Testing the null hypothesis $H_{0ES,TIME}$ is equivalent to testing the linear contrast $H_{0ES,TIME*}$ in (4).

$$H_{0ES,TIME*} : \beta_{ES} + \beta_{ES:TIME} = 0 \tag{4}$$

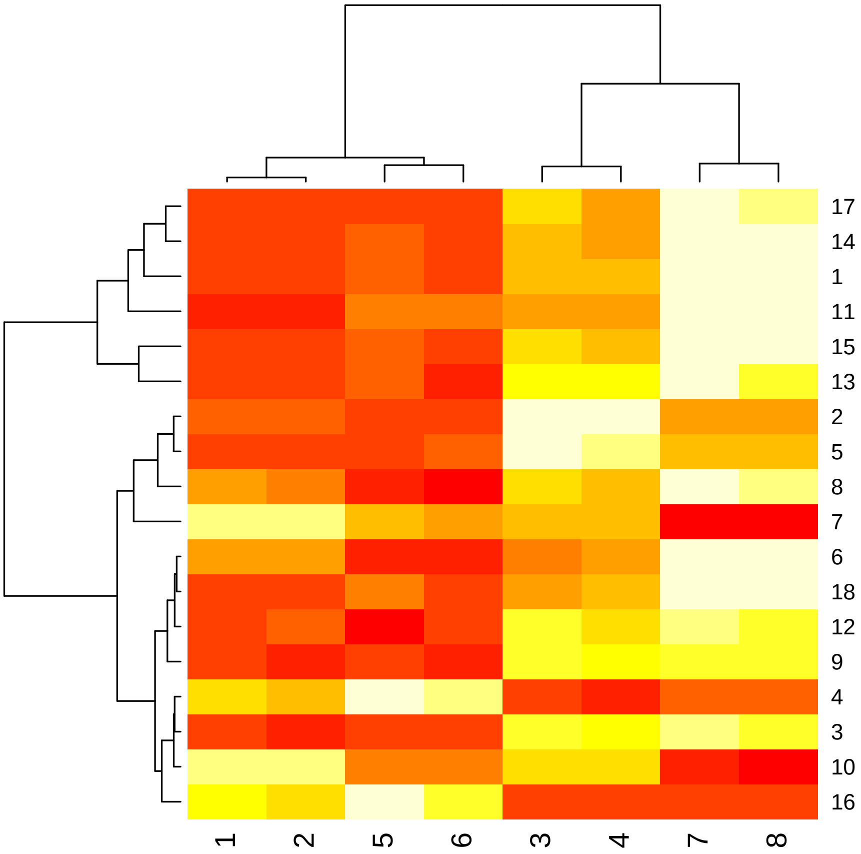
The technique for testing this linear contrast follows from straightforward linear model theory.

```
> lambdaEST <- par2lambda(betaNames, list(c("ESP", "ESP:TIME48h")),
+     list(c(1, 1)))
```

```
> ESTcontrast <- function(x) contrastTest(x, lambdaEST)[[1]]
> ESTgenes <- sapply(lm.f.Fsub, FUN = ESTcontrast)
> sum(ESTgenes == "REJECT")

[1] 18
```

Again, we can use a heatmap to look at genes for which we rejected $H_{ES,TIME*}$.

```
> heatmap(exprs(estrogen.Fsub)[ESTgenes == "REJECT", ], Colv = 1:8)
```



After genes are selected according to contrast tests of interest, the annotation information available in other Bioconductor packages allows for more in-depth research on specific genes.

Using linear models for factorial designed microarray experiments enables investigators to extend analyses beyond basic gene filtering according to fold change. Genes can be selected in a high-throughput manner with biologically interpretable parameters and quantifiable measures of confidence. This lab investigated the effects of estrogen on breast cancer cells, but the principles behind this specific example are applicable to any carefully designed microarray study.