1 Statistical Methods

1.1 Linear Model Approach to Compare Two Generations of Custom Microarray Chips for Salamanders

Let Y^g_{ijk} be the expression level for *gene* g on *chip generation* i ($i = 1$: Affy_001, $i = 2$: Affy_002), under *treatment* j ($j = 1$: group D18 after thyroid treatment, $j = 2$: group D0 after thyroid treatment), from *subject* (salamander) $k(j)$ ($k = 1, \ldots, 3$). The notation $k(j)$ indicates that subjects are nested within treatments.

A gene-wise linear statistical model for the effect of *chip generation* and *treatment* on gene expression, taking into account the *subject* (salamander) effect can be formulated as follows.

$$Y^g_{ijk} = \mu^g + \alpha^g_i + \beta^g_j + (\alpha\beta)^g_{ij} + D^g_{k(j)} + \varepsilon^g_{ijk} \tag{1.1}$$

Here, α^g_i and β^g_j denote the fixed effects due to *chip generation* and *treatment*, respectively, whereas $(\alpha\beta)^g_{ij}$ is the fixed two-way interaction between these effects. All fixed main effects and interactions are assumed to satisfy the usual identifiability constraints. The *subject* effect $D^g_{k(j)}$ due to variation between individual salamanders is assumed to follow a normal distribution with variance $\sigma^2_{D^g}$, that is, $D^g_{k(j)} i.i.d. \sim N(0, \sigma^2_{D^g})$. Finally, the random measurement error is denoted by $\varepsilon^g_{ijk} i.i.d. \sim N(0, \sigma^2_{\varepsilon^g})$, and the random terms $D^g_{k(j)}$ and ε^g_{ijk} are assumed to be independent of each other. The superscript g indicates that models are fit separately for each gene.

The experimental design can be described as a two factor study with repeated measures, or as a split plot design where the experimental units (subjects) are the salamanders. See the data schema in Table 1.

Table 1: Data Schema for Comparison of Two Chip Generations

		Chip Generation		
Treatment	Subject	old	new	difference
	$k = 1$	Y^g_{111}	Y^g_{211}	$Y^g_{211} - Y^g_{111} = Z^g_{11}$
$j = 1$	$k = 2$	Y^g_{112}	Y^g_{212}	$Y^g_{212} - Y^g_{112} = Z^g_{12}$
	$k = 3$	Y^g_{113}	Y^g_{213}	$Y^g_{213} - Y^g_{113} = Z^g_{13}$
	$k = 4$	Y^g_{121}	Y^g_{221}	$Y^g_{221} - Y^g_{121} = Z^g_{21}$
$j = 2$	$k = 5$	Y^g_{122}	Y^g_{222}	$Y^g_{222} - Y^g_{122} = Z^g_{22}$
	$k = 6$	Y^g_{123}	Y^g_{223}	$Y^g_{223} - Y^g_{123} = Z^g_{23}$

The hypothesis that *treatment* differences do not depend on *chip generation* can be formulated as no interaction between the factors *chip generation* and *treatment*, $H_0 : (\alpha\beta)^g_{ij} \equiv 0$. It can be tested using the corresponding analysis of variance F test, or equivalently using the differences $Z^g_{jk} = Y^g_{2jk} - Y^g_{1jk}$.

Using differences to test for interaction.

The expected value of the differences Z^g_{jk} is $\alpha^g_2 - \alpha^g_1 + (\alpha\beta)^g_{2j} - (\alpha\beta)^g_{1j}$, and the expected value of the differences of means $\bar{Z}^g_{1\cdot} - \bar{Z}^g_{2\cdot}$ is $(\alpha\beta)^g_{21} - (\alpha\beta)^g_{11} - (\alpha\beta)^g_{22} + (\alpha\beta)^g_{12}$, which equals zero under the null hypothesis of no interaction effect. Under model (1.1), the variance of the difference of means is $4/3\sigma^2_{\varepsilon^g}$, which can be consistently estimated by $(s^2_{Z^g_{1k}} + s^2_{Z^g_{2k}})/3$, where $s^2_{Z^g_{jk}}$ is the sample variance of the differences Z^g_{jk} in group j

. We can thus test $H_0 : \bar{Z}^g_{1\cdot} - \bar{Z}^g_{2\cdot} = 0$ in a two-sample t-test assuming equal variance in both treatment groups. Alternatively, we can relax the assumption of equal error variances and estimate adjusted degrees of freedom df using the Welch-Satterthwaite approximation

$$df = \frac{[(s^2_{Z^g_{1k}} + s^2_{Z^g_{2k}})/3]^2}{[(s^2_{Z^g_{1k}}/3)^2 + (s^2_{Z^g_{2k}}/3)^2]/2}. \tag{1.2}$$

Using analysis of variance F tests for interaction

The analysis of variance table for the design above is (see, e.g., Kutner, Nachtsheim, Neter, Li, "Applied Linear Statistical Models, 5e", McGraw-Hill, New York; Ch. 27) given in Table 2.

Table 2: Analysis of Variance Table for Two-Factor Experiment with Repeated Measures

Factor	Sums of Squares	df	Error Term
Treatment	$SST = 6\sum_j(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots})^2$	1	Subjects
Subjects	$SSS = 2\sum_j\sum_k(\bar{Y}_{\cdot jk} - \bar{Y}_{\cdot j\cdot})^2$	4	
Chip Generation	$SSG = 6\sum_i(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})^2$	1	Error
Interaction	$SSTG = 3\sum_i\sum_j(\bar{Y}_{ij\cdot} - \bar{Y}_{\cdot j\cdot} - \bar{Y}_{i\cdot\cdot} + \bar{Y}_{\cdots})^2$	1	Error
Error	$SSE = \sum_i\sum_j\sum_k(Y_{ijk} - \bar{Y}_{ij\cdot} - \bar{Y}_{\cdot jk} + \bar{Y}_{\cdot j\cdot})^2$	4	
Total	$SSTo = \sum_i\sum_j\sum_k(Y_{ijk} - \bar{Y}_{\cdots})^2$	11	

According to the table, the test for interaction is performed by calculating F_{obs} as the ratio of $SSTG$ and $SSE/4$ and comparing it to an F-distribution with numerator and denominator degrees of freedom of 1 and 4, respectively.

1.2 Estimating the proportion of false null hypotheses

Meinshausen and Rice (2006) suggest a procedure that estimates a lower bound for the proportion of false hypotheses in a large set of independent hypothesis tests. Letting n denote the number of genes, and $F(t)$ the empirical distribution of p-values, the estimated proportion λ of false hypotheses is

$$\hat{\lambda} = \sup_{t\in(0,1)} \frac{F(t) - t - \sqrt{\frac{1}{2n}\log 2/\alpha}}{1-t}. \tag{1.3}$$

If p-values are independent and uniformly distributed when their corresponding null hypotheses are true, then the estimator is conservative (i.e., $\hat{\lambda} < \lambda$) with probability $1 - \alpha$.

1.3 Clustering of time series expression profiles

The expression data from Amby_002 were examined to identify genes that were differentially expressed between the three times points (D0, D8 and D18) using the statistical software SAS 9.2 (SAS Institute Inc, Cary, NC, USA). Standard ANOVA F-tests were performed on genes for which the equal variance assumption was

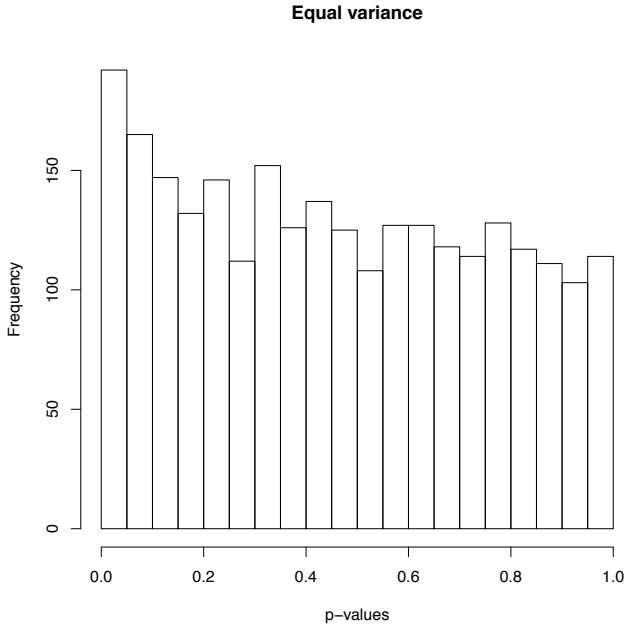

Figure 1: Histogram of p-values of t-tests for no treatment-by-generation interaction assuming equal variance

not rejected by Bartlett's test at $\alpha = .01$. Otherwise, differences between time points were tested using the variance-weighted one-way ANOVA according to (Welch 1951). All following analyses were conducted using the software R (version 2.9.2, 2009 The R Foundation for Statistical Computing). For genes that were identified as differentially expressed, *post-hoc* unequal-variance t-tests with degrees of freedom calculated according to the Welch-Satterthwaite approximation were performed. F-tests and subsequent *post-hoc* comparisons were conducted at $\alpha = 0.1, 0.05$ and 0.01, final results are presented for $\alpha = 0.05$. Based on the direction of significant gene expression changes between time points, DEGs were classified into 9 expression profiles. For example, an expression change between two consecutive time points can be described as up, down, or unchanged (i.e., no significant difference). Thus, each gene was assigned to one of the 3 x 3 = 9 expression profiles (e.g. up up; up down, up unchanged, etc).

2 Statistical Results

2.1 Using differences to test for interaction.

The distribution of p-values across 2601 genes for the two-sample test is shown in Figures 1 (assuming equal variance) and 2 (allowing unequal error variances). In the first case, 192 genes out of 2601 have a significant interaction at the α-level of 0.05. Using the Welch-Satterthwaite estimation of degrees of freedom, 130 genes have a significant interaction. For the equal-variance t-tests, the Meinshausen-Rice procedure estimates 139 (5.3%) false hypotheses (i.e., true interactions) while it estimates 74 (2.9%) false hypotheses for the t-tests modeling unequal variances.

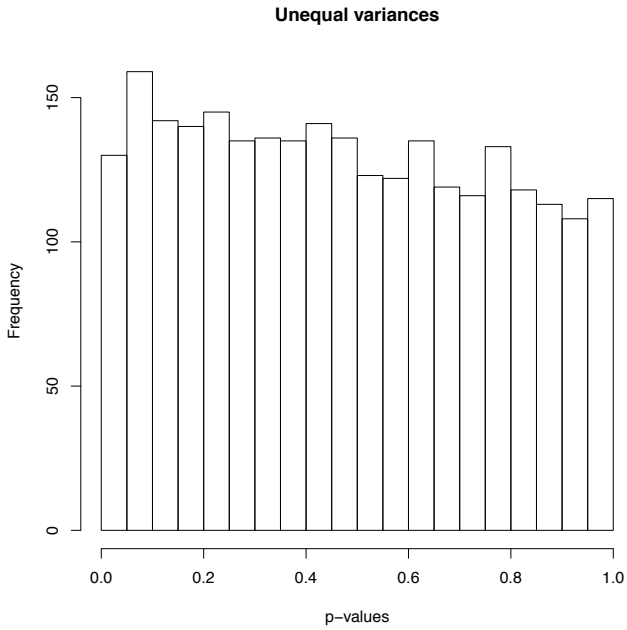

Figure 2: Histogram of p-values of t-tests for no treatment-by-generation interaction allowing for unequal variances