This is to certify that the
thesis entitled

Differential Expression Analysis of DNA Microarray data with Application to the Heat Shock Response of Arabidopsis thaliana

presented by

William R. Swindell

has been accepted towards fulfillment
of the requirements for the

M.S. degree in Statistics and Probability

Marianne Huebner
Major Professor's Signature

4-25-2007
Date

DIFFERENTIAL EXPRESSION ANALYSIS OF DNA MICROARRAY DATA WITH APPLICATION TO THE HEAT SHOCK RESPONSE OF ARABIDOPSIS THALIANA

By

William R. Swindell

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Statistics and Probability

2007

ABSTRACT

DIFFERENTIAL EXPRESSION ANALYSIS OF DNA MICROARRAY DATA WITH APPLICATION TO THE HEAT SHOCK RESPONSE OF ARABIDOPSIS THALIANA

By

William R. Swindell

DNA microarrays are widely used research tools that allow the expression level of thousands of genes to be monitored simultaneously. A common interest arising in the context of microarray data is to determine whether a gene's expression level differs between two conditions. Differential expression anlaysis provides a means for addressing such interests, but several aspects of microarray data complicate the application of standard two-sample methods. The Limma software package utilizes a Bayesian linear model approach to differential expression analysis, which has gained considerable popularity within the research community. In Chapter 1, the primary statistical challenges associated with differential expression analysis are reviewed, and the Bayesian linear model approach to these challenges is outlined. It is concluded that the statistical methods implemented in Limma are useful in the absence of ideal procedures, but that attention should be paid to several key assumptions that may not be satisfied by the data. Chapter 2 presents a detailed application of the Limma procedure (and other methods) to microarray data generated from experiments performed with *Arabidopsis thaliana*. In particular, the heat shock transcription factor and protein network of *Arabidopsis* is profiled under a wide range of abiotic and biotic stress treatments in multiple cell types. The analysis characterizes the interaction between *Arabidopsis* heat shock genes with heat and other types of stress, and identifies several heat shock gene expression patterns that have not been previously described.

ACKNOWLEDGEMENTS

I would like to thank Prof. Marianne Huebner for having provided me with encouragement throughout my graduate studies in Statistics, as well as clear guidance in developing new skills in the analysis of DNA microarray datasets. My research experiences working with Marianne were an important contribution to my graduate studies at Michigan State University, and will be of considerable value in subsequent stages of my career. I would also like to thank Prof. Andreas Weber (Plant Biology) for the considerable amount of time he spent working with Marianne and I on several projects. Andreas was of great assistance in having identified database resources critical to the completion of this thesis research, and was always welcoming of questions that arose during the course of our analyses. I thank Prof. Vince Melfi for taking the time to serve on my thesis committee and for having provided comments on this manuscript.

I am especially grateful to the Michigan State University Department of Probability and Statistics for having provided me with financial support in the form of teaching assistantships. In addition, I acknowledge the Michigan State University Quantitative Biology and Modeling Initiative (QBMI) for having provided research assistantships that have enabled the completion of several projects.

The gene expression data analyzed in chapter two of this thesis was generated by the AtGenExpress consortium and the *Arabidopsis* Functional Genomics Network (AFGN). The public availability of this data was critical to the work presented in this thesis and provides a valuable resource for the *Arabidopsis* research community.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures in this thesis are presented in color.

CHAPTER 1

DIFFERENTIAL EXPRESSION ANALYSIS OF MICROARRAY DATA WITH LINEAR MODELS AND BAYESIAN STATISTICS

Summary

DNA microarrays are increasingly popular research tools that allow the expression level of thousands of genes to be monitored simultaneously. Statistical tests of differential expression are widely used in conjunction with DNA microarray data, but there are several approaches to identifying differentially expressed genes. The goal in differential expression analysis is to determine whether sufficient evidence exists to declare that a given gene's expression level differs between two conditions of measurement. The Limma software package is a recently developed tool that utilizes a Bayesian approach to differential expression analysis. The software, moreover, is implemented within a highly flexible linear model framework, which allows investigators to evaluate complicated microarray experiments with little difficultly. In this chapter, an overview of the challenges entailed by differential expression analysis is provided, and the statistical methods by which the Limma procedure addresses these challenges are presented. In particular, the basic *t*-test as applied to differential expression is developed within the context of linear models. The main ideas underlying the moderated *t*-statistic and posterior odds are then presented, with the aim of drawing attention to the key data assumptions upon which statistical inferences depend. The performance of the Limma procedure as evaluated by Smyth (2004) is then discussed in conjunction with remaining issues that await resolution. The main conclusion emerging from this review is that statistical inferences based upon the Limma procedure are dependent upon several assumptions that may not be satisfied by microarray datasets. In the absence of ideal methods, therefore, inferences generated by Limma should be made with appropriate caution. These considerations underscore the challenges entailed by the complexity of DNA microarray data.

Section 1 Introduction

DNA microarrays are a widely used tool for simultaneously monitoring the expression level of thousands of genes throughout a genome (Brown and Botstein 1999; Lander 1999). As the application of microarray technology has become increasingly widespread, the statistical analysis of microarray data has become a topic of considerable importance. Critical statistical issues arise in almost every phase of microarray processing, including the experimental design stage (Yang and Speed 2002), extraction of expression intensities from microarray images (Jain et al. 2002), data preprocessing and normalization (Quackenbush 2002), and most especially, the post-processing phase during which a multitude of analyses are possible depending on the biological questions under consideration (e.g., Eisen et al. 1998, Alter et al. 2000, Brown 2000). At the post-processing stage, standard differential expression analysis is perhaps the most basic procedure that is most widely implemented. The objective in differential expression analysis is to determine whether there is sufficient evidence to claim that expression levels differ between two samples (e.g., tumor vs. benign tissue) for individual genes. Although this basic problem is analogous to that commonly addressed by the two-sample t-test procedure (Welch 1947), several complicating factors arise in the context of microarray data, such that only rough agreement now exists regarding the best approach for identification of differentially expressed genes (Allison et al. 2006).

The biological and technical variability commonly associated with gene expression estimates is, at a fundamental level, a key factor that complicates differential expression analysis. If such variability was not present or trivial, simple analytical methods would prove robust when applied to microarray datasets. Unfortunately,

however, gene expression estimates are inherently variable, since mixtures of cell types and possibly genotypes are common RNA sources in experiments, and the technical processes of mRNA extraction, amplification, and probe hybridization are subject to various sources of noise and interference. In light of these challenges, a wide range of approaches have been suggested and implemented in the biological literature. The first and most straightforward approach is to avoid hypothesis testing altogether, and identify genes on the basis of their expression level fold-change between samples. There are, in fact, several advantages to this approach, including a focus on effect size, non-dependency on modeling assumptions, minimal estimation of parameters, and importantly, straightforward interpretation. Ultimately, however, this approach is problematic since it does not account for the sources of variation mentioned above, which can impact different genes to varying extents. The standard two sample t-test is perhaps the next simplest alternative to fold-change, since it does account for biological and technical variability. The ordinary t-statistic, however, varies inversely with the standard deviation of expression estimates, such that significant statistics can arise for effect sizes that are arbitrarily small. Furthermore, the standard t-test requires that the variance associated with gene expression levels (σ^2) be estimated for each gene based upon what is typically a small number of replicates. While the former difficulty can be addressed by adding a constant term to the denominator of the t-statistic (e.g., Efron et al. 2001; Tusher et al. 2001; Broberg 2003), alternative methods are necessary to avoid independent estimation of σ^2 for all genes represented on a microarray.

Bayes and empirical Bayes methods provide one possible way of circumventing the gene-by-gene estimation of σ^2. Bayesian methods assume a prior distribution that

characterizes how σ^2 (and possibly other parameters) varies across all genes considered in an analysis. The key advantage of this approach lies in its effective use of the replication that exists among all genes represented on a microarray, which is not otherwise utilized in the standard t-test procedure. The empirical Bayes approach, in particular, minimizes strict reliance on priors by estimating the parameters of prior distributions (i.e., the hyperparameters) from the microarray dataset under consideration. This Bayesian framework is the basis of the differential expression procedures implemented in the Limma software package (Smyth 2004), which is part of the R Bioconductor project and freely available online at http://www.bioconductor.org (Gentleman et al. 2004). The Limma package implements Bayesian approaches within a highly flexible linear model framework, which allows complex experiments to be analyzed with relative ease by researchers without extensive statistical background. This practical simplicity is a primary factor underlying the increasingly widespread application Limma to a wide range of research problems involving gene expression data (e.g., Boutros et al. 2004; Golden and Melov 2004; Renn et al. 2004; Rodriguez et al. 2004; Peart et al. 2005; Rensink et al. 2005).

In this chapter, the basic features of the linear model framework that underlie the methodology implemented in Limma are explained, and the empirical Bayesian model utilized in differential expression testing is outlined. Particular attention is given to the assumptions upon which inferences are based. In section 1, the basic t-test as applied to differential expression analysis is presented within the context of linear models. In section 2, the main ideas underlying the posterior odds and moderated *t*-statistic are presented. The final section provides a discussion of simulations and comparative

analyses in which Limma has been applied and provides an overview of persisting problems associated with differential expression analysis.

Section 2 Linear Models

Linear models represent an efficient way of combining data from multiple arrays into a single analysis, and provide considerable flexibility that allows for the analysis of microarray experiments of arbitrary complexity. In this section, the simple two-sample t-test will be developed in the framework of linear models. The analysis will consist of a gene-by-gene approach in which a linear model is fit to the expression values associated with each gene individually. The response variable under consideration will thus be represented as y_g, where the subscript denotes that the linear model is being fit to expression values associated with gene g. The meaning of y_g will differ depending upon whether the gene expression data has been generated from two-color or one-color (e.g., Affymetrix) microarray platforms. It will generally be assumed below that expression data has been generated from a two-color microarray platform, such that the response y_g represents the relative expression intensities of two RNA sources that have been labeled with red (R_g) or green (G_g) dyes. In particular, y_g will represent the log-ratio of the distinct expression intensities.

$$y_g = \log_2\left(\frac{R_g}{G_g}\right) = \log_2(R_g) - \log_2(G_g)$$

For the case of high density oligonucleotide platforms, the response variable y_g will be absolute log-intensities rather than relative log-ratios. This modification, however, does not alter the formulation to follow, which will focus on the two-color case. It will be assumed that each gene has been represented on a total of n arrays, such that n

replicate values of y_g are available. In vector format, these *n* replicate measurements are represented by the random vector $\mathbf{y}_g^T = (y_{g1},...,y_{gn})$. Replicates represent RNA samples that have been obtained from independent sources (biological replicates), rather than multiple samples that have been taken from the same individual (technical replicates). It is further assumed that the data have been normalized to prevent systematic variation from influencing the results of the differential expression analysis.

The first step towards building a linear model is construction of an appropriate design matrix (*X*) that serves to specify the RNA targets used on arrays. The design matrix used is generally non-unique and can vary according to the parameters included in the model. The systematic component of the model is represented by $X\boldsymbol{\alpha}_g$, where $\boldsymbol{\alpha}_g$ is a vector of coefficients representing a series of contrasts. The coefficient vector $\boldsymbol{\alpha}_g$ can be viewed as the *effects* of interest in the linear model analysis. In the familiar context of multiple regression, these coefficients are the β_i most often estimated by least-squares methods. The model partially explains the variance among the elements of $\mathbf{y}_g$ with $E(\mathbf{y}_g) = X\boldsymbol{\alpha}_g$. The $\boldsymbol{\alpha}_g$ can be estimated by least-squares, maximum likelihood, or robust regression.

Consider as an example a saturated direct design for two-color microarrays with three sources of RNA. This example was provided by Smyth (2004) and is illustrated in Figure 1-1. Given this design, one possible design matrix and coefficient vector are the following.

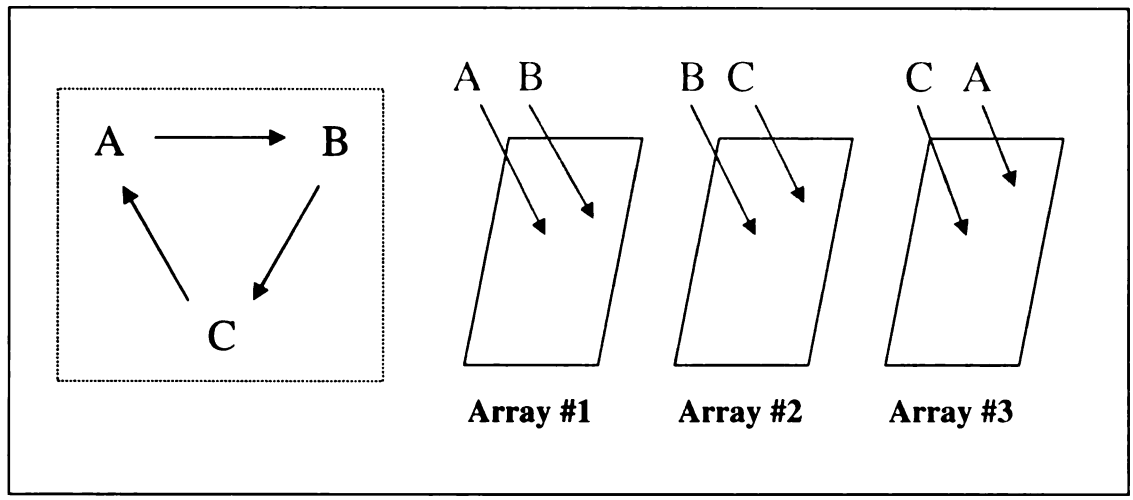

Figure 1-1. Saturated direct design for two-color microarray with three samples. Three RNA sources are considered (A, B, and C) using three microarrays (1, 2, and 3). The figure inset on the left provides a summary of the RNA-array hybridization scheme, in which the RNA source at the base of each arrow is labeled green, while the source at the tip of each arrow is labeled red (example from Smyth 2004).

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \qquad \boldsymbol{\alpha}_g = \begin{pmatrix} B-A \\ C-B \end{pmatrix}$$

The design matrix and coefficient vector given above can then be used to calculate the expected value of the random response vector $\mathbf{y}_g$. In accordance with the definitions of y_g presented above for two-color microarray platforms, this expectation represents relative log-ratios among the three sources of RNA.

$$E\begin{pmatrix} y_{g1} \\ y_{g2} \\ y_{g3} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} B-A \\ C-B \end{pmatrix} = \begin{pmatrix} B-A \\ C-A \\ A-C \end{pmatrix}$$

The contrast matrix (C) serves to specify which comparisons are of interest. In general, only some of a larger number of possible contrasts of the coefficients will be of biological importance for any one experiment. These selected contrasts are represented by $\boldsymbol{\beta}_g$, which is obtained by applying the transposed contrast matrix to the coefficient vector $\boldsymbol{\alpha}_g$.

$$\boldsymbol{\beta}_g = C^T \boldsymbol{\alpha}_g$$

Returning to the example presented in Figure 1-1, the simplest possible case is that in which only one specific contrast is of interest. Suppose, for instance, that a researcher wishes to compare the RNA abundance in sample C to that of sample A. This particular contrast is obtained by choosing the vector (1, 1) as the contrast matrix.

$$C = \begin{pmatrix} 1 & 1 \end{pmatrix}$$

Applying this contrast matrix (vector) to the coefficient vector then yields the desired comparison.

$$\boldsymbol{\beta}_g = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} B - A \\ C - B \end{pmatrix} = C - A$$

Typically, more than one contrast among RNA sources will be of interest. Such scenarios, however, are easily accommodated by an alternative specification of the

contrast matrix. For example, referring again to Figure 1-1, it may be of interest to evaluate all possible contrasts among the three sources of RNA. In this case, the following contrast matrix yields the appropriate set of contrasts $\boldsymbol{\beta}_g$.

$$C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

The contrast matrix yields the following comparisons of interest.

$$\boldsymbol{\beta}_g = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} B-A \\ C-B \end{pmatrix} = \begin{pmatrix} B-A \\ C-B \\ C-A \end{pmatrix}$$

The numerical values of $\boldsymbol{\beta}_g$ are simply the log-ratio fold changes that are of interest. In the above example, therefore, $\boldsymbol{\beta}_g$ is a vector that represents the $\log_2$ fold-change in expression between RNA sources B/A, C/B, and C/A. While more complex hypotheses are possible, the main interest is to determine whether the elements of $\boldsymbol{\beta}_g$ are significantly different from zero, i.e., whether expression values differ between RNA samples. For the *j*th entry of $\boldsymbol{\beta}_g$, therefore, it is of interest to test the null hypothesis $H_0 : \beta_{gj} = 0$ versus the alternative $H_A : \beta_{gj} \neq 0$.

The standard t-statistic makes two basic assumptions regarding the distribution of the β_{gj} and s_g^2 for any given gene. Additional assumptions will be required in the context of the moderated t-statistic (see section 3). Let $\hat{\boldsymbol{\alpha}}_g$ represent the coefficient

estimator of $\boldsymbol{\alpha}_g$, and suppose that $\hat{\boldsymbol{\alpha}}_g$ has covariance matrix $V_g s_g^2$. Since we wish to make inferences regarding the elements of $\boldsymbol{\beta}_g$, it is necessary to obtain the covariance matrix of $\hat{\boldsymbol{\beta}}_g$ (Smyth 2004).

$$\text{var}(\hat{\boldsymbol{\beta}}_g) = \text{var}\left(C^T \hat{\boldsymbol{\alpha}}_g\right) = C^T \text{var}\left(\hat{\boldsymbol{\alpha}}_g\right) C = C^T V_g C s_g^2$$

Hence, $\hat{\boldsymbol{\beta}}_g$ has covariance matrix $C^T V_g C s_g^2$. The elements of this matrix are a sub-sample of those from $V_g s_g^2$, and thus depend on $VAR(\hat{\boldsymbol{\alpha}})$ for individual effects that are of interest (as specified by the contrast matrix). The variance associated with the *j*th contrast of interest (i.e., β_{gj}) is therefore proportional to σ_g^2 times the *j*th diagonal element of $C^T V_g C$. Letting this latter quantity be represented by v_{gj}, the variance associated with the *j*th contrast of interest is $v_{gj}\sigma_g^2$. It is assumed that the estimated value of the *j*th contrast of interest approximates a normal distribution with mean β_{gj} and variance $v_{gj}\sigma_g^2$ (see Smyth 2004).

$$\hat{\beta}_{gj} | \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$$

If d_g represents the degrees of freedom associated with error term in the linear model, then $d_g s_g^2 / \sigma_g^2 \sim \chi_{d_g}^2$ from sampling theory (Rice 1995). Rearranging this distributional equality yields the sampling distribution of s_g^2.

$$s_g^2 \mid \sigma_g^2 \sim \left(\frac{\sigma_g^2}{d_g}\right)\chi_{dg}^2$$

The elements of $\hat{\boldsymbol{\beta}}_g$ therefore approximate a normal distribution, and the variance of each element follows a scaled chi-square distribution. The ratio of $\hat{\beta}_{gj}$ to $s_g\sqrt{v_{gj}}$ thus represents a normally distributed random variable divided by a chi-square random variable with d_g degrees of freedom, and is therefore a t-distribution with d_g degrees of freedom (Rice 1995).

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}}$$

The above represents the ordinary t-statistic and has been used in a number of studies in which the interest has been to determine whether differential expression has occurred between two RNA sources (e.g., tumor versus normal tissue). The two primary problems with this statistic are immediately upon inspection. First, if the value of $s_g\sqrt{v_{gj}}$ is small by chance for a particular gene, then very large values of t_{gj} may result even if the estimated fold-change ($\hat{\beta}_{gj}$) is relatively small. Such genes are unlikely to be of biological importance, but would nonetheless be identified as differentially expressed based on the above statistic. The second problem lies in the estimation of s_g^2 based upon what is a sample size equal to the number of arrays involved in the study, which would

most often not exceed six in a simple two sample case. Although additional assumptions regarding the data are required, these issues are addressed by the statistics implemented in the Limma linear modeling procedure.

Section 3 The moderated t and posterior odds statistics

The Limma linear modeling package evaluates hypotheses associated with contrasts of interest based upon two different test statistics. The moderated t-statistic ($\tilde{t}$) is an extension of the standard t-statistic introduced in section 1, while the posterior odds statistic (O) is based upon the work of Lönnstedt and Speed (2002), and provides an indication of the likelihood of differential expression with respect to a given contrast. The moderated t-statistic is generally used as a basis for obtaining p-values, while the posterior odds is primarily used for ranking genes according to evidence for differential expression. Gene rankings based upon the posterior odds are identical to those based upon the moderated t-statistic if the microarray data under consideration does not have missing values. Of the two statistics, the moderated t statistic may be of greater practical utility, since it is based upon fewer assumptions in comparison to the posterior odds. As discussed below, both statistics are based upon certain *hyperparameters* that must be estimated from the data. One particular hyperparameter upon which the posterior odds is based requires knowledge regarding the proportion of genes that are differentially expressed, which is unknown and difficult to estimate from the data. The posterior odds is therefore ultimately dependent upon an assumed value regarding the proportion of differentially expressed genes. The moderated t-statistic, however, is not dependent upon v_{0j}. Another advantage of the moderated t-statistic is that it is easily extendable to multiple testing of multiple contrasts using the F distribution. The central ideas leading to the moderated t-statistic and posterior odds statistics are discussed below, with particular attention to the assumptions that underlie their probabilistic interpretation. It should be mentioned that some notations and concepts are introduced below without immediately

stating a complete and precise definition. This has been done for clarity and allows for an initial focus on the main results and ideas, subsequently followed by details where appropriate.

The moderated t-statistic

The central difficulty associated with the standard t-statistic is the low number of observations available for individual genes. This low replication leads to poor estimates of the standard deviation associated with each gene (s_g). In microarray datasets, however, while replication associated with individual genes is low, there is ample replication with respect to the thousands of genes that are represented on an array. The advantage of the moderated t-statistic over the standard t-statistic lies in the manner in which this source of replication is utilized to obtain most stable estimates of s_g for individual genes. The moderated t-statistic ($\tilde{t}$) is thus identical to the standard version, except that an improved estimator of s_g ($\tilde{s}_g$) is substituted into the denominator (Smyth 2004).

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}$$

The estimator $\tilde{s}_g$ is improved in the sense that it represents a weighted average between the observed $\hat{s}_g$ for each individual gene and a prior value s_0. The critical choice of s_0 is made within an empirical Bayesian framework, utilizing information from

all genes, and is explained below. Ultimately, this shrinkage towards a prior estimate s_0 prevents problems associated with very large or very small $\hat{s}_g$, thereby increasing the stability of associated with the denominator of $\tilde{t}$. A reflection of this added stability is that $\tilde{t}$ is now associated with degrees of freedom $d_g + d_0$, where d_0 represents the prior degrees of freedom associated with s_0. For a fixed level of α, therefore, a smaller observed value of $|\tilde{t}|$ is required to obtain a significant result.

The $\tilde{s}_g^2$ is a weighted value between the observed and prior variance for individual genes, with the weighting determined by the relative magnitudes of d_g and d_0 (Smyth 2004).

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

The value of d_g represents the residual degrees of freedom associated with the linear model fit to gene *g*, and is therefore proportional to the replication that is available. The value of d_g will be identical among all genes if no values are missing from the data. The prior degrees of freedom d_0 is estimated based upon how $\hat{s}_g$ are distributed among all genes represented on arrays (see below). In particular, d_0 will be large when variances are homogenous among all genes, and conversely, d_0 will be small in magnitude if variances differ greatly among genes. If the $\hat{s}_g^2$ are similar among genes, therefore, $\tilde{s}_g^2$ is nearly equal to the prior s_0^2 for all genes, in which case $\tilde{t}$ is directly

proportional to the observed fold-change. If the $\hat{s}_g^2$ differs considerably among genes, however, $\tilde{s}_g^2$ is determined by the $\hat{s}_g^2$ associated with individual genes to a greater extent, shrinkage towards the prior s_0^2 is limited, and $\tilde{t}$ becomes increasingly similar to the standard *t*-statistic discussed in section 1.

The moderated t-statistic can be viewed as a type of offset *t*-statistic along the lines of those developed by Tusher et al. (2001), Efron et al. (2001) and Brobery (2003). Offset *t*-statistics include an added (constant) term in the denominator to prevent small variances from leading to significant differential expression calls. In particular, for the moderated *t*-statistic, if $d_0 < \infty$ and $d_g > 0$ then,

$$\tilde{t}_{gj} = \left(\frac{d_0 + d_g}{d_g}\right)^{1/2} \frac{\hat{\beta}_{gj}}{\sqrt{s_{*,g}^2 v_{gj}}}, \text{ where}$$

$$s_{*,g}^2 = s_g^2 + \left(\frac{d_0}{d_g}\right) s_0^2$$

In the above formulation of the moderated t-statistic, the moderated variance $s_{*,g}^2$ includes the term $(d_0 / d_g)s_0^2$, which sets a lower limit on the magnitude of the denominator. In contrast to the statistics proposed by Tusher et al. 2001, Efron et al. 2001 and Brobery 2003, however, the constant term $(d_0 / d_g)s_0^2$ is not arbitrary, but is instead estimated from the data based upon an underlying distributional theory (described below).

All advantages of $\tilde{t}$ over t cited above are attained at the expense of an additional assumption, as well as the having to estimate hyperparameters upon which the term $\tilde{s}_g^2$ depends (i.e., s_0^2 and d_0). The primary assumption introduced is that the estimators $\hat{\boldsymbol{\beta}}_g$ and s_g^2 are independent for different genes, which is not likely to be valid for most microarray datasets. The estimation of s_0^2 and d_0 is somewhat less problematic, due to the well developed distributional framework in which these parameters are embedded, which allows stable estimates based on information borrowed from all genes considered in the analysis. A sketch of the manner in which estimators of s_0^2 and d_0 are derived is provided below, further details of which can be found in Smyth (2004).

A hierarchical model is assumed that describes the manner in which σ_g^2 varies across all genes. Given the prior estimator s_0^2 with d_0 degrees of freedom, and noting the relationship $ks^2/\sigma^2 \sim \chi_k^2$ expected under sampling theory (Rice 1995), it is assumed that σ_g^2 follows a scaled chi-square distribution across genes.

$$\sigma_g^2 \sim \frac{d_0 s_0^2}{\chi_{d_0}^2}$$

If it is further assumed that the β_{gj} approximate a normal distribution, the joint distribution of the observed $\hat{s}_g^2$ and $\hat{\beta}_{gj}$ under the null hypothesis can be derived by carrying out the following integration (Smyth 2004).

$$P(\hat{\beta}_{gj}, s_g^2 | \beta_{gj} = 0) = \int P(\hat{\beta}_{gj} | \sigma^2, \beta_{gj} = 0) P(s_g^2 | \sigma^2) P(\sigma^2) d\sigma^2$$

All three probability terms in the integrand are known by the assumptions stated above. Following substitution of the appropriate normal or chi-square distributions and integration of the resulting term, it can be shown based upon the joint distribution $P(\hat{\beta}_{gj}, s_g^2 | \beta_{gj} = 0)$ that that the s_g^2 among genes is expected to follow a surprisingly simple distribution (Smyth 2004).

$$s_g^2 \sim s_0^2 F_{d_g, d_0}$$

The distributional result stated above allows for estimation of the s_0^2 and d_0 hyperparameters based upon the observed distribution of s_g^2 among all genes. In particular, if $z_g = \log s_g^2$ and $f(\circ)$ to represents the digamma function (Smyth 2004), define the following quantity e_g for each gene.

$$e_g = z_g - f(d_g / 2) + \log(d_g / 2)$$

Lastly, letting G represent the number of genes present on arrays and $h(\circ)$ represent the trigamma function, a closed form estimators of d_0 and s_0^2 are provided by the following equations (Smyth 2004).

$$d_0 = 2h^{-1}\left[mean\left\{ G\left(\frac{e_g - \bar{e}}{G-1} \right) - h(d_g/2) \right\} \right]$$

$$s_0^2 = \exp\{\bar{e} + f(d_0/2) - \log(d_0/2)\}$$

The posterior odds statistic

The posterior odds statistic provides a convenient summary measure for comparing the probability that a gene is differentially expressed ($\beta_{gj} \neq 0$) to the probability that a gene is not differentially expressed ($\beta_{gj} = 0$) with respect to the *jth* contrast of interest. This statistic was originally developed by Lönnstedt and Speed (2002) and is primarily intended to provide a means of ranking genes, rather than as a device upon which formal inferences can be based. The limitations of the posterior odds (O_{gj}) statistic are immediately apparent from its definition below (Smyth 2004), in which p_j represents the proportion of genes that are differentially expressed.

$$O_{gj} = \frac{P(\beta_{gj} \neq 0 | \tilde{t}_{gj}, s_g^2)}{P(\beta_{gj} = 0 | \tilde{t}_{gj}, s_g^2)} = \frac{P(\beta_{gj} \neq 0, \tilde{t}_{gj}, s_g^2)}{P(\beta_{gj} = 0, \tilde{t}_{gj}, s_g^2)} = \left(\frac{p_j}{1-p_j} \right) \frac{P(\tilde{t}_{gj} | \beta_{gj} \neq 0)}{P(\tilde{t}_{gj} | \beta_{gj} = 0)}$$

The last ratio reveals that the posterior odds are dependent upon the value of p_j, which is unknown and cannot be reliably estimated from the data. Moreover, substituting the distribution of $\tilde{t}_{gj}$ into the last ratio reveals that O_{gj} is dependent upon an additional hyperparameter v_{0j}.

$$O_{gj} = \left(\frac{p_j}{1-p_j}\right)\left(\frac{v_{gj}}{v_{gj}+v_{0j}}\right)^{1/2}\left(\frac{\tilde{t}_{gj}^2 + d_0 + d_g}{\tilde{t}_{gj}^2 \dfrac{v_{gj}}{v_{gj}+v_{0j}} + d_0 + d_g}\right)^{(1+d_0+d_g)/2}$$

The value v_{0j} represents the unscaled variance of β_{gj}, such that $v_{0j}^{1/2}\sigma_g$ is the standard deviation of the log fold-changes associated with differentially expressed genes. Smyth (2004) derived a closed form estimator of v_{0j}, which had not previously been obtained in the original formulation of Lönnstedt and Speed (2002). This estimator, however, is based only on genes that are differentially expressed, rather than on all genes represented on the array (as with d_0 and s_0^2). Given that estimates of v_{0j} are based upon a smaller number of genes, it is likely that these estimates are less stable than those obtained for other hyperparameters. In support of this conjecture, low stability of v_{0j} estimates was, in fact, observed in simulation studies carried out by Smyth (2004). These concerns, however, are compounded by the dependency of O_{gj} on p_j, for which no estimator is available.

Arguably, it is conceivable that researchers could, in some cases, have a rough *a priori* idea regarding what fraction of genes (p_j) should be differentially expressed in a particular experiment. It is clear that some treatments should have a larger impact on a transcriptome than other treatments. It is reasonable, for example, to suppose that treating an *Arabidopsis* plant with a genotoxic stress will have much smaller effect on expression levels than prolonged temperature stress. Nevertheless, there does seem to be little basis

for specifying particular values of p_j, such as whether 10%, 20% or 30% of the genome should be differentially expressed by genotoxic stress. These considerations become especially awkward given the expected dependence of p_j on the number of arrays used in an experiment and consequent levels of statistical power.

In light of the above considerations, O_{gj} is directly proportional to $\tilde{t}_{gj}$, and for this reason, rankings of genes based upon O_{gj} are in agreement of those based upon $\tilde{t}_{gj}$ (when there are no missing values in the data). The statistic O_{gj} therefore remains a useful tool for ranking genes, despite its limitations as a tool for drawing statistical inferences.

Section 4 Discussion

Statistical tests for differential expression are widely used throughout the research community and are a starting point in the analysis of most microarray datasets. A number of procedures have now been developed that represent considerable improvements over methods employed in early studies of gene expression. The Limma package allows researchers to test hypotheses for two sample comparisons, but also provides a flexible linear model implementation that permits testing of hypotheses that are of considerably greater complexity. This chapter has reviewed the basic framework of linear models as applied to microarray data, as well as the statistics used by Limma to rank genes and carry out inferences regarding differential expression. Perhaps the most critical point emerging from this review is that statistical inferences based upon Limma's statistics are dependent upon key distributional assumptions, and should therefore be applied with some caution. It is clear, for example, that the posterior odds statistic should only be used as a tool for ranking genes with respect to evidence for differential expression. The moderated *t*-statistic, in contrast, is of greater potential utility as an inferential statistic. However, as noted by Smyth (2004), the moderated *t*-statistic assumes that variances associated with gene expression measurements are independent across genes, in addition to a number of distributional requirements that need to be approximately valid for the dataset under consideration. Like the posterior odds, therefore, the moderated *t*-statistic is best used as a means for ranking genes according to evidence of differential expression. Overall, these considerations underscore the challenges entailed by the complexity of microarray data. For some experimental designs, especially those for which a small

number of arrays have been used, ideal methods for drawing inferences regarding differential expression remain to be developed.

Simulation analyses are an important part of evaluating the efficiency of statistical procedures, and provide a useful means to compare new and pre-existing methodologies. Smyth (2004) carried out simulation analyses to explore how results obtained using the statistics implemented in Limma compared with those based upon fold-change, the ordinary *t*-statistic, an offset *t*-statistic (Efron et al. 2001), and the *B*-statistic developed by Lönnstedt and Speed (2002) (implemented in the SMA package for R). Several different sets of simulated data were generated, with each set exhibiting a different pattern of variance (σ^2) among genes. The results of these investigations revealed that, when variances are homogenous among genes, alternative methods were comparable in terms of false discovery rates. Differences among alternative methods, however, were more substantial when variances were heterogeneous among genes. In such cases, it was found that the moderated *t*-statistic exhibited superior performance to alternative methods in terms of false discovery rates and the area underneath ROC curves. The second main finding from Smyth (2004)'s simulation analyses was that estimation of the hyperparameters s_0^2 and d_0 was very accurate, while in contrast, estimation of v_0 could be poor if the proportion of differentially expressed genes is specified incorrectly. In particular, if the proportion of differentially expressed genes is not correctly specified, then v_0 is generally overestimated by as much as 73% (when variances are highly similar among genes). Since the posterior odds statistic is dependent upon the v_0 hyperparameter, the simulations illustrate why this statistic provides a poor basis for inferences about differential expression.

It should be noted that the simulations carried out by Smyth (2004) used data generated under the assumptions of the hierarchical model, such that variances among genes followed an inverse chi-square distribution and contrast coefficients were approximately normal in distribution. In an important sense, therefore, simulations were tailored to the statistics implemented in the Limma package, such that the relative performance of the Limma procedure could be inflated beyond that typically observed on real datasets. A second important point is that the P-values associated with the moderated *t*-statistic were not explicitly checked for accuracy. Hence, although the relative performance of the moderated *t*-statistic was superior to that of other methods, the accuracy associated with p-values is not guaranteed (especially with respect to real datasets for which hierarchical model assumptions may not be valid).

In addition to simulation analyses, evaluating the relative performance of new methods on real datasets is critical to judge how well underlying assumptions conform to biological variability. When applied to real datasets, Smyth (2004) found that the moderated *t*-statistic and posterior odds yielded rankings that were consistent with prior biological knowledge associated with the genes being analyzed. Alternative methods, in contrast, were in some cases not consistent with preexisting biological knowledge. In Figures 1-2 and 1-3, the associations between gene rankings based upon fold-change, the ordinary *t*-statistic, and the moderated *t*-statistic are shown. These plots were constructed from the top 30 genes that emerged from Smyth (2004)'s analysis of the Swirl dataset (Dudoit and Yang 2003). There was a weak overall association between ranks based upon fold-change and the moderated *t*-statistic ($r = 0.402$), although there was an overall stronger association between ranks based upon the ordinary and moderated *t*-statistics (*r*

= 0.765). An interesting difference evident from Figures 1-2 and 1-3 is that, with respect to fold-change, differences from Limma rankings are of the same magnitude for both large and small rankings. With respect to the ordinary *t*-statistic, however, differences from Limma rankings are large for strongly upregulated genes, but rather small for strongly downregulated genes. This distinction most likely reflects the shrinkage of large ordinary *t*-statistics that arise from spuriously low variance estimates associated with some genes.

An ideal procedure for differential expression analysis would be powerful enough to reliably detect differentially expressed genes, but at the same time, not critically dependent upon distributional assumptions that are unlikely to be valid for most microarray datasets. It could be argued, for example, that if p-values generated by a differential expression procedure are not meaningful due to heavy dependence upon shaky assumptions, the purpose behind developing a probabilistic framework has been defeated. At the other extreme, however, two-sample non-parametric methods, which are free of restrictive assumptions, often lack the power to detect many differentially expressed genes of biological relevance and are therefore not useful. The moderated *t*-statistic and posterior odds have a number of positive aspects that represent improvements over earlier approaches, even if all properties associated with the hypothetical "ideal statistic" are not satisfied. Since differential expression analysis remains an active area of investigation, there are prospects for continued improvement through the development of new methodologies (e.g., Pan 2003; Lu et al. 2005; Zou and Hastie 2005).

Although differential expression analysis has been the primary focus of this review, it is important to note that, in most experiments, biologically important transcriptional changes will occur that are undetected. Differential expression analysis effectively divides the transcriptome into "important" versus "non-important" genes (with regard to the biological phenomena under consideration). While this schism is a convenient summary device for the human mind seeking to understand a given process, it may often be an inappropriate characterization. The true importance of genes with regard to a particular process (e.g., stress response, cancer proliferation) is likely to be as continuous as the p-values generated by a differential expression analysis. As the number of arrays used to investigate a particular treatment increases, the effect sizes declared significant by differential expression analysis become increasingly small, and correspondingly, the list of "important genes" will grow in length. The concept of differential expression, therefore, is relative in a key sense, and in most cases, it will be misleading to believe that genes not declared differentially expressed are unimportant. An important future goal is to develop investigative methodologies that go beyond differential expression. Such methods will utilize whole-genome datasets more efficiently by recognizing small transcriptional changes below the differential expression threshold, yielding outputs of greater depth than the "list of genes" approach, which (hopefully) can be interpreted in a biologically meaningful way.

An additional issue relevant to all differential expression analysis methods is the means by which multiple comparisons are accounted for. Regardless of the test statistic that is used, differential expression analysis involves performing as many tests as there are genes represented on a given array. The p-values generated by a given test statistic

must therefore be adjusted to account for the large number of tests that are being performed. The Bonferroni method is the simplest approach to controlling for a large number of hypothesis tests. Bonferroni p-value adjustments control the family-wise error rate by dividing p-values associated with each gene by n, where n is the number of genes being tested in the differential expression analysis. Following this adjustment, genes associated with P-values less than a nominal type-I error rate of α can be declared as differentially expressed at level α. The Bonferroni approach seems to perform moderately well in many applications in the biological sciences (Sokal and Rohlf 1995). In the context of differential expression analysis, however, it is generally agreed that it yields results that are far too conservative, and may lead to a failure to detect genes that are of biological importance (Allison et al. 2006). To remedy this short-coming, it was proposed that controlling the false discovery rate would be more appropriate in the context of differential expression analysis (Benjamini and Hochberg 1995). This proposal has been well-received, such that the Benjamini-Hochberg approach to adjusting for multiple testing via control of the false discovery rate has become widely applied. The Benjamini-Hochberg method, however, assumes that p-values generated for each individual gene are independent of one another. This assumption is unlikely to be valid in most cases, although it has been argued that the method is robust to certain types of dependency structures among genes (Reiner et al. 2003). Some recent approaches have attempted to relax the independence assumption by developing resampling methods to control for false discovery rate (Pollard and van der Laan 2004; van der Laan et al. 2004). In addition, a large number of methods have been developed based upon mixture model distributions (Pounds and Morris 2003; Datta and Datta 2005; Do et al. 2005), which are

not free of assumptions, but are generally more powerful than the Benjamini-Hochberg method (Allison et al. 2006).

The generation of biological knowledge from DNA microarray technology depends critically on the statistical validity that underlies differential expression procedures. Differential expression, however, is just one of several issues that arise in the context of DNA microarray data analysis. Apart from differential expression, the reliability associated with data preprocessing and normalization, clustering methods, and classification algorithms are also of considerable importance. There are several reasons for optimism regarding the prospects for future improvements of procedures applied to all phases of microarray data analysis. The transition time between algorithm development and implementation, for example, should be considerably reduced in light of a freely available user-friendly statistical computing environment (Gentleman et al. 2004). A second positive development is the widespread adoption of the MIAME (minimum information about a microarray experiment) standards (Brazma et al. 2001), which will increase the efficiency with which analyses can be carried out, and when necessary, independently repeated by new investigators. With the abundance of tools now available, the volumes of data generated by microarrays will serve as an increasingly valuable resource for addressing biologically significant research questions.

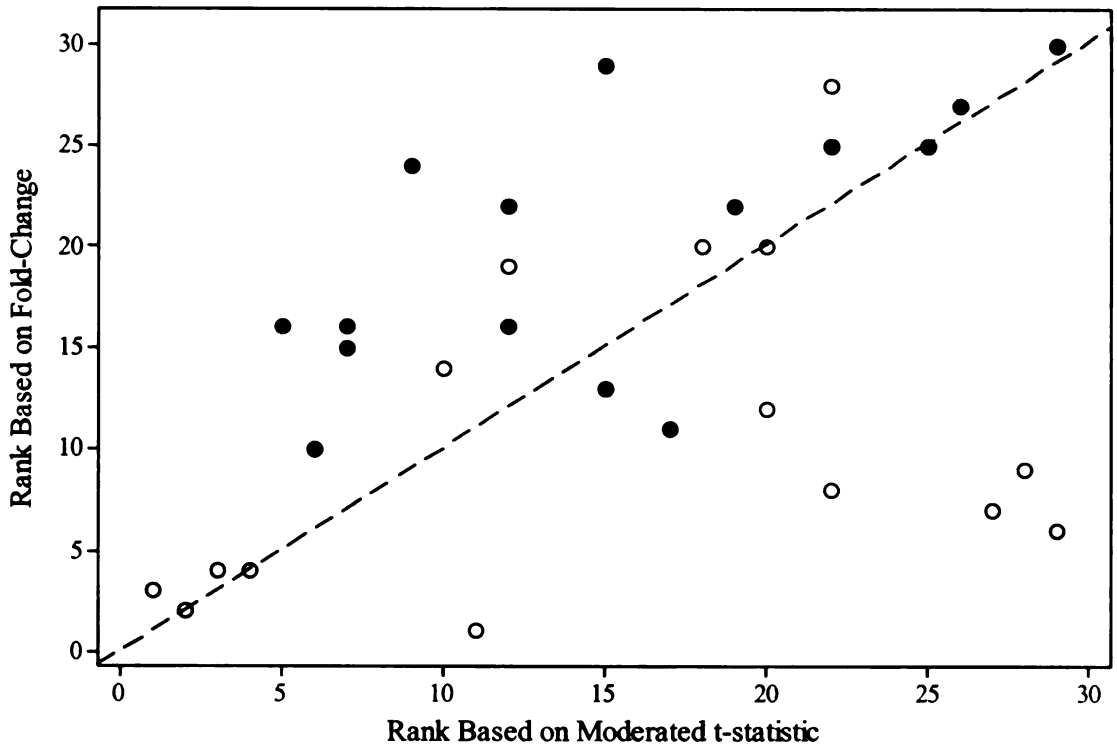

Figure 1-2. Comparison of ranks based on the moderated t-statistic with ranks based on fold-change. Smyth (2004) calculated both the moderated t-statistic and fold-change for genes within the Swirl dataset (Dudoit and Yang 2003). The scatterplot above displays the association among ranks calculated based on the two methods for the top 30 genes. Open circles represent strongly down-regulated genes, while closed circles represent strongly up-regulated genes.

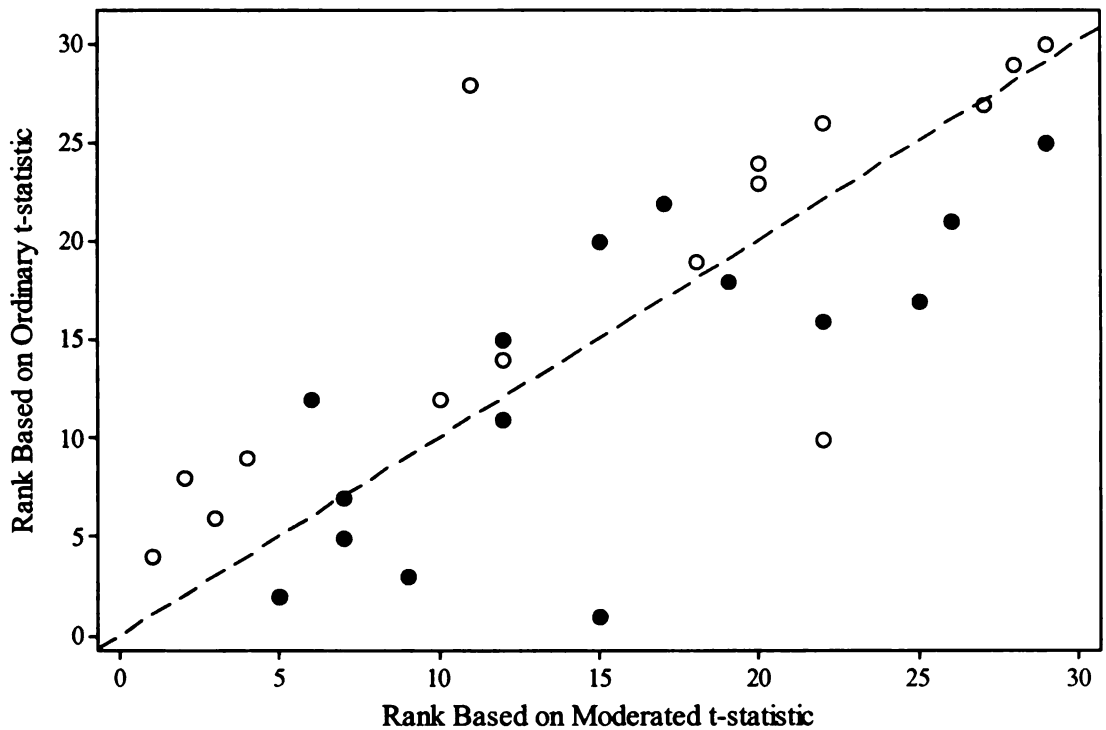

Figure 1-3. Comparison of ranks based on the moderated t-statistic with ranks based on ordinary t-stastistic. Smyth (2004) calculated both the moderated t-statistic and the ordinary t-statistic for genes within the Swirl dataset (Dudoit and Yang 2003). The scatterplot above displays the association among ranks calculated based on the two methods for the top 30 genes. Open circles represent strongly down-regulated genes, while closed circles represent strongly up-regulated genes.

Bibliography

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

Alter, O., P. Brown, and D. Botstein. 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA 97: 10101-10106.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a powerful and practical approach to multiple testing. J. Roy. Stat. Soc. B. 57: 289-300.

Boutros, P. C., I. D. Moffat, M. A. Franc, N. Tijet, J. Tuomisto, R. Pohjanvirta, and A. B. Okey. 2004. Identification of the DRE-II gene battery by phylogenetic footprinting. Biophys. Res. Commun. 312(3): 707-715.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitiz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. 2001. Minimum information about a microarray experiment (MIAME) – towards standards for microarray data. Nat. Genet. 29: 365-371.

Broberg, P. 2003. Statistical methods for ranking differentially expressed genes. Genome Biol. 4: R41.

Brown, P. O., and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. Nature Genet. 21(suppl.): 33-37.

Brown, M. P., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA 97(1): 262-262.

Datta, S., and S. Datta. 2005. Empirical bayes screening of many p-values with applications to microarray studies. Bioinform. 21: 1987-1994.

Do, K. A., P. Mueller, and F. A. Tang. 2005. A nonparametric Bayesian mixture model for gene expression. J. R. Stat. Soc. Ser. C. 54: 1-18.

Dudoit, S., and Y. H. Yang. 2003. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. Pp. 73-101 *in* G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, eds. The analysis of gene expression data: methods and software. Springer, New York.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher. 2001. Empirical bayes analysis of a microarray experiment. J. Amer. Stat. Assoc. 96: 1151-1160.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95(25): 14863-14868.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. 2004. Bioconductor: open source software development for computational biology and bioinformatics. Genome Biol. 5: R80.

Golden, R. T., and S. Melov. 2004. Microarray analysis of gene expression with age in individual nematodes. Aging Cell 3: 111-124.

Jain, A. D., T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel. 2002. Fully automatic quantification of microarray image data. Genome Res. 12: 325-332.

Lander, E. S. 1999. Array of hope. Nat. Genet. 21(suppl.): 3-4.

Lönnstedt, I., and T. P. Speed. 2002. Replicated microarray data. Statist. Sin. 12: 31-46.

Lu, Y., P. Y. Liu, P. Xiao, and H. W. Deng. 2005. Hotelling's T2 multivariate profiling for detecting differential expression of microarrays. Bioinform. 21: 3105-3113.

Pan, W. 2003. On the use of permutation in and the performance of nonparametric methods to detect differential gene expression. Bioinform. 19: 1333-1340.

Peart, M. J., G. K. Smyth, R. K. van Laar, V. M. Richon, A. J. Holloway, and R. W. Johnstone. Identification and functional significance of genes regulated by structurally diverse histone deacetylase inhibitors. Proc. Natl. Acad. Sci. USA 102(10): 3697-3702.

Pollard, K. S., and M. J. van der Laan. 2004. Choice of a null distribution in resampling-based multiple testing. J. Stat. Plann. Infer. 125: 85-100.

Pounds, S., and S. W. Morris. 2003. Estimating the occurrence of false-positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinform. 19: 1236-1242.

Quackenbush, J. 2002. Microarray data normalization and transformation. Nat. Genet. 32(suppl.): 496-501.

Reiner, A., D. Yekutieli, and Y. Benjamini. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinform. 19: 368-375.

Renn, S. C. P., N. Aubin-Horth, and H. A. Hofmann. 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cdna microarray. BMC Genomics 5(1): 42.

Rensink, W. A., S. Lobst, A. Hart, S. Stegalkina, J. Liu, and C. R. Buell. 2005. Gene expression profiling of potato responses to cold, heat, and salt stress. Funct. Integ. Genom. 5: 201-207.

Rice, J. A. 1995. Mathematical statistics and data analysis. Duxbury Press, Belmont, California.

Rodriguez, M. W., A. C. Paquet, Y. H. Yang, and D. J. Erle. 2004. Differential gene expression by integrin $\beta 7+$ and $\beta 7-$ memory T helper cells. BMC Immun. 5:13.

Smyth, G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3(1):3.

Sokal, R. R., and Rohlf, F. J. 1995. Biomety. W. H. Freeman, New York.

Tusher, V. G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. 98: 5116-5121.

van der Laan, M. J., S. Dudoit, and K. S. Pollard. 2004. Multiple testing part I: single-step procedures for control of general type I error rates. Stat. Appl. Genet. Mol. Biol. 3: 13.

Welch, B. L. 1947. The generalization of 'students' problem when several different population variances are involved. Biometrika 34: 28-35.

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

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. JR. Stat. Soc. Ser. B. Stat. Methodol. 67: 301-320.

CHAPTER 2

TRANSCRIPTIONAL PROFILING OF ARABIDOPSIS HEAT SHOCK GENES

Summary

The heat shock response of *Arabidopsis thaliana* is dependent upon a complex regulatory network involving twenty-one known transcription factors and several heat shock protein families. It is known that heat shock proteins (Hsps) and transcription factors (Hsfs) are involved in cellular response to various forms of stress besides heat. However, the role of Hsps and Hsfs under cold and non-thermal stress conditions is not well understood, and it is unclear which types of stress interact least and most strongly with Hsp and Hsf response pathways. We have examined transcriptional response profiles of *Arabidopsis* Hsfs and Hsps to a range of abiotic and biotic stress treatments (heat, cold, osmotic stress, salt, drought, genotoxic stress, ultraviolet light, oxidative stress, wounding, and pathogen infection) in three different types of tissue (roots, shoots, leaves). Our findings demonstrate that nearly all stress treatments interact with Hsf and Hsp response pathways, suggesting considerable cross-talk between heat and non-heat stress regulatory networks. We identified several heat shock gene expression patterns that have not been previously described. First, with respect to the Hsp20 protein family, large expression responses occurred under all types of stress, with striking similarity among expression response profiles. Second, a number of Hsp20, Hsp70 and Hsp100 genes were specifically upregulated twelve hours after wounding in root tissue, and exhibited a similar expression response pattern during recovery from heat stress. Lastly, among all protein families, large expression responses occurred under ultraviolet-B light stress in shoot tissue but not root tissue. These findings have implications regarding the molecular basis of cross-tolerance in plant species, and raise several new questions to be pursued in future experimental studies of the *Arabidopsis* heat shock response network.

Section 1 Introduction

The heat shock response network of *Arabidopsis thaliana* involves temperature perception mechanisms, an intricate array of signal transduction networks, and twenty-one known transcription factors that activate heat shock proteins and other heat-stress related genes (Nover et al. 2001; Sung et al. 2003; Larkindale et al. 2005). The availability of genome sequence data has considerably advanced our understanding of this heat shock response pathway, as well as the molecular basis of regulatory networks that underlie other forms of environmental stress in *Arabidopsis* (e.g., cold, salinity, drought). One result of this development has been increased recognition of the cross-talk or overlap that exists among cellular responses to different environmental stress treatments (e.g., Cheong et al. 2002; Rensink et al. 2005; Ma et al. 2006; Mittler 2006; Rossel et al. 2006). In this respect, heat shock proteins (and their associated transcription factors) are of special interest. Heat shock proteins are molecular chaperones that regulate the folding, localization, accumulation, and degradation of protein molecules in both plant and animal species (Feder and Hofmann 1999). Heat shock proteins are therefore believed to play a broad role in many cellular processes, which may impart a generalized role in tolerance to multiple environmental stress treatments apart from heat stress. Understanding the role of heat shock proteins under cold and non-thermal stress conditions may therefore provide insight into multiple stress tolerance mechanisms (Hoffmann and Parsons 1991). This may be of considerable importance for improving the production of agriculturally important crop species under field conditions, which are best characterized as an interaction of several different types of stress, rather than just a single stress treatment in isolation (Mittler 2006).

The *Arabidopsis* heat shock proteins (Hsps) and transcription factors (Hsfs) have been well characterized on the basis of genome sequence information (Agarwal et al. 2001; Lin et al. 2001; Nover et al. 2001; Krishna and Gloor 2001; Scharf et al. 2001). In addition to the twenty-one known transcription factors (Nover et al. 2001), the *Arabidopsis* heat shock response is partly mediated by thirteen Hsp20 proteins (Scharf et al. 2001), eighteen Hsp70 proteins (Lin et al. 2001), seven Hsp90 proteins (Krishna and Gloor 2001), and up to eight members of the Hsp100 protein family (Agarwal et al. 2001). The molecular pathways leading to Hsp expression are not entirely understood (Sung et al. 2003), but involve temperature perception mechanisms coupled with multiple signal transduction pathways (Larkindale et al. 2005), which together lead to the activation of Hsfs that induce expression of heat shock genes by binding to heat shock elements (Schöffl et al. 1998). There are several levels at which this molecular pathway may overlap with those underlying response to cold and non-thermal stress treatments. However, since Hsps play a uniquely broad role in cellular processes, they are particularly likely to underlie interactions between heat and non-heat stress response pathways. A role of heat shock proteins in cellular response to cold and non-heat stress treatments, for instance, has been supported by several gene expression studies. In *Arabidopsis* and other plant species, various Hsps have been induced by low temperature (Sabehat et al. 1998), osmotic stress (Sun et al. 2001), salt (Liu et al. 2006a), oxidative stress (Banzet et al. 1998; Lee et al. 2000; Volkov et al. 2006), desiccation (Liu et al. 2006b), exposure to intense light (Desikan et al. 2001; Hihara et al. 2001; Rossel et al. 2002), wounding (Cheong et al. 2002), and heavy metal exposure (Györgyey et al. 1991).

While a number of studies have shown the Hsp expression can be induced under cold and non-thermal stress treatments, no comparative analyses have been carried out to identify which particular stress treatments are the weakest and strongest inducers of Hsp expression. It therefore remains unclear which stress-response pathways overlap most extensively with this important part of the *Arabidopsis* heat shock regulatory network. If the primary stress conditions interacting with Hsp response pathways can be identified, it would be of considerable interest to understand how Hsfs and Hsps contribute to tolerance under such stress conditions. The physiological role of Hsfs and Hsps in promoting tolerance may differ depending on the nature of the stress imposed upon the cell. Heat stress, for instance, leads directly to denaturation of cellular proteins. It is therefore clear how molecular chaperone activity may contribute to high temperature tolerance via prevention of deleterious protein conformations and elimination of non-native aggregations. With respect to cold and non-thermal stress treatments, however, the impact on cellular protein conformations is less direct and not as well understood. The role of Hsps as molecular chaperones, therefore, may not strictly parallel their function under heat stress, and it is possible that their cellular function extends beyond the chaperone activity that has been well characterized *in vitro* (e.g., Lee et al. 1997; Lee and Vierling 2000). One possibility, for example, is that Hsps limit damage resulting from accumulation of reactive oxygen species, which are generated as messengers and elements of signal transduction pathways under a wide range of stress conditions (Pastori and Foyer 2002). In both plant and animal species, for instance, there is evidence to suggest that Hsps protect against reactive oxygen species (Mehlen et al. 1993; Fleming et al. 1992; Wheeler et al. 1995; Härndahl et al. 1999; Preville et al. 1999; Ding and Keller

2001; Martindale and Holbrook 2002; Kregel 2002; Neta-Sharir et al. 2005). This hypothesis is particularly intriguing in light of the considerable interconnectivity that exists between heat shock and oxidative stress response pathways in plant species (Desikan et al. 2001; Panchuk et al. 2002; Miller and Mittler 2006; Volkov et al. 2006).

DNA microarray technology offers a promising approach for better understanding the functional role of *Arabidopsis* heat shock proteins and transcription factors under both heat and non-heat stress conditions. Recently, a number of genome-wide microarray datasets have been generated and made publicly available by the AtGenExpress consortium. These resources provide an opportunity to profile Hsf and Hsp expression over a wide range of stress conditions simultaneously. In this study, we utilized AtGenExpress datasets to analyze transcriptional responses of *Arabidopsis* Hsfs and Hsps to a total of ten different abiotic and biotic stress treatments (cold, osmotic stress, salt, drought, genotoxic stress, ultraviolet light, oxidative stress, wounding, high temperature, and pathogen infection). The gene expression data we consider was generated from three different types of plant tissue (roots, shoots, leaves), with expression measurements obtained at up to six different time points of stress exposure (0.5, 1, 3, 6, 12, and 24 hours). With respect to each of five protein families (Hsf, Hsp20, Hsp70, Hsp90, and Hsp100), we evaluated whether expression responses of each family to each stress were significantly large in comparison to other *Arabidopsis* genes. This analysis provided indication of which types of stress interacted most and least with each protein family. In addition, we characterized Hsf and Hsp stress-response patterns at the level of protein families, as well as among individual genes within protein families. This allowed identification of family-level expression patterns under each stress, gene sub-groups

within families exhibiting similar expression patterns, and individual Hsf/Hsp genes with large expression responses to several different stress treatments.

Section 2 Results

An overview of how strongly each stress treatment impacted expression levels for each heat shock gene family is provided in Tables 1 and 2 for root and shoot tissue, respectively. To compare the effect of each stress, a summary statistic was developed (T) that represents the median level of fold-change induced by each stress among members of a given protein family (see Equation 2 in Methods). For the Hsp20, Hsp70 and Hsp90 protein families, the largest expression responses were associated with the high temperature treatment, with median levels of fold-change in each family (T) ranging from one to above four (Tables 1 and 2). However, for the Hsf and Hsp100 protein families, the largest magnitude expression responses were associated with osmotic stress (Table 1). For each stress, it was of interest to determine whether median-level expression responses of each gene family were large in comparison to all other genes represented on the ATH1 array. A resampling procedure was therefore carried out to evaluate the likelihood of observed T statistics under a null hypothesis of random sampling from the genome (see Table 1 caption and Methods). Significant T statistics were found with respect to each type of stress we considered, indicating that for one or more protein families, each stress induced expression responses that were large in comparison to other *Arabidopsis* genes (see Tables 1 and 2). High temperature was associated with a significant T statistic for all protein families except Hsp100 in both roots and shoot. The second strongest elicitor of expression responses was oxidative stress, since it was associated with a significant T statistic for most families in both tissue types. Interestingly, for the pathogen stress treatment (not shown in Tables 1 and 2), significant T statistics were associated with

respect to the Hsp70 and Hsp90 families (P = 0.001 and 0.018, respectively), while *T* statistics were non-significant for the Hsf, Hsp20 and Hsp100 families (P > 0.096).

Protein families exhibiting strong expression responses to many stress treatments exhibit a generalized expression response pattern. The Hsf and Hsp20 were associated with the most stress-general expression patterns, since for both families, *T* was significant for nearly all types of stress (see Tables 1 and 2). In contrast, the Hsp70, Hsp90, and Hsp100 families were not so widely responsive across all stress treatments. Family-level expression response patterns specific to each tissue-treatment combination and variation among individual heat shock genes are described in the following sections.

Heat Shock Transcription Factors

Heat shock transcription factors were most strongly upregulated under heat, cold, osmotic, and salt stress treatments. Figure 2-1 displays gene expression response profiles for all Hsf genes in roots, while Figure 2-2 displays response profiles of Hsf genes in shoots. In both tissues, expression responses to cold, osmotic, and salt treatment primarily occur over the late stages of stress exposure between 6 and 24 hours (see Figs. 2-1 and 2-2, parts A – C). This pattern contrasts with that observed under heat stress treatment, in which Hsfs were strongly up-regulated during early stages of stress exposure, with the effect diminishing after the heat stress was removed beyond the 6 hr. time point (Figs. 2-1 and 2-2, part I). A general trend among all five heat shock gene groups was a difference between the effects of UV-B stress in shoot tissue in comparison to root tissue. With respect to the Hsf genes, UV-B stress induced strong up-regulation over most time points

in shoot tissue (Fig 2-2G), but yielded comparatively low gene expression responses in roots (Fig. 2-1G).

Considerable variation was observed among expression response patterns associated with individual Hsf genes. To discern which Hsf genes were the least and most stress-responsive across all stress treatments, we ranked genes according to an index (*d*) (see Table 3). The value of *d* represents the mean proportion of time points, among all stress and tissue types we considered, at which a gene was differentially expressed (see Methods). Highly stress-responsive genes were associated with large values of *d*, while genes less responsive to stress were associated with low values of *d*. The seven least-stress responsive Hsf genes were all Class A type Hsfs, and were associated with values of *d* less than or equal to 0.167. The most stress responsive Hsf gene, in contrast, was the one class C transcription factor in the Hsf family (HsfC1, $d = 0.456$). Cluster analysis using the HOPACH algorithm identified three multi-member clusters of Hsf genes with respect to stress-responses across all tissue-treatment-time combinations (stress-clusters 451, 452, and 470) (see Table 2). The heatmap corresponding to this clustering solution is provided in section 1C of supplemental data file 1. For comparison, the Hsfs were also clustered with respect to their expression patterns across the developmental series conditions analyzed by Schmid et al. (2005) (see Table 3). Members of developmental-cluster 60 (see Table 2) exhibited a pattern of tissue-specificity that was found among certain genes from each of the four Hsp families. The expression pattern was characterized by strong upregulation specific to seed stages 6 – 10 (ATGE conditions 81 – 84), roots (17 days) (ATGE condition 9), flowers stage 12 (ATGE conditions 34 – 37),

and flowers stage 15 (ATGE conditions 41 – 45) (see section 1C of supplemental data file 1).

Hsp20 Protein Family

The Hsp20 protein family exhibited the strongest overall responsiveness to environmental stress treatments, as well as the most cohesive family-level expression patterns among member genes. Expression response profiles are displayed in Figures 2-3 and 4 for all Hsp20 proteins in root and shoot tissues respectively. Tissue-specific patterns of Hsp20 stress-response can be discerned from a comparison of Figures 2-3 and 2-4. With respect to the UV-B treatment, for example, strong downregulation of Hsp20 genes occurred between the 3 – 6 hr. time points in roots (Fig. 2-3G). In shoots, however, UV-B stress induced strong upregulation over this same time period (Fig. 2-4G). With respect to the cold stress treatment, Hsp20 genes were downregulated between the 3 – 6 hr. time points in roots (Fig. 2-3A), while no such response pattern was associated with shoots (Fig. 2-4A). More subtle tissue differences were associated with wounding and heat stress. Hsp20 proteins were responsive to both stress treatments, but the temporal dynamics of expression differed between the two tissue types (see Figs. 2-3H, 2-3I, 2-4H and 2-4I).

The expression responses of Hsp20 proteins under wounding and heat stress revealed surprising family-level patterns within root tissue. Nearly all Hsp20 proteins exhibited strong upregulation 12 hrs. following wounding of root tissue (see Fig. 2-3H). Under the heat treatment, Hsp20 upregulation also occurred at the 12 hr. time point (Fig 2-3I), which represented the heat stress recovery period (9 hrs. following cessation of

heat stress). These expression responses during heat stress recovery were a unique aspect of the Hsp20 family, since generally, all other heat shock genes were responsive only while heat stress was directly applied (0.5 – 3 hrs.). For a number of Hsp20 genes, moreover, the 12 hr. upregulation under heat strongly coincided with that observed under the wounding stress treatment. This synchrony between expression response profiles under wounding and heat treatment in root tissue is evident from Figure 2-5, which displays response profiles of nine Hsp20 genes under wounding and heat stress treatments.

Most Hsp20 proteins exhibited strong expression responses to several types of stress. AtHsp14.2-P(r) exhibited the weakest overall responsiveness to stress (d = 0.096), while in contrast, AtHsp18.5-Cl(r) showed the strongest expression responses to stress treatments (d = 0.325) (see Table 4). Cluster analyses revealed strong similarities among Hsp20 genes with respect to stress-response patterns and the developmental series of Schmid et al. (2005) (see section 2C of supplemental data file 1). Members of stress-cluster 30 (AtHsp23.6-M, AtHsp25.4-P) and stress-cluster 42 (AtHsp26.5-P(r), AtHsp15.7-Cl(r), AtHsp22.0-ER) were highly responsive to stress treatments in root tissue. The other multi-gene stress-cluster (21) consisted of Hsp17 proteins entirely (AtHsp17.4-CI, AtHsp17.6-CII, AtHsp17.6B-CI, AtHsp17.6C-CI, AtHsp17.7-CII), and similar to stress-clusters 30 and 42, exhibited large expression responses to all stress treatments, except that strong responses were present in both roots and shoots. The members of all three of these stress-clusters, and most Hsp20 proteins in general, exhibited a similar expression profile among developmental stages (see section 2C of supplemental data file 1). As among certain Hsfs, this developmental expression profile

consisted of high expression levels with respect to roots (17 days), flowers stage 12, flowers stage 15, and seed stages 6 – 10.

Hsp70, Hsp90, and Hsp100 Protein Families

The Hsp70, Hsp90, and Hsp100 protein families were generally associated with smaller magnitude expression responses across stress conditions. However, members of these families were stress-responsive, since differential expression occurred under most stress conditions for nearly every Hsp within these families. Members of Hsp70, Hsp90, and Hsp100 families were most strongly induced by heat, primarily over the early portion of the time course (0.5 -3 hrs.), although several genes within each family exhibited large responses to the cold, osmotic, and salt stress treatments. The Hsp70, Hsp90, and Hsp100 families were all associated with a similar tissue-specific pattern under the UV-B stress condition (see Fig. 2-6). In particular, expression levels of member genes increased at the 3 – 6 hr. time points in shoot tissue, with little or no transcriptional induction in root tissue. In addition, the expression response pattern identified following wounding and heat stress in root tissue was also evident with respect to AtHsp70-5, AtHsp70-8, AtHsp100-1, and to a lesser extent, AtHsp90-1 (see Fig. 2-7).

The individual gene members of the Hsp70, Hsp90, and Hsp100 families are listed in tables 4, 5, and 6, respectively. On the basis of differential expression, these families contained both the least and most stress responsive *Arabidopsis* Hsps. The most stress-responsive was AtHsp70-4 ($d = 0.439$), which was differentially expressed under all stress treatments, including all three time points of exposure to pathogen stress. In contrast, AtHsp100-2 was the least stress-responsive *Arabidopsis* Hsp ($d = 0.009$), and

was differentially expressed with respect to just one time point under the heat stress treatment.

Clustering of Hsp90 genes with respect to stress-response patterns assigned five members to one group (AtHsp90-2, 4, 5, 6, and 7), since these genes were all associated with highly similar (and weak) expression response patterns in root tissue (see section 4C of supplemental data file 1). The remaining AtHsp90-1 exhibited a strong expression response pattern distinct from all other AtHsp90 genes, and was therefore assigned to a singleton cluster (see Table 6). Within Hsp70 and Hsp100 families, clustering with respect to stress-response patterns identified few sub-groups among member genes (see Tables 5 and 7).

Various members of the Hsp70, Hsp90, and Hsp100 families were associated with the same developmental expression pattern found among certain Hsf and Hsp20 genes. This pattern was best exhibited by AtHsp70-4, AtHsp70-11, AtHsp90-1, and AtHsp100-1, all of which were highly expressed in roots (17 days), flowers stage 12, flowers stage 15, and seed stages 6 – 10 (see sections 3C, 4C, and 5C of supplemental data file 1).

Table 1. Oveview. Root tissue. Values of the T statistic associated with each tissue-treatment combination with respect to each of five protein families (Hsf, Hsp20, Hsp70, Hsp90, Hsp100) in the root tissue type. The value of T is proportional to the median level of $\log_2$ fold-change induced by a given stress treatment among the n gene members within a protein family (see Equation 2 in Methods). The P-values associated with each statistic were obtained by genome-wide resampling and represent the probability of obtaining an equal or larger value of T based on 10,000 random samples of n genes from the $N = 22746$ genes represented on the ATH1 array. P-values exceeding 0.0245 in the table below are non-significant following the Benjamini-Hochberg adjustment for multiple testing (with nominal type I error rate of $\alpha = 0.05$).

Treatment	Hsf	Hsp20	Hsp70	Hsp90	Hsp100
	($n = 21$)	($n = 18$)	($n = 13$)	($n = 6$)	($n = 7$)
cold	0.52(0.001)	1.24(< 0.001)	0.45(0.026)	0.32(0.364)	0.25(0.660)
osmotic	0.75(< 0.001)	2.12(< 0.001)	0.48(0.089)	0.74(0.025)	0.66(0.040)
salt	1.17(< 0.001)	1.70(< 0.001)	0.45(0.341)	0.64(0.133)	0.52(0.269)
drought	0.33(0.001)	0.56(< 0.001)	0.26(0.087)	0.15(0.824)	0.20(0.432)
genotoxic	0.24(0.246)	0.72(< 0.001)	0.31(0.061)	0.27(0.258)	0.27(0.243)
oxidative	0.23(0.035)	0.61(< 0.001)	0.26(0.021)	0.19(0.417)	0.31(0.021)
UV-B	0.27(0.050)	0.96(< 0.001)	0.27(0.107)	0.17(0.797)	0.24(0.288)
wounding	0.27(0.043)	1.18(< 0.001)	0.35(0.007)	0.32(0.069)	0.35(0.030)
heat	0.49(0.003)	4.32(< 0.001)	1.55(< 0.001)	1.02(< 0.001)	0.50(0.048)

Table 2. Oveview. Shoot tissue. See Table 1 caption.

Treatment	Hsf	Hsp20	Hsp70	Hsp90	Hsp100
	(n = 21)	(n = 18)	(n = 13)	(n = 6)	(n = 7)
cold	0.62(0.001)	0.51(0.023)	0.44(0.136)	0.57(0.070)	0.35(0.423)
osmotic	0.83(< 0.001)	1.35(< 0.001)	0.76(0.008)	0.46(0.276)	0.65(0.078)
salt	0.52(0.001)	1.00(< 0.001)	0.44(0.017)	0.39(0.116)	0.31(0.275)
drought	0.36(0.002)	0.66(< 0.001)	0.31(0.057)	0.28(0.198)	0.28(0.192)
genotoxic	0.22(0.343)	0.35(0.001)	0.20(0.607)	0.23(0.374)	0.23(0.363)
oxidative	0.29(0.005)	0.78(< 0.001)	0.48(< 0.001)	0.22(0.344)	0.33(0.016)
UV-B	0.50(0.011)	0.69(< 0.001)	0.53(0.024)	0.69(0.020)	0.32(0.499)
wounding	0.33(0.018)	0.62(< 0.001)	0.27(0.225)	0.35(0.099)	0.28(0.248)
heat	0.55(< 0.001)	4.66(< 0.001)	1.34(< 0.001)	1.45(< 0.001)	0.32(0.339)

Table 3. Hsf protein family. Genes are ordered from least to most stress-responsive (according to *d*). The value of *d* represents the mean proportion of time points, among the 19 tissue-treatment combinations considered, at which a gene was differentially expressed. Cluster IDs represent gene groupings determined by the HOPACH clustering algorithm (see Methods). The development cluster analysis was carried out with respect to the developmental series conditions of Schmid et al. (2005). The stress cluster analysis was carried out with respect to expression responses observed under each of the 111 tissue-treatment-time combinations examined in this study. The *j*th digit in each cluster ID indicates the group to which a gene was assigned in the *j*th iteration of the HOPACH algorithm (see Pollard and van der Laan 2005).

Gene Name	Cluster ID (Development)	Cluster ID (Stress)	*d*
HsfA9	31	452	0.018
HsfA5	50	452	0.061
HsfA1a	32	452	0.070
HsfA7b	60	440	0.088
HsfA7a	80	300	0.123
HsfA6a	31	420	0.149
HsfA1b	70	452	0.167
HsfB3	40	430	0.184
HsfA4c	40	452	0.193
HsfA1d	10	451	0.202
HsfA2	60	200	0.228
HsfB4	90	451	0.228
HsfA1e	60	460	0.237
HsfA3	20	410	0.272
HsfA6b	31	100	0.307
HsfB2b	60	470	0.316
HsfA4a	20	490	0.351
HsfA8	20	480	0.395
HsfB2a	60	470	0.439
HsfB1	40	500	0.447
HsfC1	80	600	0.456

Table 4. Hsp20 protein family. Genes are ordered from least to most stress-responsive (according to *d*). See Table 3 caption for an explanation of clustering procedures and the value of *d*.

Gene Name	Cluster ID (Development)	Cluster ID (Stress)	*d*
AtHsp14.2-P(r)	50	60	0.096
AtHsp25.4-P	24	30	0.167
AtHsp17.6-CII	25	21	0.184
AtHsp23.6-M	22	30	0.184
AtHsp22.0-ER	22	42	0.193
AtHsp23.5-M	30	10	0.193
AtHsp17.6B-Cl	28	21	0.202
AtHsp17.7-CII	27	21	0.219
AtHsp21.7-Cl(r)	70	44	0.219
AtHsp26.5-P(r)	26	42	0.219
AtHsp17.6A-Cl	23	22	0.228
AtHsp18.1-Cl	40	41	0.237
AtHsp17.6C-Cl	27	21	0.254
AtHsp17.4-Cl	27	21	0.263
AtHsp15.7-Cl(r)	21	42	0.263
AtHsp17.4-CIII	10	43	0.307
AtHsp15.4-Cl(r)	60	70	0.316
AtHsp18.5-Cl(r)	70	50	0.325

Table 5. Hsp70 protein family. Genes are ordered from least to most stress-responsive (according to *d*). See Table 2 caption for an explanation of clustering procedures and the value of *d*.

Gene Name	Cluster ID (Development)	Cluster ID (Stress)	*d*
AtHsp70-15	5	54	0.202
AtHsp70-5	2	80	0.211
AtHsp70-8	2	70	0.219
AtHsp70-6	7	53	0.228
AtHsp70-17	6	54	0.228
AtHsp70-9	5	52	0.254
AtHsp70-1	5	54	0.272
AtHsp70-10	4	40	0.281
AtHsp70-11	1	60	0.333
AtHsp70-7	7	51	0.351
AtHsp70-3	4	20	0.368
AtHsp70-2	3	30	0.386
AtHsp70-4	1	21	0.439

Table 6. Hsp90 protein family. Genes are ordered from least to most stress-responsive (according to *d*). See Table 2 caption for an explanation of clustering procedures and the value of *d*.

Gene Name	Cluster ID (Development)	Cluster ID (Stress)	*d*
AtHsp90-6	3	2	0.281
AtHsp90-4	3	2	0.298
AtHsp90-5	2	2	0.307
AtHsp90-7	4	2	0.316
AtHsp90-1	1	1	0.333
AtHsp90-2	3	2	0.342

Table 7. Hsp100 protein family. Genes are ordered from least to most stress-responsive (according to *d*). See Table 2 caption for an explanation of clustering procedures and the value of *d*.

Gene Name	Cluster ID (Development)	Cluster ID (Stress)	*d*
AtHsp100-2	2	3	0.009
AtHsp100-5	3	3	0.140
AtHsp100-1	1	1	0.228
AtHsp100-8	3	5	0.246
AtHsp100-4	1	2	0.254
AtHsp100-3	4	3	0.316
AtHsp100-7	3	4	0.368

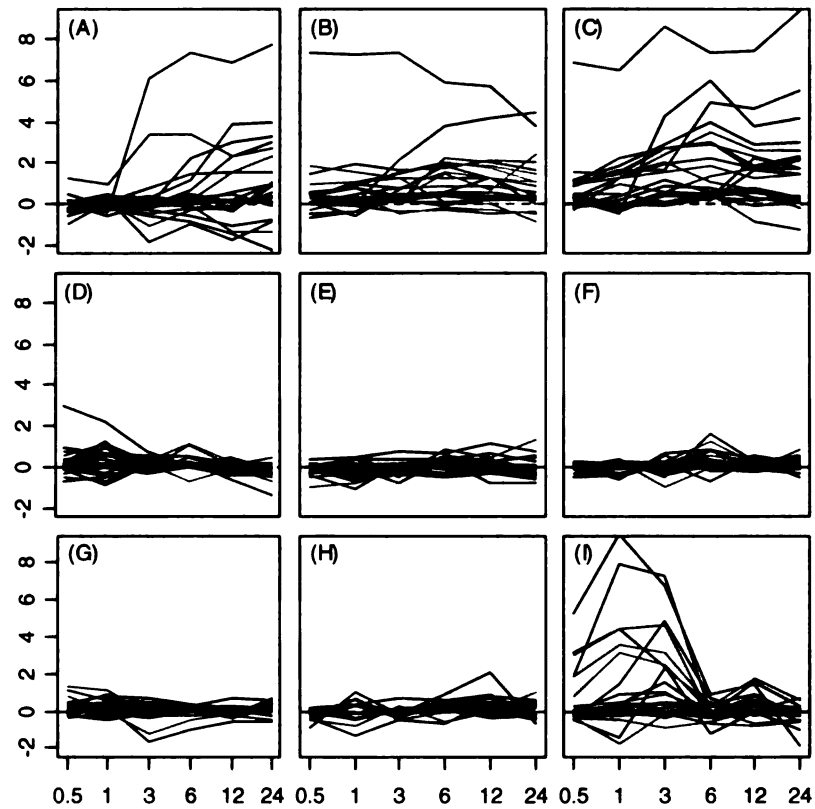

Figure 2-1. Hsf expression response profiles in roots. Plots show profiles associated with (A) cold stress, (B) osmotic stress, (C) salt stress, (D) drought, (E) genotoxic stress, (F) oxidative stress, (G) UV-B light, (H) wounding, and (I) heat. Class A, B, and C transcription factors are represented by black, red, and blue lines, respectively. The horizontal axis of each subplot corresponds to time points at which gene expression measurements were taken under each stress treatments (0.5, 1, 3, 6, and 12 hrs.). The vertical axis of each subplot indicates the $\log_2$ fold-change associated with each Hsf (see Equation 1). The dotted horizontal line in each plot indicates a $\log_2$ fold-change of zero (no expression response to stress). [This image is presented in color]

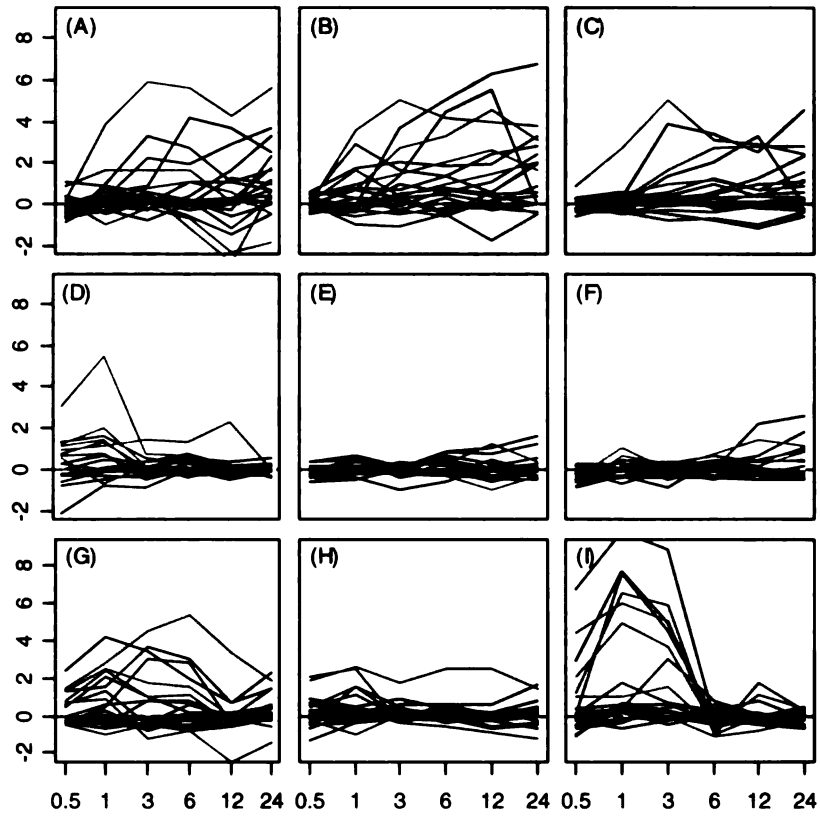

Figure 2-2. Hsf expression response profiles in shoots. See Figure 2-1 caption.

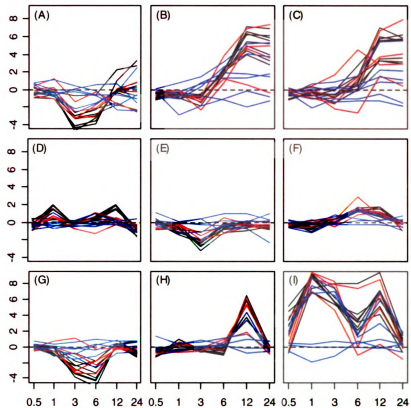

Figure 2-3. Hsp20 expression response profiles in roots. Plots show profiles associated with (A) cold stress, (B) osmotic stress, (C) salt stress, (D) drought, (E) genotoxic stress, (F) oxidative stress, (G) UV-B light, (H) wounding, and (I) heat. The cytoplamic/nuclear Hsp20s (classes I – III) are represented by black lines. Plastidial, endoplasmic reticulum, and mitochondrial Hsp20s (classes P, ER, and M) are represented by red lines. Class I and Class P related Hsp20s are indicated by blue lines. The horizontal axis of each subplot corresponds to time points at which gene expression measurements were taken under each stress treatments (0.5, 1, 3, 6, and 12 hrs.). The vertical axis of each subplot indicates $\log_2$ fold-change (see Equation 1). The dotted horizontal line in each plot indicates a $\log_2$ fold-change of zero (no expression response to stress).

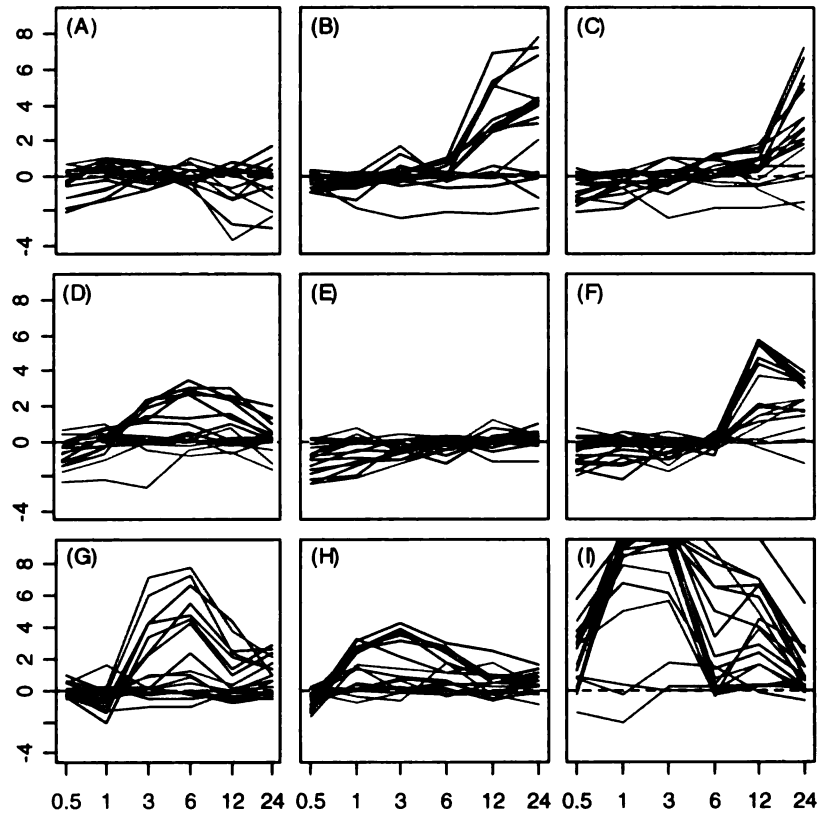

Figure 2-4. Hsp20 expression response profiles in shoots. See Figure 2-3 caption.

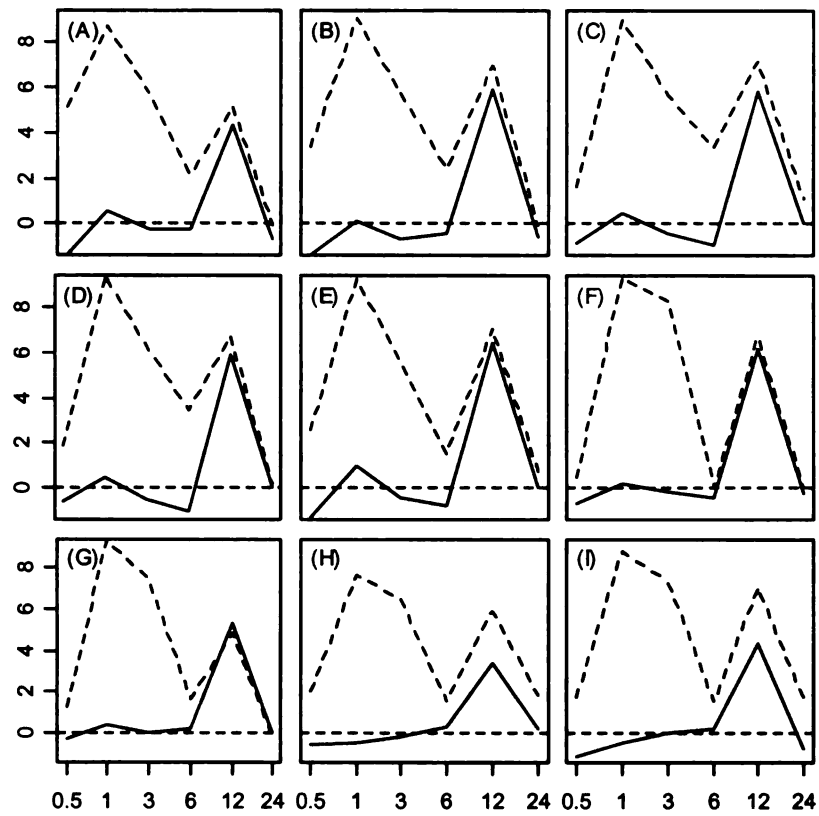

Figure 2-5. Expression response profiles of select Hsp20 genes under wounding and heat stress treatments. Expression response profiles of nine selected Hsp20 proteins under wounding (solid line) and heat (dotted line) treatments are shown. Subplots display response profiles associated with (A) 17.6A-CI, (B) 17.4-CI, (C) 17.6C-CI, (D) 17.6-CII, (E) 17.7-CII, (F) 25.4-P, (G) 23.6-M, (H) 15.7-CI(r), and (I) 26.5-P(r). The horizontal axis corresponds to time points at which gene expression measurements were obtained, while the vertical axis indicates the $\log_2$ fold-change. The dotted horizontal line in each plot indicates a $\log_2$ fold-change of zero (no expression response to stress). For the heat stress treatment, roots were exposed to heat until the 3 hr. time point, such that the 3-24 hr time interval represents a recovery period.

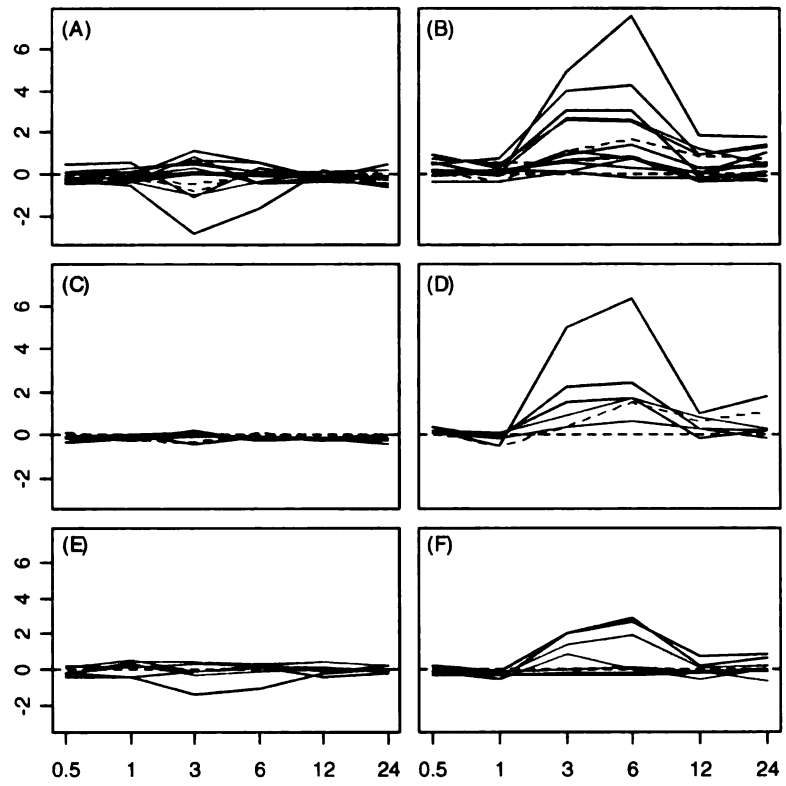

Figure 2-6. Hsp70, Hsp90 and Hsp100 expression response profiles under ultraviolet-B light stress treatment. Expression response profiles associated with all members of the Hsp70 family (A and B), Hsp90 family (C and D), and Hsp100 family (E and F). Profiles associated with root tissue are shown in A, C, and E, while expression response profiles associated with shoot tissue are shown in B, D, and F. The horizontal axis corresponds to time points at which genes expression measurements were obtained, while the vertical axis indicates the $\log_2$ fold-change. The dotted horizontal line in each plot indicates a $\log_2$ fold-change of zero (no expression response to UV-B light). Hsps were localized to the cytoplasm (black lines), plastid (red lines), chloroplast (green line), mitochondria (blue lines), or endoplasmic reticulum (dashed blue line).

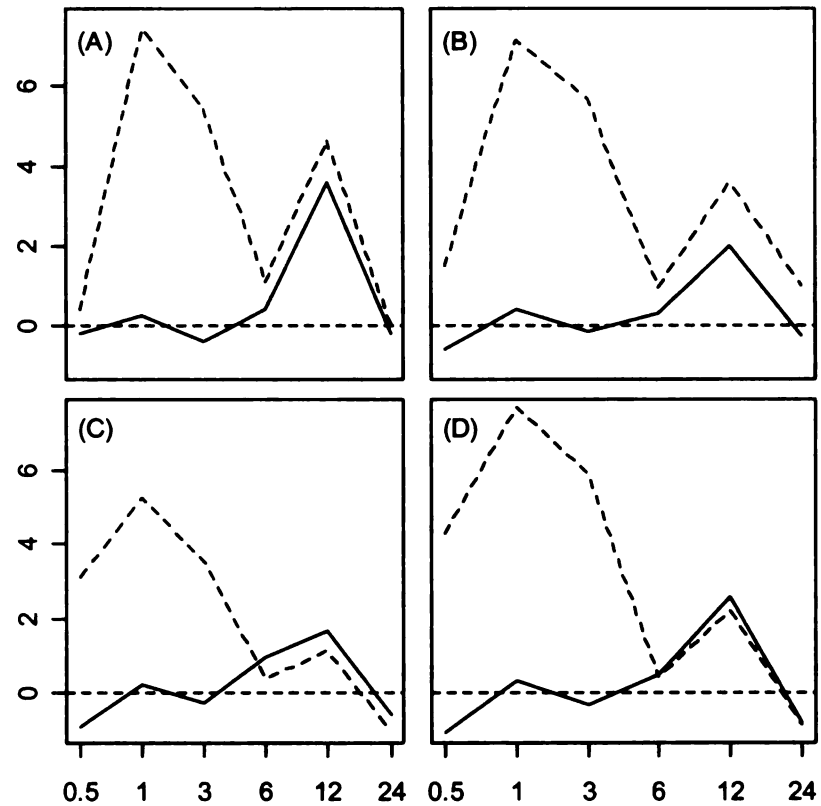

Figure 2-7. Expression response profiles of selected Hsp70, Hsp90 and Hsp100 genes under wounding and heat stress treatments. Expression response profiles of four selected proteins (Hsp70, Hsp90, or Hsp100) under wounding (solid line) and heat (dotted line) treatments are shown. Subplots display response profiles associated with (A) AtHsp70-5, (B) AtHsp70-8, (C) AtHsp90-1, and (D) AtHsp100-1. The horizontal axis corresponds to time points at which gene expression measurements were obtained, while the vertical axis indicates the $\log_2$ fold-change. The dotted horizontal line in each plot indicates a $\log_2$ fold-change of zero (no expression response to stress). For the heat stress treatment, roots were exposed to heat until the 3 hr. time point, such that the 3-24 hr time interval represents a recovery period.

Section 3 Discussion

Heat shock proteins (Hsps) and transcription factors (Hsfs) are central components of the *Arabidopsis thaliana* heat shock regulatory network. It has long been recognized that these elements are also involved in response to cold and non-thermal stress treatments (Feder and Hofmann 1999), but the types of stress that most strongly elicit Hsp/Hsf expression responses have not been identified, and the physiological role of these proteins under non-heat stress treatments is unclear. The findings of this study support the hypothesis that Hsps and Hsfs represent an intersection point between heat and non-heat stress response pathways. Our results indicate that, to varying extents, each of nine cold and non-thermal stress treatments interact with Hsfs and Hsps at the level of gene expression. Several prominent family-level expression response patterns were identified. These included highly similar stress-response profiles among Hsp20 proteins, a number of Hsps specifically upregulated 12 hours after wounding and during recovery following heat stress, and upregulation of heat shock genes to UV-B stress in shoot but not root tissue. Our findings raise important questions to be pursued in future experimental studies of the *Arabidopsis* heat shock response network.

Genome-wide transcriptional profiling allowed the expression of Hsf and Hsp genes under many stress conditions to be examined within the same context. This facilitated identification of which stressors interact with each protein family most strongly, which provides insight into the nature and degree of cross-talk that exists between heat and other forms of stress. The osmotic, cold, and salt treatments were among the strongest inducers of heat shock gene expression. These stress treatments induced expression responses of heat shock genes that were large in comparison to other

Arabidopsis genes (see Table 1), and also large in an absolute sense, since these stressors induced strong fold-changes and differential expression of individual Hsf and Hsp genes (see Figs. 2-1 to 2-4 and supplemental data). Expression response patterns were very similar under each of these treatments, with upregulation primarily occurring over the late stages of stress exposure (3 – 24 hours). Since osmotic, cold, and salt stress treatments are each believed to have a deleterious impact on cellular water potential (Kreps et al. 2002; Verslues et al. 2006), it is possible that their impact on heat shock genes is related to this common effect. In support of this notion, several previous studies in plant species have implicated Hsp20 proteins in tolerance to water stress treatments (Almoguera et al. 1993; Coca et al. 1996; Sun et al. 2001). Among other stress treatments, wounding and UV-B stress induced moderately large expression responses of heat shock genes (with strong differences among families and between tissue types). The pathogen infection treatment was unique, since in contrast to other types of stress, it elicited strong expression responses among the Hsp70, Hsp90, and Hsp100 families, while most members of the Hsf and Hsp20 family were not responsive. Overall, drought and genotoxic stress treatments were associated with weak induction of heat shock genes, although some individual genes can be cited as an exception to this trend (e.g., HsfA8, AtHsp15.4-CI(r), AtHsp100-7).

The degree to which oxidative stress impacted heat shock gene expression is difficult to discern. In comparison to other *Arabidopsis* genes, all protein families (except Hsp90) exhibited large expression responses to oxidative stress (see Table 1). However, since genomic expression responses to oxidative stress were generally small, absolute fold-changes induced by oxidative stress were nonetheless of small magnitude. Only one

transcription factor, for instance, was differentially expressed under oxidative stress (HsfA1d). These results were surprising, since there is considerable evidence that heat shock transcription factors can function as reactive oxygen species sensors in plants (reviewed by Miller and Mittler 2006), and extensive interactions have been identified between heat and oxidative stress molecular pathways (e.g., Härndahl et al. 1999; Panchuk et al. 2002; Pnueli et al. 2003; Davletova et al. 2004; Neta-Sharir et al. 2005). Moreover, since the generation of reactive oxygen species is a general response under many types of stress (Pastori and Foyer 2002), Hsf activation by reactive oxygen species may provide the best hypothesis to explain why heat shock genes are induced by so many stress treatments. In view of this, an important factor to consider is the means by which oxidative stress was experimentally induced. For data we analyzed, oxidative stress was induced by exogenous application of methyl viologen, which is a generator of superoxide anion radical (Laloi et al. 2004). The impact of this reactive oxygen species on heat shock gene expression may differ from that of others (op den Camp et al. 2003; Gadjev et al. 2006). In a recent study, for instance, Gadjev et al. (2006) demonstrated that among genes upregulated more than two-fold under heat stress, relatively few were responsive to superoxide anion radical, while most were instead responsive to the singlet oxygen reactive oxygen species. These considerations suggest that, although the oxidative stress treatment examined by this study may not have had a strong impact on heat shock genes, the production of different types of reactive oxygen species (e.g., H_2O_2), leading to Hsf activation and consequently Hsp expression, remains a pathway through which cellular responses to heat and other forms of stress may be linked.

Heat shock transcription factors are of fundamental importance to understanding stress response networks, since these proteins coordinate the expression of Hsps and other stress-responsive genes. The *Arabidopsis* Hsf family is larger than that which has been described in any other plant or animal system (Nover et al. 2001), and at present, no single Hsf has been identified as a primary trigger of the heat shock response. The emerging picture is one of considerable complexity, with extensive interactions among individual Hsfs and sensitivity to a diverse range of environmental signals (Miller and Mittler 2006). We found that seven Hsfs (six class A, one class B) exhibited very weak expression responses to heat and all other stress conditions (see Table 2, stress-clusters 451 and 452), while the remaining 14 Hsfs were strongly induced by several stress treatments. The most distinctive expression response patterns we observed were associated with HsfA6b and HsfC1 (see supplemental data file 2). In root tissue, HsfA6b exhibited approximately five-fold induction to salt and osmotic treatments across all time points of gene expression measurement (0.5 – 24 hours). This pattern contrasted with that observed among other Hsfs, most of which responded to salt and osmotic stress over the late stages of stress exposure only. This early response of HsfA6b to salt and cold treatments was, in fact, unique among all the heat shock genes that we examined, suggesting that HsfA6b may interact with elements outside of the Hsf/Hsp response pathway. On the basis of differential expression analysis, HsfC1 was the most stress-responsive of all Hsfs. Among all treatments and tissues that we examined, this transcription factor was, on average, differentially expressed with respect to nearly half of the time points at which gene expression was measured. This strong expression response

pattern is particularly noteworthy in light of the large structural dissimilarities between HsfC1 and all other *Arabidopsis* Hsfs (Nover et al. 2001).

The Hsp20 family exhibited the most stress-general expression response pattern of all the protein groups that we examined. Our results therefore suggest that this protein family is of potential importance as a factor contributing to multiple stress tolerance in plant species. These findings are also consistent with those of previous studies, which have found that certain Hsp20 proteins are involved in cellular responses to a wide variety of environmental treatments besides heat, such as alcohol (Kuo et al. 2000), cold (Sabehat et al. 1998), heavy metals (Lin et al. 1984; Tseng et al. 1993; Guan et al. 2004; Sun et al. 2002), osmotic stress (Sun et al. 2001), desiccation (Wehmeyer and Vierling 2000), and oxidative stress (Neta-Sharir et al. 2005). At present, little is known regarding how Hsp20 proteins are integrated with molecular networks that underlie cellular responses to these stress treatments. Increasingly, it has been recognized that Hsp20 proteins can engage in a wide range of cellular processes under stress, including ATP-independent stabilization of substrate proteins undergoing conformational disruption (Sun et al. 2002), or associating with lipid molecules to regulate fluidity of the membrane structure (Tsvetkova et al. 2002). This latter function suggests that Hsp20s could be involved in the perception of stressful stimuli leading to the activation of signal transduction pathways. Under temperature extremes, the role of membrane fluidity as a means of stress perception and activation of signal transduction pathways has been well established (Sun et al. 2002). However, since non-thermal stressors may also alter membrane fluidity or lead to various types of membrane damage, interactions of Hsp20s

with membranes could partly account for the overall stress-responsiveness of the Hsp20 family.

A striking aspect of the Hsp20 family was the similarity among the expression response patterns of member genes. This similarity was demonstrated by our clustering analysis (see section 2C of supplemental data 1), which interestingly, revealed a cluster of five 17 kDa Hsp20 proteins that included both class I and II nuclear/cytosolic proteins. This result is consistent with findings of previous studies, which have identified functional similarities between class I and II Hsp20s (e.g., Löw et al. 2000), despite marked differences between the amino acid sequences of the two classes (Vierling 1991). If analysis is restricted to stress responses occurring in the root tissue type only, the overall homology of Hsp20 expression response patterns is considerably enhanced. In root tissue, expression patterns of 17 kDa Hsp20s are very similar to those of the 18 – 20 kDa Hsp20s, including those localized to the mitochondria and endoplasmic reticulum (see section 2C of supplemental data file 1). The similarity of expression patterns among the Hsp20 proteins may reflect shared induction mechanisms, and possibly extensive coordination among Hsp20s as cellular chaperones, such as that observed during the formation of heat-stress granules (Nover et al. 1983). Shared induction mechanisms among Hsp20 proteins may include accumulation of denatured proteins in the cytoplasm (Sung et al. 2003), generation of reactive oxygen species (Miller and Mittler 2006), or changes in membrane lipid composition and fluidity (Tsvetkova et al. 2002). These processes are thought to be upstream signals leading to the activation of critical Hsfs, which are most likely the direct inducers of Hsp20 expression under stress.

A number of Hsps were upregulated 12 hours after wounding, with a parallel expression response pattern during recovery from heat stress in root tissue. While the majority of these proteins were members of the Hsp20 family (see Fig.2-5), some members of the Hsp70, Hsp90, and Hsp100 families also exhibited this distinctive expression pattern (see Fig. 2-7). The upregulation of multiple Hsps following wounding and during heat stress recovery has not been previously documented in *Arabidopsis* or other plant species, and is therefore an important finding of this study. The first indication that Hsps are involved in the wounding response pathway was provided by the study of Cheong et al. (2002), in which the effect of wounding on expression levels of 8,200 *Arabidopsis* genes was surveyed in leaf tissue. Cheong et al. (2002) identified one Hsf upregulated 0.5 hours following wounding (AtHsfA4a), along with another upregulated both 0.5 and 6 hours after wounding (AtHsfB1). In addition, three Hsp70 proteins were upregulated 6 hours after wounding (HSP70, HSC70-G8, HSC70-G7), as well as two 17 kDa sHSPs (AtHsp17.8-CII and AtHsp17.7-CII). We found that the most interesting wounding-response patterns occurred in root tissue, but our results are consistent with those of Cheong et al. (2002), since Hsp upregulation also occurred after wounding in aerial shoot tissue. Overall, our findings suggest that Hsp involvement in wounding response is greater than previously recognized, and by profiling Hsps simultaneously under multiple types of stress, our results show that late wound-responsive genes are also active during recovery from heat stress. These results point to a broad role of some Hsps during stress recovery or acclimation, i.e., the process by which plants increase stress-tolerance following an initial period of exposure. Following wounding of plant tissue, both local and systemic signals are generated that coordinate defense responses aimed at

limiting further injury (e.g., by pathogen) (Leon et al. 2001). Acclimation to heat stress has been especially well studied, and is characterized by an increased tolerance or hardening to high temperatures following initial exposure (Sung et al. 2003). The functional role of Hsps upregulated as part of the post-wounding and post-heat stress response is unclear and warrants further investigation. With respect to heat stress recovery, one recent study found that mutant plants lacking a 32 kDa heat shock associated protein (Hsa32) exhibited an elevated decay in thermotolerance following exposure to heat stress (Charng et al. 2006).

Ultraviolet-B radiation resulted in upregulation of heat shock proteins and transcription factors in shoots, but did not have this effect in root tissue. This distinction between aerial and subterranean tissue types was most marked with respect to the Hsp20 group, in which nearly all Hsp20s were upregulated in shoots and downregulated in roots. Similar to other stress treatments, exposure to ultraviolet-B light has been associated with the production of reactive oxygen species (Arnott and Murphy 1991; Green and Fluhr 1995). Specifically, ultraviolet light stress has been found to increase cellular concentrations of H_2O_2 (Shiu and Lee 2005), which has been thought to activate Hsf expression (Miller and Mittler 2006), especially that of HsfA4a and HsfA8 (Davletova et al. 2005). We found that both HsfA4a and HsfA8 were strongly induced by UV-B stress in shoots but not in roots (see supplemental data file 2). These observations are consistent with the notion that Hsp expression in shoots results from UV-B induced activation of Hsfs, possibly HsfA4a and HsfA8, with the generation of H_2O_2 as an intermediary signal. Given the tissue-specific effect we observed, however, the generation of H_2O_2 could be dependent upon interactions between UV-B stress and photosynthetic processes taking

place in chloroplast. In previous models, it has been suggested that UV-B generated reactive oxygen species are upstream components that act upon photosynthetic genes (i.e., $H_2O_2 \rightarrow$ photosynthesis) (A.-H.-Mackerness et al. 1999; John et al. 2001). Our results, however, suggest that the reverse is also plausible, in which photosynthetic processes are upstream components leading to the generation of reactive oxygen species under UV-B stress (i.e., photosynthesis $\rightarrow H_2O_2$).

Recently, it has been emphasized that the generation of agricultural varieties tolerant to a range of stress conditions should be a primary goal in biotechnological applications, since under field conditions, plants may encounter different types of stress in combination (Mittler 2006). Focusing on overlapping elements among response pathways that underlie diverse forms of stress may advance our knowledge of cross-tolerance in plant species (Bowler and Fluhr 2000). The *Arabidopsis* heat shock proteins and transcription factors exhibit expression responses under a wide range of stressful stimuli, and are therefore a natural model for developing our understanding of integration between regulatory networks associated with different kinds of stress. The findings of this study have identified which types of stress interact least and most strongly with Hsfs each Hsp family at the transcriptional level. In addition, new family-level expression response patterns related to wounding and ultraviolet-B light stress have been uncovered. These results provide insight into the nature and degree of cross talk between heat and non-heat stress conditions, and represent a basis for further experimental investigations into the involvement of Hsf and Hsp proteins under cold and non-thermal stress.

Section 4 Methods

All microarray data analyzed in this study were generated using the ATH1 Affymetrix microarray platform (Hennig *et al.* 2003; Redman *et al.* 2004), with expression estimates obtained by gcRMA normalization (Wu *et al.* 2004). A total of 22,810 probes were included on the ATH1 platform, along with 64 control probes not corresponding to *Arabidopsis* genes. Our analysis is therefore based on a total of 22,746 genes, representing approximately 80% of all known *Arabidopsis* genes (Schmid *et al.* 2005). Gene expression datasets were downloaded from AtGenExpress at http://www.weigelworld.org/resources/microarray/AtGenExpress/ (abiotic stress and pathogen series). Complete protocols associated with these data can be obtained from TAIR (http://www.arabidopsis.org/) (submission numbers: ME00325, ME00326, ME00327, ME00328, ME00329, ME00330, ME00338, ME00339, ME00340, ME00342). In brief, the abiotic stress series data consists of gene expression measurements performed on *Arabidopsis thaliana* (col-0) roots and shoots under a benign control condition and nine environmental stress conditions. For each stress condition, expression measurements were obtained from 16 to 18-day old plants at six different time points of stress-exposure (1/2, 1, 3, 6, 12, and 24 hours). All expression measurements were performed with duplicate biological replications. Stress treatments included cold (4°), osmotic stress (300 mM Mannitol), salt (150 mM NaCl), drought (15 min. dry air stream leading to 10% loss of fresh weight), genotoxic stress (1.5 μg/ml bleomycin, 22 μg/ml mitomycin), oxidative stress (10 μM methyl viologen), ultraviolet-B light stress (15 min. exposure, 1.18 W/m^2 Phillips TL40W/12), wounding (pin puncture), and high temperature (3 hrs. at 38° followed by 21 hrs. recovery at 25°). From the pathogen series

dataset, we considered experiments involving *P. infestans* infection of 5-week old *Arabidopsis* leaves, along with corresponding control treatments in which H_2O was applied to leaves. Expression measurements were obtained from three biological replications at each of three post-infection time points (6, 12, and 24 hours). Pathogen infections used 10^{-8} cfu/ml in $MgCl_2$ with 5 x 10^5 *P. infestans* spores applied to leaf surfaces.

The heat shock proteins and transcription factors analyzed in this study were selected based upon the genomic sequence analyses performed by Agarwal et al. (2001), Lin et al. (2001), Nover et al. (2001), Krishna and Gloor (2001), and Scharf et al. (2001). Our analysis includes all of the twenty-one Hsfs identified by Nover et al. (2001). Several Hsps identified by the above-cited studies were not represented on the ATH1 array (AtHsp17.8-Cl, AtHsp70-12, AtHsp70-13, AtHsp70-14, AtHsp70-16, AtHsp70-18, AtHsp90-3, and AtHsp100-6), and therefore could not be included in this study. In total, our heat shock protein analysis is based upon 18 of 19 members of the Hsp20 family (12 sHsps and 6 related sHsp-like proteins), 13 of 17 members of the Hsp70 family (11 DnaK and 2 SSE subfamily), 6 of 7 members of the Hsp90 family, and 7 of 8 members of the Hsp100 family (AtHsp100-1 and six homologues). The expression response patterns of each Hsf and Hsp gene was analyzed with respect to nine abiotic stress treatments (applied to root and shoot tissue), in addition to pathogen infection treatment (applied to leaf tissue). In total, therefore, the expression response of each Hsf and Hsp was examined under 19 tissue-treatment combinations.

The T statistic represents the median level of fold-change induced by a given stress treatment among members of a given protein family. Let $\overline{X}_{ijkl}$ represent the mean

gcRMA normalized expression intensity of the *i*th gene under the *j*th experimental treatment (abiotic stress or pathogen) within the *k*th tissue following *t* hours of stress exposure ($i = 1 \ldots N, j = 1 \ldots 10, k = 1 \ldots 3$, and $t = 0.5 \ldots 24$). For every tissue-treatment-time combination, values of $\overline{X}_{ijkt}$ were associated with a corresponding control measurement designated as $\overline{X}_{i0kt}$ ($j = 0$). Log_2 fold-changes (M) at each tissue-treatment-time combination were thus calculated as the difference between expression intensities in the *j*th stress treatment and corresponding control treatments.

$$M_{ijkt} = \overline{X}_{ijkt} - \overline{X}_{i0kt} \qquad (1)$$

The average value of $|M|$ occurring over all time points under a given tissue-treatment combination reflects the overall stress-responsiveness associated with a gene's expression profile. Letting this average value for gene *i* under treatment *j* in tissue *k* be represented by $|\overline{M}_{ijk\cdot}|$, a test statistic T was defined as the median value of $|\overline{M}_{ijk\cdot}|$ among the *n* gene members of a given protein family.

$$T_{jk} = \underset{i=1\ldots n}{\text{median}}(|\overline{M}_{ijk\cdot}|) \qquad (2)$$

The magnitude of T reflects how large expression responses of a protein family are, on average, with respect to a given tissue-treatment combination. The significance of observed T statistics was evaluated under the null hypothesis that the *n* genes in each protein family are a random sample of the $N = 22746$ genes represented on the ATH1

array, versus the alternative that the n genes are a non-random sample yielding a T statistic larger than expected within a random sample. This hypothesis was evaluated by the following resampling procedure. With respect to each tissue-treatment combination and each protein family, a total of 10^3 random samples of n genes were drawn from among all $N = 22746$ genes, and the value of T was calculated from each of the 10^3 random samples. This yielded null distributions specific to each tissue-treatment combination and protein family, which were used to evaluate the significance of observed T statistics. An observed T statistic was significant if the proportion of random samples yielding a larger or equal T statistic was less than $\alpha = 0.05$. A significant T statistic indicates that the expression responses among the n members of a protein family (with respect to a given tissue-treatment combination) are larger than expected within a random sample of n genes.

Hsf and Hsp expression response patterns within protein families and among individual genes were analyzed by differential expression analysis and clustering. Differential expression analysis was carried out using the Limma linear modeling package available in the R Bioconductor software suite (Smyth 2004). In this approach, a linear model was fit for all genes with respect to each of the 19 tissue-treatment combinations. This allowed heat shock related genes to be tested for differential expression at every time point associated with each tissue-treatment combination. For each of the 19 linear model analyses performed, P-values were adjusted for multiple comparisons using the Benjamini and Hochberg method (Benjamini & Hochberg 1995; Reiner et al. 2003). The differential expression analysis was used to construct the index (d) introduced in Results section.

The hierarchical ordered partitioning and collapsing hybrid (HOPACH) clustering algorithm was used to identify sub-groups of genes with similar expression response patterns in each protein family (van der Laan and Pollard 2003). In this algorithm, the number of clusters appropriate in the final clustering solution is determined automatically according the median split silhouette criterion (Pollard and van der Laan 2005). The HOPACH algorithm is particularly well-suited for finding homogenous clusters of small size among a limited number of genes. Stress-clusters were formed by grouping Hsf/Hsp genes with respect to their expression responses (M) under all 111 tissue-treatment-time combinations included in our analysis (18 tissue-treatment combinations with 6 time points + 1 tissue-treatment combination with 3 time points). The Euclidean distance metric was used to measure similarity between vectors of expression responses (M) associated with each Hsf/Hsp gene. To form developmental-clusters, genes were centered to have a mean expression intensity of zero across the 79 developmental series conditions, and the cosine angle similarity metric was used to cluster expression profiles of Hsf/Hsp genes within each family.

Bibliography

A.-H.-Mackerness, S., S. L. Surplus, P. Blake, C. F. John, V. Buchanan-Wollaston, B. R. Jordan, and B. Thomas. 1999. Ultraviolet-B-induced stress and changes in gene expression in *Arabidopsis thaliana*: role of signaling pathways controlled by jasmonic acid, ethylene and reactive oxygen species. Plant Cell Environ. 22: 1413-1423.

Agarwal, M., S. Katiyar-Agarwal, C. Sahi, D. R. Gallie, and A. Grover. 2001. *Arabidopsis thaliana* Hsp100 proteins: kith and kin. Cell Stress Chap. 6: 219-224.

Almoguera, C., M. A. Coca, and J. Jordano. 1993. Tissue-specific expression of sunflower heat-shock proteins in response to water-stress. Plant J. 4: 947-958.

Arnott, T., and T. M. Murphy. 1991. A comparison of the effects of a fungal elicitor and ultraviolet-radiation on ion-transport and hydrogen-peroxide synthesis by rose cells. Env. Exp. Bot. 31: 209-216.

Banzet, N., C. Richaud, Y. Deveaux, M. Kazmaier, J. Gagnon, and C. Triantaphylides. 1998. Accumulation of small heat shock proteins, including mitochondrial Hsp22, induced by oxidative stress and adaptive response in tomato cells. Plant J. 13: 519-527.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a powerful and practical approach to multiple testing. J. Roy. Stat. Soc. B. 57: 289-300.

Bowler, C., and R. Fluhr. 2000. The role of calcium and activated oxygens as signals for controlling cross-tolerance. Trends Plant Sci. 5: 241-246.

Charng, Y. Y., H. C. Liu, N. Y. Liu, F. C. Hsu, and S. S. Ko. 2006. Arabidopsis Hsa32, a novel heat shock protein, is essential for acquired thermotolerance during long recovery after acclimation. Plant Physiol. 140: 1297-1305.

Cheong, Y., H. Chang, R. Gupta, X. Wang, T. Zhu and S. Luan. 2002 Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal response in *Arabidopsis*. Plant Phys. 129: 661-677.

Coca, M. A., C. Almoguera, T. L. Thomas, and J. Jordano. 1996. Differential regulation of small heat-shock genes in plants: analysis of a water-stress-inducible and developmentally activated sunflower promoter. Plant Mol. Biol. 31: 863-876.

Davletova S., K. Schlauch, J. Coutu, and R. Mittler. 2005. The zinc-finger protein Zat12 plays a central role in reactive oxygen and abiotic stress signaling in *Arabidopsis*. Plant Phys. 139: 847-856.

Desikan, R., S. A.-H.-Mackerness, J. T. Hancock, and S. J. Neill. 2001. Regulation of the *Arabidopsis* transcriptome by oxidative stress. Plant Phys. 127: 159-172.

Ding, Q. X., and J. N. Keller. 2001. Proteasome inhibition in oxidative stress neurotoxicity: implications for heat shock proteins. J. Neurochem. 77: 1010-1017.

Feder, M. E., and G. E. Hofmann. 1999. Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. Ann. Rev. Physiol. 61: 243-282.

Fleming, J. E., I. Reveillaud, and A. Niedzwiecki. 1992. Role of oxidative stress in *Drosophila* aging. Mutat. Res. 275: 267-279.

Gadjev, I., S. Vanderauwera, T. S. Gechev, C. Laloi, I. N. Minkov, V. Shulaey, K. Apel, D. Inze, R. Mittler and F. Van Breusegem. 2006. Transcriptomic footprints disclose specificity of reactive oxygen species signaling in *Arabidopsis*. Plant Phys. 141: 436-445.

Green, R., and R. Fluhr. 1995. UV-B-induced PR-1 accumulation is mediated by active oxygen species. Plant Cell 7: 203-212.

Guan, J. C., T. L. Jinn, C. H. Yeh, S. P. Feng, Y. M. Chen, and C. Y. Lin. 2004. Characterization of the genomic structures and selective expression profiles of nine class I small heat shock protein genes clustered on two chromosomes in rice (Oryza sativa L.). Plant Mol. Biol. 56: 795-809.

Györgyey, J., A. Gartner, K. Nemeth, Z. Magyar, H. Hirt, E. Heberlebors, and D. Dudits. 1991. Plant Mol. Biol. 16: 999-1007.

Härndahl, U., R. B. Hall, K. W. Osteryoung, E. Vierling, J. F. Bornman, and C. Sundby. 1999. The chloroplast small heat shock protein undergoes oxidation-dependent conformational changes and may protect plants from oxidative stress. Cell Stress Chap. 4: 129-138.

Hennig, L., M. Menges, J. A. Murray and W. Gruissen, 2003 *Arabidopsis* transcript profiling on Affymetrix GeneChip arrays. Plant Mol. Biol. 53: 457-465.

Hihara, Y., A. Kamei, M. Kanehisa, A. Kaplan, and M. Ikeuchi. 2001. DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. Plant Cell 13: 793-806.

Hoffmann, A. A., and P. A. Parsons. 1991 Evolutionary genetics and environmental stress. Oxford University Press, Oxford.

John, C. F., K. Morris, B. R. Jordan, B. Thomas, and S. A.-H.-Mackerness. 2001. Ultraviolet-B exposure leads to up-regulation of senescence-associated genes in *Arabidopsis thaliana*. J. Exp. Bot. 52: 1367-1373.

Krishna, P., and G. Gloor. 2001. The Hsp90 family of proteins in *Arabidopsis thaliana*. Cell Stress Chap. 6: 238-246.

Kreps, J. A., Y. Wu, C. Hur-Song, T. Zhu, X. Wang, and J. F. Harper. 2002 Transcriptomic changes for *Arabidopsis* in response to salt, osmotic, and cold stress. Plant Phys. 130: 2129-2141.

Kregel, K. C. 2002. Heat shock proteins: modifying factors in physiological stress responses and acquired thermotolerance. J. Appl. Physiol. 92: 2177-2186.

Kuo, H. F., Y. F. Tsai, L. S. Young, and C. Y. Lin. 2000. Ethanol treatment triggers a heat shock-like response but no thermotolerance in soybean (Glycine max cv. Kaohsiung No.8) seedlings. Plant Cell Environ. 23: 1099-1108.

Laloi, C., K. Apel, and A. Danon. 2004. Reactive oxygen signaling: the latest news. Curr. Opin. Plant Biol. 7: 323-328.

Larkindale, J., J. D. Hall, M. R. Knight, and E. Vierling. 2005. Heat stress phenotypes of *Arabidopsis* mutants implicate multiple signaling pathways in the acquisition of thermotolerance. Plant Phys. 138: 882-897.

Lee, G. J., A. M. Roseman, H. R. Saibil, and E. Vierling. 1997. A small heat shock protein stably binds heat-denatured model substrates and can maintain a substrate in a folding-competent state. Embo J. 16: 659-671.

Lee, G. J., and E. Vierling. 2000. A small heat shock protein cooperates with heat shock protein 70 systems to reactivate a heat-denatured protein. Plant Phys. 122: 189-197.

Lee, B. H., S. H. Won, H. S. Lee, M. Miyao, W. I. Chung, I. J. Kim, and J. Jo. 2000. Expression of the chloroplast-localized small heat shock protein by oxidative stress in rice. Gene 245: 283-290.

Leon, J., E. Rojo, and J.J. Sanchez-Serrano. 2001. Wound signalling in plants. J. Exp. Bot. 52: 1-9.

Lin, C. Y., J. K. Roberts, and J. L. Key. 1984. Acquisition of thermotolerance in soybean seedlings: synthesis and accumulation of heat shock proteins and their cellular localization. Plant Physiol. 74: 152-160.

Lin, B.-L., J.-S. Wang, H.-C. Liu, R.-W. Chen, Y. Meyer, A. Barakat, and M. Delseny. 2001. Genomic analysis of the Hsp70 superfamily in *Arabidopsis thaliana*. Cell Stress Chap. 6: 201-208.

Liu, N. Y., S. S. Ko, K. C. Yeh, and Y. Y. Charng. 2006a. Isolation and characterization of tomato Hsa32 encoding a novel heat-shock protein. Plant Science 170: 976-985.

Liu, D., X. Zhang, Y. Cheng, T. Takano, and S. Liu. 2006b. rHsp90 gene expression in response to several environmental stresses in rice (Oryza sativa L.). Plant Physiol. Biochem 44: 380-386.

Löw, D., K. Brändle, L. Nover, and C. Forreiter. 2000. Cytosolic heat-stress proteins Hsp17.7 class I and Hsp17.3 class II of tomato act as molecular chaperones in vivo. Planta 211: 575-582.

Ma, S., Q. Gong, and H. J. Bohnert, 2006 Dissecting salt stress pathways. J. Exp. Bot. 57: 1097-1107.

Martindale, J. L., and N. J. Holbrook. 2002. Cellular response to oxidative stress: signaling for suicide and survival. J. Cell. Phys. 192: 1-15.

Mehlen, P., J. Briolay, L. Smith, C. Diazlatoud, N. Fabre, D. Pauli, and A. P. Arrigo. 1993. Analysis of the resistance to heat and hydrogen peroxide stresses in cos cells transiently expressing wild-type or deletion mutants of the *Drosophila* 27-kda heat-shock protein. Europ. J. Biochem. 215: 277-284.

Miller, G., and R. Mittler. 2006. Could heat shock transcription factors function as hydrogen peroxide sensors in plants? Ann. Bot. 98: 279-288.

Mittler, R. 2006. Abiotic stress, the field environment and stress combination. Trends Plant Sci. 11: 15-19.

Neta-Sharir, I., T. Isaacson, S. Lurie, and D. Weiss. 2005. Dual role for tomato heat shock protein 21: protecting photosystem II from oxidative stress and promoting color changes during fruit maturation. Plant Cell 17: 1829-1838.

Nover, L., K. D. Scharf, and D. Neumann. 1983. Formation of cytoplasmic heat-shock granules in tomato cell-cultures and leaves. Mol. Cell. Biol. 3: 1648-1655.

Nover, L., K. Bharti, P. Döring, S. K. Mishra, A. Ganguli, and K.-D. Scharf. 2001. *Arabidopsis* and the heat stress transcription factor world: how many heat stress transcription factors do we need? Cell Stress Chap. 6: 177-189.

Op den Camp, R. G. L., D. Przybyla, C. Ochsenbein, C. Laloi, C. Kim, A. Danon, D. Wagner, E. Hideg, C. Göbel, I. Feussner, M. Nater, and K. Apel. 2003. Rapid induction of distinct stress responses after release of singlet oxygen in *Arabidopsis*. Plant Cell 15: 2320-2332.

Panchuk, I. I., R. A. Volkov, and F. Schöffl. 2002. Heat stress and heat shock transcription factor dependent expression and activity of ascorbate peroxidase in *Arabidopsis*. Plant Phys. 129: 838-853.

Pastori, G. M., and H. Foyer. 2002 Common components, networks, and pathways of cross-tolerance to stress. The central role of "redox" and abscisic acid-mediated controls. Plant Phys. 129: 460-468.

Pnueli, L., H. Liang, M. Rozenberg, and R. Mittler. 2003. Growth suspension, altered stomatal responses, and augmented induction of heat shock proteins in cytosolic ascorbate peroxidase (Apx1)-deficient *Arabidopsis* plants. Plant J 34: 187-203.

Pollard, K. and M. van der Laan. 2005. Cluster analysis of genomic data. Pp. 209-228 *in* R. Gentleman, W. Huber, V. J. Carey, R. A. Irizarry, and S. Dudoit, eds. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York.

Preville, X., F. Salvemini, S. Giraud, S. Chaufour, C. Paul, G. Stepien, M. V. Ursini, and A. P. Arrigo. 1999. Mammalian small stress proteins protect against oxidative stress through their ability to increase glucose-6-phosphate dehydrogenase activity and by maintaining optimal cellular detoxifying machinery. Exp. Cell Res. 247: 61-78.

Redman, J. C., B. J. Haas, G. Tanimoto, and C. D. Town. 2004 Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. Plant J. 38: 545-561.

Reiner, A., D. Yekutieli, and Y. Benjamini. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinform. 19: 368-375.

Rensink, W. A., S. Lobst, A. Hart, S. Stegalkina, J. Liu, and C. R. Buell. 2005. Gene expression profiling of potato responses to cold, heat, and salt stress. Funct. Integ. Genom. 5: 201-207.

Rossel, J. B., I. W. Wilson, and B. J. Pogson. 2002. Global changes in gene expression in response to high light in *Arabidopsis*. Plant Phys. 130: 1109-1120.

Rossel, J. B., P. B. Walter, L. Hendrickson, W. S. Chow, A. Poole, P. M. Mullineaux, and B. J. Pogson. 2006. A mutation affecting ascorbate peroxidase 2 gene expression reveals a link between responses to high light and drought tolerance. Plant Cell Environ. 29: 269-281.

Sabehat, A., S. Lurie, and D. Weiss. 1998. Expression of small heat-shock proteins at low temperatures – A possible role in protecting against chilling injuries. Plant Phys. 117: 651-658.

Scharf, K.-D., M. Siddique, and E. Vierling. 2001. The expanding family of *Arabidopsis thaliana* small heat stress proteins and a new family of proteins containing α-crystallin domains (Acd proteins). Cell Stress Chap. 6: 225-237.

Schmid, M., T. S. Davison, S.R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, and J. U. Lohmann. 2005. A gene expression map of Arabidopsis thaliana development. Nat. Genet. 37: 501-506.

Schöffl, F., R. Prandl, and A. Reindl. 1998. Regulation of the heat-shock response. Plant Phys. 117: 1135-1141.

Shiu, C. T., and T. M. Lee. 2005. Ultraviolet-B-induced oxidative stress and responses of the ascorbate-glutathione cycle in a marine macroalga Ulva fasciata. J. Exp. Bot. 56: 2851-2865.

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

Sun, W. N., C. Bernard, B. van de Cotte, M. Van Montagu, and N. Verbruggen. 2001. At-Hsp17.6A, encoding a small heat-shock protein in *Arabidopsis*, can enhance osmotolerance upon overexpression. Plant J. 27: 407-415.

Sun, W. N., M. Van Montagu, and N. Verbruggen. 2002. Small heat shock proteins and stress tolerance in plants. Gene Struct. Expr. 1577: 1-9.

Sung, D. Y., F. Kaplan, K.-J. Lee, and G. L. Guy. 2003. Acquired tolerance to temperature extremes. Trends Plant Sci 8: 179-187.

Tseng, T. S., S. S. Tzeng, K. W. Yeh, C. H. Yeh, F. C. Chang, Y. M. Chen, and C. Y. Lin. 1993. The heat-shock response in rice seedlings – isolation and expression of cDNAs that encode class-I low-molecular-weight heat-shock protein. Plant Cell Phys. 34: 165-168.

Tsvetkova, N. M., I. Horvath, Z. Torok, W. F. Wolkers, Z. Balogi, N. Shigapova, L. M. Crowe, F. Tablin, E. Vierling, J. H. Crowe, and L. Vigh. 2002. Small heat-shock proteins regulate membrane lipid polymorphism. Proc. Natl. Acad. Sci. 99: 13504-13509.

Van der Lann, M., and K. Pollard. 2003. Hybrid clustering of gene expression data with visualization and bootstrap. J. Stat. Plan. Inf. 117: 275-303.

Verslues, P. E., M. Agarwal, S. Katiyar-Agarway, J. H. Zhu, and J. K. Zhu. 2006. Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status. Plant J. 45: 523-539.

Vierling, E. 1991. The roles of heat-shock proteins in plants. Ann. Rev. Plant Phys. Mol. Biol. 42: 579-620.

Volkov, R. A., I. I. Panchuk, P. M. Mullineaux, and F. Schöffl. 2006. Heat stress-induced H_2O_2 is required for effective expression of heat shock genes in *Arabidopsis*. Plant Mol. Biol. 61: 733-746.

Wehmeyer, N., and E. Vierling. 2000. The expression of small heat shock proteins in seeds responds to discrete developmental signals and suggests a general protective role in desiccation tolerance. Plant Phys. 122: 1099-1108.

Wheeler, J. C., E. T. Bieschke, and J. Tower. 1995. Muscle-specific expression of *Drosophila* hsp70 in response to aging and oxidative stress. Proc. Natl. Acad. Sci. 92: 10408-10412.

Wu, Z., R. Irizarry, R. Gentleman, F. Martinez Murillo, and F. Spencer. 2004. A model based background adjustment for oligonucleotide expression arrays. J. Amer. Stat. 99: 909-917.

Appendix

The following R code was used to carry out differential expression analyses using the LIMMA linear modeling package. The matrix "X" is a 22746 x 24 matrix that contains all expression data associated with a single stress-tissue combination (stress and control treatments, 6 time points, 2 replicates per time point).

```
> library(limma)
> X = read.table("ExpressionMatrix.txt")
> eset = X
> design = model.matrix(~ -1+factor(c(1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10,
10, 11, 11, 12, 12)))
> colnames(design) =
c("control1","control2","control3","control4","control5","control6","time1",
"time2","time3","time4","time5","time6")
> fit = lmFit(eset, design)
> contrast.matrix = makeContrasts(time1-control1, time2-control2, time3-control3,
time4-control4, time5-control5, time6-control6,levels=design)
> fit2 = contrasts.fit(fit, contrast.matrix)
> fit2 = eBayes(fit2)
> result = decideTests(fit2, method = "separate", adjust.method = "BH", p.value=0.05)
> result1 = abs(result[,1])
> result2 = abs(result[,2])
> result3 = abs(result[,3])
> result4 = abs(result[,4])
> result5 = abs(result[,5])
> result6 = abs(result[,6])
```

The following is an example of the R-Code used to construct Figures 2-1 to 2-4 of Chapter 2. The matrices X1, X2, …, X9 have six rows each (one for each time point at which measurements were obtained). The column number of each matrix corresponds to the number of genes represented in a given figure (i.e., 21 in Figures 2-1 and 2-2, 18 in Figures 2-3 and 2-4).

```
> par(mfrow = c(3,3))
> linetypes = rep(1,21)
> thecolors = c(rep(1,15),rep(2,5),4)
>
> par(mai = c(0.05, 0.27, 0.10, 0))
> X1 = data.frame(X1)
> row.names(X1) = c("0.5","1","3","6","12","24")
> matplot(X1, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> axis(2)
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(A)", pos=4)
>
> par(mai = c(0.05, 0.10, 0.10, 0.05))
> X2 = data.frame(X2)
> row.names(X2) = c("0.5","1","3","6","12","24")
> matplot(X2, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(B)", pos=4)
>
> par(mai = c(0.05, 0.05, 0.10, 0.10))
> X3 = data.frame(X3)
> row.names(X3) = c("0.5","1","3","6","12","24")
> matplot(X3, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(C)", pos=4)
>
> par(mai = c(0.10, 0.25, 0.05, 0))
> X4 = data.frame(X4)
> row.names(X4) = c("0.5","1","3","6","12","24")
> matplot(X4, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> axis(2)
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(D)", pos=4)
>
> par(mai = c(0.10, 0.10, 0.05, 0.05))
> X5 = data.frame(X5)
```

```
> row.names(X5) = c("0.5","1","3","6","12","24")
> matplot(X5, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(E)", pos=4)
>
> par(mai = c(0.10, 0.05, 0.05, 0.10))
> X6 = data.frame(X6)
> row.names(X6) = c("0.5","1","3","6","12","24")
> matplot(X6, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(F)", pos=4)
>
> par(mai = c(0.27, 0.27, 0, 0))
> X7 = data.frame(X7)
> row.names(X7) = c("0.5","1","3","6","12","24")
> matplot(X7, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> axis(2)
> axis(1,1:6,row.names(X7))
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(G)", pos=4)
>
> par(mai = c(0.27, 0.10, 0, 0.05))
> X8 = data.frame(X8)
> row.names(X8) = c("0.5","1","3","6","12","24")
> matplot(X8, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> axis(1,1:6,row.names(X8))
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(H)", pos=4)
>
> par(mai = c(0.27, 0.05, 0, 0.10))
> X9 = data.frame(X9)
> row.names(X9) = c("0.5","1","3","6","12","24")
> matplot(X9, axes=F, frame=T, lty=linetypes, col=thecolors, type="l",
xlab="Time(hrs.)",ylab="M", ylim=c(-2,9))
> axis(1,1:6,row.names(X9))
> abline(h = 0, lty = 2)
> text(0.75, 8.6, labels="(I)", pos=4)
```