A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R

W. Gregory Alvord, Jean A. Roayaei, Octavio A. Quiñones and Katherine T. Schneider

Submitted: 1st May 2007; Received (in revised form): 31st August 2007

Abstract

This article describes specific procedures for conducting quality assessment of Affymetrix GeneChip® soybean genome data and for performing analyses to determine differential gene expression using the open-source R programming environment in conjunction with the open-source Bioconductor software. We describe procedures for extracting those Affymetrix probe set IDs related specifically to the soybean genome on the Affymetrix soybean chip and demonstrate the use of exploratory plots including images of raw probe-level data, boxplots, density plots and M versus A plots. RNA degradation and recommended procedures from Affymetrix for quality control are discussed. An appropriate probe-level model provides an excellent quality assessment tool. To demonstrate this, we discuss and display chip pseudo-images of weights, residuals and signed residuals and additional probe-level modeling plots that may be used to identify aberrant chips. The Robust Multichip Averaging (RMA) procedure was used for background correction, normalization and summarization of the AffyBatch probe-level data to obtain expression level data and to discover differentially expressed genes. Examples of boxplots and MA plots are presented for the expression level data. Volcano plots and heatmaps are used to demonstrate the use of (log) fold changes in conjunction with ordinary and moderated t-statistics for determining interesting genes. We show, with real data, how implementation of functions in R and Bioconductor successfully identified differentially expressed genes that may play a role in soybean resistance to a fungal pathogen, Phakopsora pachyrhizi. Complete source code for performing all quality assessment and statistical procedures may be downloaded from our web source: http://css.ncifcrf.gov/services/download/MicroarraySoybean.zip.

Keywords: microarray; Affymetrix GeneChip®; probe-level data preprocessing; quality control; differential gene expression analysis

Corresponding author. W. Gregory Alvord, PhD, Director, Statistical Consulting Services, Data Management Services, Inc. (DMS), National Cancer Institute at Frederick (NCI-Frederick), PO Box B, Frederick, MD 21702-1201, USA. Tel: 301.846.5101; Fax: 301.846.6196; E-mail: gwa@css.ncifcrf.gov

W. Gregory Alvord is Director of the Statistical Consulting Services Department at the NCI—Frederick campus. He received his PhD in Statistics in 1983 from the University of Maryland. His research interests include applications of statistical methodology to diverse areas of biomedical research; current interests include bioinformatics, microarray gene expression analysis, clinical proteomics and computational biology.

Jean Roayaei gained his PhD in Econometrics and Statistical Methods in 1984 from Florida State University. He finished his Postdoctoral degree in Biostatistics at UNC-Chapel Hill, School of Public Health, in 2000. He is currently at DMS, NCI—Frederick. His main areas of research are microarray gene expression, clinical proteomics and statistical genetics.

Octavio A. Quiñones gained his MSPH in Biostatistics from the College of Public Health at the University of South Florida in 2000. He is currently a Statistician at DMS, NCI—Frederick. His main interests include programming of computer algorithms in numerous statistical languages, clinical proteomics and microarray data analysis.

Katherine T. Schneider gained her PhD in 2002 from the Genetics Department of North Carolina State University. Dr Schneider is currently a Research Molecular Biologist with the USDA-ARS Foreign Disease-Weed Science Research Unit at Ft Detrick, MD, USA. Her main interests are plant resistance to disease and molecular crop improvement.
INTRODUCTION
This article describes specific procedures for conducting quality assessment of Affymetrix GeneChip® Soybean Genome data and for performing analyses to determine differential gene expression using the R programming environment [1] and Bioconductor software [2]. We begin by describing procedures necessary for extracting those Affymetrix probe set IDs specifically related to the soybean genome on the Affymetrix Soybean GeneChip®. The use of exploratory plots including images of raw probe-level data, boxplots, density plots and M versus A plots, followed by a discussion of RNA degradation and recommended procedures from Affymetrix for quality control are then presented. We next describe how an appropriate probe-level model provides further excellent quality assessment tools. Chip pseudo-images of weights, residuals and signed residuals, relative log expression plots and normalized unscaled standard error plots that may be used to identify aberrant chips are discussed and displayed. Following the discussion of quality assessment of the chips, we present our analyses to discover differentially expressed genes. We briefly review the Robust Multichip Averaging (RMA) procedure for background correction, normalization and summarization of the AffyBatch probe-level data to obtain expression level data [3, 4]. Boxplots and MA plots for the expression level data are then described. The use of (log) fold changes in conjunction with both ordinary and moderated t-statistics for determining the best summary statistic for ranking interesting genes is explained. We demonstrate these procedures with volcano plots and a heatmap. All analyses are performed using the open-source R programming environment (version 2.5.1) in conjunction with the open-source Bioconductor software (version 1.9.9). Complete source code and related R objects/workspaces for performing all quality assessment and statistical procedures may be downloaded from our website: http://css.ncifcrf.gov/services/download/Microarray Soybean.zip, from the document entitled ‘Soybean R 2-5-1 Code Chunks.doc’. The data discussed in this article have been deposited in NCBI’s Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession number GSE8432. Readers interested in accessing the specific .CEL files used in this analysis may go to the NCBI GEO site, or to our website indicated above. Executing the R code produces Figures 2 through 10 in color. All Figures, 1 through 10, are available in color in the PowerPoint presentation entitled ‘Soybean Paper.ppt’.

APPLICATION TO THE SOYBEAN GENOME
The fungus Phakopsora pachyrhizi causes a devastating disease in soybeans called soybean rust [5–7]. A microarray experiment was conducted to identify genes that may play a role in soybean resistance to P. pachyrhizi. There are many isolates of P. pachyrhizi, named by their geographic origin and year of discovery. This study used the Hawaii 94-1 and Taiwan 80-2 isolates because they produce distinct physiological responses in the host. The interaction of the soybean with the fungus may produce either a resistant or a susceptible reaction. In our case, the interaction of a specific soybean cultivar with isolate Hawaii 94-1 produces the resistant reaction characterized by the lack of visible symptoms (Figure 1A). The interaction of the same soybean cultivar with rust isolate Taiwan 80-2 produces a susceptible reaction (Figure 1B). The susceptible reaction is characterized by tan lesions, yellowing of the leaves, and the production of numerous spores. The soybean genotype in each reaction is identical; hence, identification of over- or under-expression in specific soybean genes may help in identifying which genes may play a role in the production of the resistant reaction. For this analysis we discuss data collected at 12 h after the fungal infection begins.

Preprocessing of RNA samples
The quality of our RNA samples was verified at three steps during the sample preparation process: after extraction, after labeling, and after fragmentation of the sample. The samples were tested for purity and integrity using an Agilent 2100 BioAnalyzer (Agilent, Santa Clara, CA, USA). We verified that the starting samples were free from contaminating proteins, that the quantity of biotin-labeled cRNA produced was adequate and that the samples were sufficiently fragmented before hybridization. We also verified the presence and expected concentrations of four hybridization control RNAs that are provided at specific, known concentrations by Affymetrix.

Quality assessment
Importing and accessing probe-level data
The Affymetrix GeneChip® Soybean Array was designed in collaboration with the Soybean Research
Community as part of the GeneChip® Consortia Program [8]. Affymetrix GeneChip arrays use short oligonucleotides to probe for genes in an RNA sample. The Soybean Array uses 11 probe pairs, referred to as a probeset for each gene [8]. One component of these probe pairs is referred to as the perfect match probe (PM) and is designed to hybridize only with transcripts from the intended gene, i.e. to measure specific hybridization. However, hybridization to the PM probes by other mRNA species cannot be avoided. Therefore, Affymetrix provides another component of a probe pair, referred to as the mismatch probe (MM) that is intended to measure nonspecific hybridization. Affymetrix’s strategy is to make MM probes identical to their perfect match counterpart except that the middle base is exchanged with its complement [9, 10].

The Bioconductor software is designed to read the .CEL files that contain data on the intensity at each probe on the GeneChip, along with other values. To import the .CEL file data into the Bioconductor software, one uses the ReadAffy function. The following abbreviated R code chunk accomplishes this.

```r
> library('affy')
> scy.ab <- ReadAffy('<pathname>/A3.CEL', ...
> new.sampleNames <- c('hr.a3.12',
>                       'hr.b3.12', 'hr.c3.12',
>                       + 'ts.a4.12', 'ts.b4.12', 'ts.c4.12')
> sampleNames(scy.ab) <- new.sampleNames
```

(Complete source code to read in the .CEL files and other commands may be downloaded from our website: http://css.ncifcrf.gov/services/download/MicroarraySoybean.zip, from the document entitled ‘Soybean R 2-5-1 Code Chunks.doc’.)

The scy.ab object contains the probe-level data and related phenotypic information and is an object of class AffyBatch [11]. For example, the pm and mm accessor functions return the PM and MM probe intensities. To see, for example, the intensities of the first three probes for probe set ID Gma.7559.1.S1_s_at we execute the following R code.

```r
> pm(scy.ab, 'Gma.7559.1.S1_s_at')[1:3,]
   hr.a3.12 hr.b3.12
Gma.7559.1.S1_s_at1  559    706
Gma.7559.1.S1_s_at2  586    788
Gma.7559.1.S1_s_at3  255    564
```

![Figure 1: (A) 10× magnified soybean leaf that displayed the resistant (Hawaii 94-1) reaction to the fungus Phakopsora pachyrhizi. (B) 10× magnified soybean leaf that displayed the susceptible (Taiwan 80-2) reaction to the fungus P. pachyrhizi.](image-url)
The phenoData slot of the *AffyBatch* class is where phenotypic data is stored. A call from the function phenoData may be used to access this information.

```r
> phenoData(soy.ab)
rowNames: hr.a3.12, hr.b3.12, ..., ts.c4.12 (6 total)
```

Additional information regarding annotation, notes, etc., is contained in *AffyBatch* objects.

**Extraction of soybean genome transcripts**

The Soybean Genome Chip contains spots representing genes from three different species. The Affymetrix GeneChip® Soybean Genome Array contains over 37 500 *Glycine max* (soybean) transcripts, approximately 15 800 transcripts for *Phytophthora sojae* (a water mold that commonly attacks soybean crops) and over 7500 transcripts representing *Heterodera glycines* (cyst nematode pathogen) [8]. Therefore, prior to quality control assessment, it is necessary to extract only those probe set IDs, referred to as affyids, which are specific to the soybean genome. Following the loading of the ‘affy’ package and reading in the .CEL files one obtains an *AffyBatch* object. However, as stated earlier, this *AffyBatch* object includes, in addition to the affyids for the soybean genome, numerous additional affyids for *P. sojae* species and for *H. glycines* pathogens. After reading in the six .CEL files for these data, we obtained the following *AffyBatch* object.

```r
> soy.ab
**AffyBatch object**
size of arrays=1164x1164 features (63517kb)
cdf=Soybean (61170 affyids)
number of samples=6
number of genes=61170
annotation=soybean
notes=
```

Of the 61 170 affyids, only 37 744 represent the soybean genome. The affyids associated with the soybean genome are of species ‘*Glycine max*’. Following the procedure in the R workspace SoybeanCutObjects.RData, that may be downloaded from our website, we executed the following commands:

```r
> Species.Affy.ID <- read.table
   ('<pathname>/SpeciesAffyID.txt',
+ header=T, sep=" ")
> load('<pathname>/SoybeanCutObjects.RData')
> source('http://www.bioconductor.org/
biocLite.R')
> biocLite('soybeanprobe')
> tv.for.glycine.max <- Species.Affy.ID$species == 'Glycine max'
> listOutProbeSets <-
+ Species.Affy.ID$affyID[tv.for.glycine.
max==FALSE]
> listOutProbeSets <- as.character
(listOutProbeSets)
> RemoveProbes (listOutProbes=NULL,
listOutProbeSets,
+ cdfpackagename, probepackagename)
and obtained the following *AffyBatch* object for the same data set.

```r
> soy.ab
**AffyBatch object**
size of arrays=1164x1164 features (63517kb)
cdf=Soybean (37744 affyids)
number of samples=6
number of genes=37744
annotation=soybean
notes=
```

(More complete source code for this chunk may be downloaded from our website: http://css.ncifcrf.gov/services/download/MicroarraySoybean.zip, from the document entitled ‘Soybean R 2-5-1 Code Chunks.doc’.) It should be noted that it does not matter if one cuts out the probes before or after reading the .CEL files to an *AffyBatch* object because the function does not actually remove any data from the *AffyBatch* object. Instead, it removes the mappings from the x-, y-coordinates to the probes in the cdf environment so that when R is told to look at the environment, R cannot detect that those probe sets are there (Dr Jenny Drnevich, personal communication).

**QUALITY ASSESSMENT**

**Exploratory data analysis**

A typical first step is to look at the image plots of the (PM and MM) probe-level data to determine if any anomalies exist. What is considered an anomaly is context dependent. In general, one looks for spatial artifacts or other nonhomogeneous patterns in the image plots. Figure 2 displays the log (base 2) image
plots for the six arrays in our data, generated with the following R code.

```r
palette.gray <- c(rep(gray(0:10/10), times=seq(1,41, by=4)))
par(oma=c(3,1,3,1))
par(mfrow=c(2,3))
image(soy.ab[,1], col=palette.gray)
image(soy.ab[,2], col=palette.gray)
image(soy.ab[,3], col=palette.gray)
image(soy.ab[,4], col=palette.gray)
image(soy.ab[,5], col=palette.gray)
image(soy.ab[,6], col=palette.gray)

Viewing the image plots across all arrays may help in determining whether one or more arrays might appear abnormal. For example, a potentially defective array may appear lighter or darker than the others, or display spatial artifacts (rings, shadows, etc.) not evident in the other arrays. The image plots in Figure 2 appear similar to each other and display no obvious anomalies.

Next, it is useful to plot boxplots and density plots of the probe-level data. To determine the existence of potentially defective arrays, we look for boxplots that stand out from the others, as evidenced for example by distinctly different ranges or displaced boxes (interquartile ranges, IQR). With respect to the density plots, we look for densities that are removed from the others, or that display bimodalities, show uniquely different shapes or other abnormalities. The following R code produces Figure 3A and B.

# Construct color boxplots
> library('RColorBrewer')
> brewer.cols <- brewer.pal (6, 'Set1')
> boxplot(soy.ab, col=brewer.cols,
+ ylab='Unprocessed log (base 2) scale Probe Intensities',
+ xlab='Array Names')

# Construct density plots
> hist (soy.ab, col=brewer.cols, lty=1,
+ xlab='Log (base 2)'

Figure 2: Image plots of perfect match (PM) and mismatch (MM) probe intensities (log base 2) for six arrays. The top three image plots correspond to the resistant (Hawaii) reaction and the bottom three to the susceptible (Taiwan) reaction.
The probe-level data for the boxplots (Figure 3A) of these arrays are distributed from about 2–14 on the log (base2) scale. The IQR, represented by the boxes in the boxplots, overlap each other to a large extent. The density plot (Figure 3B) does not display any bimodalities or any other particular anomalies and there is a great deal of overlap among the individual density plots. This suggests good quality in the chips.

**MA plot for the probe-level data**

Another exploratory plot for quality assessment is the MA Plot. When two microarrays are being compared, the difference of their log intensities for each probe on each gene (usually denoted ‘M’) are plotted against their average (usually denoted ‘A’). When it is desired to compare more than two arrays, a *synthetic* array is created by taking the probe wise medians [12] across all arrays. Each microarray may now be plotted versus the synthetic array. Figure 4 shows the MA plot for the hr.a3.12 array, with the median synthetic array centered at zero. Figure 4 is produced with the command

```
> MAplot(soy.ab, which=1)
```

which automatically adds a lowess regression line to the plot. The MAplot command, without the which=1 option, produces six MA plots. Source code for this may be downloaded from our website in the document entitled ‘Soybean R 2-5-1 Code Chunks.doc’.

Quality problems are most apparent from an MA-plot in cases where the loess smoother oscillates wildly or if the variability of the M values appears to be greater in one or more arrays relative to the

---

*Figure 3: (A) Boxplots and (B) density plots of six arrays of probe-level soybean data. The first three boxplots correspond to the *resistant* (Hawaii) reaction and the last three to the *susceptible* (Taiwan) reaction.*
Affymetrix quality assessment metrics

Affymetrix software produces a number of quantities for quality assessment of GeneChip® data [13, 14]. Among these are: (i) the average background, (ii) scale factor and (iii) percent present, to be defined shortly. The ‘simpleaffy’ library in the Bioconductor software contains the qc function to extract these quantities for many Affymetrix GeneChips®. However, the qc function, at present, cannot interpret the Affymetrix Soybean GeneChip®. Thus, we had to modify the code in the qc function in library ‘simpleaffy’ to extract these quantities for the Affymetrix Soybean GeneChip®. Documentation for how to do this, and our soybean.mod.qc.affy function are provided in the document entitled ‘Soybean R 2-5-1 Code Chunks.doc’ that may be downloaded from our website.

The first quantity to be examined is the ‘average background’. The background adjustment is outlined in the [15]. Each chip is divided into a grid of \( k \) (default, \( k = 16 \)) rectangular regions. In each region, the lowest 2% of the probe intensities are used to compute the background values [15, 16]. The average background values for our data are, respectively, 43.5, 56.5, 43.7, 53.5, 48.2 and 38.4. According to the guidelines recommended by Affymetrix, the average background values should be ‘comparable’ to each other. Assuming normality for these six values, all fall within two standard deviations of the mean. We note that these values obtained from our soybean.mod.qc.affy function agree with those obtained from Affymetrix MAS 5.0 calls.

The next quantities of interest are the ‘scale factors’. In [17], normalization is accomplished by taking a baseline array to which all other arrays are scaled to have the same mean intensity. The same procedure is followed for each array. The scale factors for the six arrays we examined are, respectively, 0.46, 0.26, 0.55, 0.36, 0.49 and 0.61. Affymetrix recommends that these values be within 3-fold of each other. Our values are within 3-fold of each other, and thus satisfy these criteria. (We note that the values obtained from our modification of the qc function in ‘simpleaffy’ do not agree with those obtained from Affymetrix MAS 5.0 calls.)

The next quantities of interest are the ‘percent present’. The percent present is the percentage of probe sets identified as ‘present’ according to the Affymetrix detection algorithm. Our percent present values are, respectively, 73.2, 74.8, 70.6, 72.7, 70.3 and 70.0. These percent present values should be ‘similar’ for replicate samples with extremely low values being a possible indication of poor quality. All values are in the 70–75% range. Hence, these values are quite similar; they agree with those obtained from MAS 5.0 calls.

RNA degradation

The Bioconductor package contains algorithms that may detect RNA degradation. From [18], for every GeneChip® probe set, the individual probes are numbered sequentially from the 5’ end of the targeted transcript. When RNA degradation is sufficiently advanced, PM (perfect match) probe intensities should be systematically elevated at the 3’ end of a probe set, when compared to the 5’ end. For any single probe set, the probe effects dominate the effect of degradation. Thus, the 3’/5’ trend becomes apparent on average over large numbers of probe sets. It has been observed that the 3’/5’ trend is roughly linear for the middle probe positions and lower at the ends.
Figure 5 shows the RNA degradation plot for the six arrays in our soybean experiment. This was generated with the R commands.

```r
> RNAdeg <- AffyRNAdeg(soy.ab)
> plotAffyRNAdeg (RNAdeg, col=c(rep('blue',3), rep('red', 3)))
```

A summary of the slopes for the degradation plots is provided by the command.

```r
> summaryAffyRNAdeg(RNAdeg)
```

<table>
<thead>
<tr>
<th></th>
<th>hr.a3.12</th>
<th>hr.b3.12</th>
<th>hr.c3.12</th>
<th>ts.a4.12</th>
<th>ts.b4.12</th>
<th>ts.c4.12</th>
</tr>
</thead>
<tbody>
<tr>
<td>slope</td>
<td>2.28000</td>
<td>1.84000</td>
<td>1.010</td>
<td>1.9000</td>
<td>1.4400</td>
<td>0.817</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.00239</td>
<td>0.0139</td>
<td>0.0102</td>
<td>0.00116</td>
<td>0.0409</td>
<td>0.206</td>
</tr>
</tbody>
</table>

There are no clear guidelines on how large a degradation slope must be to determine a bad chip. Different chip types have different characteristic slopes [19]. The values of the slopes we obtained differ by approximately 1 within the resistant (Hawaii) and susceptible (Taiwan) arrays. These plots appear to be reasonably parallel as well.

It should be noted that some authors have questioned the utility of RNA degradation plots as a measure of assessing RNA integrity [20]. In addition, postings to the Bioconductor mailing list bioconductor@stat.math.ethz.ch, have suggested that the RNA degradation plots have limited utility.

**PROBE-LEVEL MODELS**

Probe-level models (PLM) may be useful in determining the quality of Affymetrix chips. A PLM is a model that is fit to probe-intensity data [21]. Specifically, a PLM provides parameter estimates for probe sets and chips (arrays) on a probe-set by probe-set (i.e. gene by gene) basis. The `affyPLM` package in Bioconductor provides functions that fit PLM through the function `fitPLM`.

The `fitPLM` function provides the user with a wide variety of methods for background correction and normalization. In our application, we chose to invoke the default options. Specifically, we chose the option ‘RNA convolution’ for background correction [21, 22], and the option ‘Quantile Normalization’ for normalization [21, 23]. We accomplished this with the following simple commands.

```r
> library('affyPLM')
> Pset1 <- fitPLM(soy.ab)
```

Having performed the convolution background correction and quantile normalization procedures, the following linear PLM for the background adjusted normalized probe-level data, $S_{gij}$, may be stated [21, 24]:

\[
(\log_2)S_{gij} = \theta_g + \phi_i + \epsilon_{gij}
\]  

(1)

where, $\theta_g$ represents the log-scale expression level for the $g$-th gene on the $i$-th array, $\phi_i$ represents the $j$-th probe on the $g$-th gene and $\epsilon_{gij}$ represents the measurement error. We note that, following convolution background correction and quantile normalization procedures described, the signal $S_{gij}$ is, in fact, the PM (perfect match) value for the $j$-th probe on the $g$-th gene on the $i$-th array. The `fitPLM` function fits model (1) by employing an M-estimator robust regression algorithm [25–27]. The fitted object contains information regarding the parameter estimates, standard errors, weights, residuals and signed residuals [25, 28].

Figure 6, panels A–D, generated with the following R commands

```r
> par (mfrow=c(2,2))
> par (oma=c(3,1,3,1))
```
show the gray scale image of the log intensities, along with the weights, residuals and signed residuals of the fitted PLM for array hr.a3.12. One looks for any anomalies, artifacts or nonhomogeneous patterns in these plots. Panel A (upper left) shows the gray scale image of the log intensities. It is identical to the upper left image in Figure 2. There are no apparent spatial artifacts in this plot. Panel B shows the PLM Weights plot for one of our six arrays. Weight images use topographical coloring so that light areas indicate high weights and dark areas (green in color plots) indicate significant down-weighting of mis-performing probes [29]. These weights are obtained from the robust regression procedure for which small weights are associated with outliers. There are no apparent spatial artifacts in this plot. Panel C shows the PLM Residuals plot for the same array. Images based on residuals are ‘dark’ for negative residuals (blue in color plots) and ‘light’ for positive residuals (red in color plots) [29]. Panel C shows that the positive and negative residuals are homogeneously spread out across the image. Finally, panel D shows the PLM Signed Residuals plot for the same array [29]. The image of the signs of the residuals report either +1 or −1 depending on whether the residual is positive or negative. This can sometimes make visible effects that might not be apparent in the other plots. These images highlight the power of the PLM procedures at detecting a subtle artifact that might otherwise be missed completely. Our data show no artifacts or nonhomogeneous patterns. There are numerous ways in which images showing the ‘weights’, ‘residuals’ and ‘signed residuals’ may display anomalies. Bolstad maintains a website (http://plmimagegallery.bmbolstad.com/) showing interesting artifacts (e.g. ‘crop circles’, ‘ring of fire’, ‘tricolor’, ‘moon shape’, etc.) observed in the chip pseudo-images produced by the PLM procedures in the Bioconductor affyPLM package.

**RELATIVE LOG EXPRESSION (RLE)**

Other useful statistics and procedures are possible following the fitting of a PLM. We first examine the RLE plot. This plot is constructed in the following manner. First, start with the log scale estimates of expression $\hat{\theta}_g$ for each gene $g$ on each array $i$ obtained from the PLM fit. Next, compute the median value across arrays for each gene, $m(g)$, and define the RLE as

$$\text{RLE}_g = \hat{\theta}_g - m(g)$$  \hspace{1cm} (2)

That is, the median value of the $g$-th gene is subtracted from each gene $g$ on each array $i$.

These relative expressions are then displayed with a boxplot for each array. In our research it is reasonable to assume that the majority of genes are not changing in expression between resistant and susceptible reactions. The majority of these nondifferential genes are displayed on the RLE plot by the boxes (IQRs). Ideally, these boxes should have small spread and be centered at RLE = 0. An array with quality problems may result in a box that has relatively greater spread or that is not centered near RLE = 0 [30]. Figure 7A, generated with the following command

```r
> Mbox(Pset1, col=brewer.cols, main='RLE Plot')
```

shows the RLE plot for our data. The arrays are nicely centered around RLE = 0, with approximately equal box sizes (i.e. IQRs) and present no quality control problems.
NORMALIZED UNSCALED STANDARD ERROR PLOT (NUSE)

Another graphical tool is the NUSE plot. For this plot, one begins with the standard error estimates obtained for each gene $g$ on each array $i$ from the PLM fit. Call this SE (\(\hat{\theta}_g\)). Variability may differ considerably among genes. To correct for this, one may standardize these standard error estimates such that the median standard error across arrays is 1 for each gene. Specifically,

\[
\text{NUSE}(\hat{\theta}_g) = \frac{SE(\hat{\theta}_g)}{\text{med}(SE(\hat{\theta}_g))}
\]

A low quality array on this plot might be indicated by a box that is significantly elevated or shows more spread relative to other arrays. The median NUSE values for the arrays provide suitable summary values for this procedure. High values of median NUSE may suggest that one has a problem with the array. However, it should be noted that these values are not comparable across data sets since NUSE is relative only within a data set [30].

Figure 7B shows the NUSE plot for our data. It appears that the arrays are reasonably centered around the median NUSE = 1, with approximately equal box sizes (i.e. IQRs). They do not appear to present any quality control problems.

A NOTE CONCERNING OVERALL CHIP QUALITY

All of the chips that were used in this research were found to be of good quality. This may be confirmed by readers who are interested in reproducing all our results by executing the R code chunks in our 'Soybean R 2-5-1 Code Chunks.doc', which may be downloaded from our website. But what should the researcher do if abnormalities are found in the quality control plots? If the situation arises in which researchers observe obvious abnormalities in the quality control plots, it would be appropriate to remove those chips from the data set before pursuing further analyses. Examples of ‘obvious abnormalities’ include: (i) gray scale image plots for chips which are lighter or darker than those in the other arrays; (ii) corresponding boxplots that show no tail below the box (IQR) in the boxplot, or that stand out from other boxplots that correspond to other arrays or that have ranges notably longer or shorter than the other boxplots; (iii) density plots that show an extreme spike at the low or at the high end of the density plot, or that show bimodalities; (iv) MA-plots for the probe-level data which show a ‘fish hook’, or other abnormal, appearance indicating wildly nonhomogeneous probe-level values in the chip, or a wildly deviating lowess regression line. Our decision criterion was to remove a chip from the analysis if it failed two or more of our specific quality control measures outlined above. In this research, all the chips passed our quality control criteria.

STATISTICAL ANALYSES FOR DIFFERENTIAL GENE EXPRESSION

Background adjustment, normalization and summarization by the Robust Multichip Averaging (RMA) algorithm

After preprocessing, it is necessary to transform the data into a final expression value for each gene. With Affymetrix data this usually involves at least three steps: (i) background adjustment, (ii) normalization and (iii) summarization. From [31],

Background adjustment is essential because parts of the measured probe intensities are due to nonspecific hybridization, noise in the optical detection system and other reasons. Observed intensities need adjustment to give accurate measurements of specific hybridization. Without proper
normalization, it is impossible to compare measurements from different array hybridizations due to many obscuring sources of variation. These include different efficiencies of reverse transcription, labeling, or hybridization reactions, physical problems with the arrays, reagent batch effects and laboratory conditions. With Affymetrix data, summarization is needed because transcripts are represented by multiple probes. For each gene, the background adjusted and normalized probe intensities need to be summarized into one value.

There are many strategies for performing these three steps. We have chosen the RMA procedure, which performs a convolution background correction, quantile normalization and summarization based on a multi-array model fit robustly using the median polish algorithm [32]. This algorithm only uses PM information [22, 33]. The RMA procedure takes the probe-level data and transforms it to produce expression level data. This results in a single expression level value for each gene (or affyid). We performed the RMA procedure and obtained the expression level values, in log (base 2) scale, with the following commands.

> eset <- rma(soy.ab)
> exprs.eset <- exprs(eset)

Boxplots and MA-plot following RMA background adjustment, normalization and summarization

Figure 8A shows boxplots of the six arrays obtained with the commands.

> Index1 <- 1:3
> Index2 <- 4:6
> Difference <- rowMeans(exprs.eset[,Index1]) - rowMeans(exprs.eset[,Index2])
> Average <- rowMeans(exprs.eset)
> exprs.eset.df <- data.frame(exprs.eset)
> boxplot(exprs.eset.df, col=brewer.colors)

One can see that these arrays are nicely aligned. We note, however, that boxplots of expression level values following RMA transformation will often look neatly aligned. This is due to the aggressive quantile normalization step in the RMA procedure. Hence, they do not represent a means for evaluating the quality of the probe-level data, but of the success of the normalization and summarization.

Figure 8B displays an MA-plot of log-fold changes. An MA-plot for the expression level data is a plot of the differences of the average log (base 2) values between the resistant (Hawaii) and susceptible...
(Taiwan) populations versus the mean log-expression across the two populations: resistant (Hawaii) and susceptible (Taiwan). Figure 8B was obtained with the commands.

```R
> plot(Average, Difference)
> lines(lowess(Average, Difference), col = 'red', lwd = 4)
> abline(h = -2)
> abline(h = 2)
```

A lowess regression line is added to the plot. A reasonably straight lowess regression line, as is seen here, provides good evidence that the normalization and summarization procedures were adequate. For log (base 2) data, a value of positive 1 (negative 1) indicates a fold change of 2 to 1 (1 to 2). For log (base 2) data, a value of positive 2 (negative 2) indicates a fold change of 4 to 1 (1 to 4). Of the 37,744 genes in the Glycine max subset, 12 have absolute log-fold changes equal to or greater than 2. Of these 6 were greater than +2. These represent possible candidates for soybean genes that are up-regulated in the resistant samples. Six (6) were or less than −2, indicating possible candidates for genes that were down-regulated in the resistant samples.

**SUMMARY STATISTICS AND TESTS FOR RANKING**

One can see from Figure 8B, that the variances are not completely homogeneous across the plot. More spread is observed in the middle than in the ends. Thus, in further assessing differences between the resistant and susceptible populations, one should use a procedure that takes into account the fact that different genes will exhibit different magnitudes of variability. A popular choice is the \( t \)-statistic. Hence, \( t \)-statistics for all 37,744 genes were computed.

We are interested in determining how much our rankings change if we use the \( t \)-statistic instead of the average log-fold change. A ‘volcano’ plot is useful to see both these quantities simultaneously [34–37]. A ‘volcano’ plot displays the ‘lod’ scores, which are the negative logs (base 10) of the \( P \)-values [i.e. \(-\log_{10} (P)\)]. For example, a lod value of 3.0 represents a probability of 0.001. We do not show the volcano plot for the lod scores versus log-fold change as produced by the standard \( t \)-statistic in this article. However, the commands are available in our document entitled ‘Soybean R 2-5-1 Code Chunks.doc’ from our website http://css.ncifcrf.gov/services/download/Microarray Soybean.zip.

Statistical significance, assessed using the \( t \)-statistic, evaluates the average difference between the two population values relative to their variances. Small average differences in the context of large variances could result in a large, and therefore, nonsignificant \( P \)-values. On the other hand, small average differences with small variances could result in significant \( P \)-values. Large average differences with small variances might result in significant \( P \)-values. And, finally, large average differences with large variances could result in nonsignificant \( P \)-values. In summary, it is possible to have large log-fold changes that are not statistically significant because the populations exhibit much variability. It also is possible to have small log-fold changes that are highly statistically significant because the populations exhibit little variability. For these reasons, Irizarry states [38]:

Researchers have proposed alternative statistics that borrow information about variability across all genes to obtain a more stable estimate of gene-specific variance. These procedures attempt to minimize the impact of genes with very large and very small variances. The resulting statistics are referred to as modified, penalized, attenuated or regularized \( t \)-statistics... An example of such a modified \( t \)-statistic is given by Smyth [39]...

A moderated \( t \)-statistic is available in the ‘limma’ package in Bioconductor. It is based on an empirical Bayes approach described in detail in [39] and [40]. The moderated \( t \)-statistics and their corresponding \( P \)-values were computed with the following commands.

```R
> library('limma')
> population.groups <- factor(c(rep('Taiwan/Susceptible', 3), + rep('Hawaii/Resistant', 3)))
> design <- model.matrix(~ population.groups)
> fit <- lmFit(eset, design)
> fit.eBayes <- eBayes(fit)

The moderated \( t \)-statistics and their corresponding \( P \)-values are contained in objects fit.eBayes$t[,2] and fit.eBayes$p.value[,2], respectively.
Figure 9 shows a volcano plot constructed with the moderated \( t \)-statistics produced with the following commands.

```r
> lodd <- -log10 (fit.eBayes$p.value[,2])
> o1 <- order (abs(Difference), decreasing =TRUE)[1:50]
> oo2 <- order(abs (fit.eBayes$t[,2]), decreasing =TRUE)[1:50]
> oo <- union (o1, oo2)
> ii <- intersect (o1, oo2)
> plot (Difference[-oo], lodd[-oo],
>       cex=.25, xlim=c (-3,3), ylim=range (lodd), xlab='Average (log)
>       Fold-change',
>       ylab='LOD score – Negative log10 of
>       P-value')
> points (Difference[o1], lodd[o1],
>        pch=18, col = 'blue', cex=1.5)
> points(Difference [oo2], lodd [oo2],
>        pch=1, col = 'red', cex=2,
>        lwd=2)
> abline (h=3)
```

In this plot, the 50 genes with the lowest \( P \)-values are shown as open circles (red in the color plot). Similarly, the 50 genes with the highest (absolute) log-fold changes are represented as diamonds (blue in the color plot). Of interest to biologists in differential gene expression studies are those genes that possess simultaneously, (i) low probability values (as determined by the moderated \( t \)-statistic) and (ii) high (absolute) log-fold changes.

Figure 9 displays graphically nine genes that satisfy these criteria.

**A NOTE REGARDING THE MULTIPLE COMPARISONS PROBLEM**

The Multiple Comparisons Problem (MCP) confronts microarray researchers attempting to identify which genes are differentially expressed between classes/populations [41]. Suppose that a researcher has computed, say, \( t \)-tests between two different classes on 10 000 genes and has identified all those genes that resulted in \( P \)-values <0.05. Under the null hypothesis of no differential expression among genes, one would find, on average, 5% of the genes to have \( P \)-values <0.05. With 10 000 genes, this would imply that 500 genes, on average, would be identified as being significant at the 0.05 level, resulting in 500 false positives. Various strategies have been proposed to deal with this problem: the significance analysis of microarrays (SAM) statistic [42], the false discovery rate (FDR) [43], the positive false discovery rate, pFDR [44] and the \( Q \)-value, a Bayesian posterior \( P \)-value or pFDR analogue of the \( P \)-value [45].

**Our decision to use Smyth’s moderated \( t \)-statistic and high (absolute) log-fold changes**

The MCP is a useful way of controlling for the problems encountered in determining differentially expressed genes when the \( P \)-values obtained from various statistical tests are the sole criterion. Our problem, however, was to determine those genes that satisfied two criteria useful to biologists: (i) low probability values and, (ii) high (absolute) log-fold changes. Hence, we chose to use Smyth’s moderated \( t \)-statistic (along with its corresponding \( P \)-value) because it provides for more stable inferences when the number of arrays is small. In addition, it is easily accessed in the ‘limma’ package in Bioconductor.

**HEATMAP**

A heatmap is a two dimensional, colored grid that displays data in the form of a rectangular matrix. The color, or gray scale, of each rectangle is determined by the value of the corresponding
entry in the matrix. The rows and columns of a heatmap can be ordered (i.e. clustered) such that similar rows are placed next to each other, and, independently, similar columns placed next to each other. Hierarchical clustering algorithms have traditionally been used in microarray analysis [46]. Hierarchical clustering algorithms may be broadly divided into two types [47]. ‘Divisive’ algorithms create a ‘tree diagram’ (dendrogram) by considering all the data as first being comprised of a single cluster (i.e. the ‘trunk’ of the tree). As the algorithm progresses, the ‘trunk’ separates into ‘limbs’, ‘branches’, ‘twigs’ and eventually into ‘leaves’. More specifically, the dendrogram is built from the ‘top down’ by recursively partitioning the data elements [48]. ‘Agglomerative’ algorithms, on the other hand, begin with the ‘leaves’, and construct their hierarchy such that ‘twigs’, ‘branches’ and ‘limbs’ merge until only a single cluster remains, the ‘trunk’ of the tree. That is, the dendrogram is built from the ‘bottom up’ by recursively combining the data elements [48]. Hierarchical clustering methods (as do all other clustering methods) require that the ‘distances’ initially be computed between all pairs of the elements. There are a number of different distance measures that may be employed, the most popular of which is ‘Euclidean’ [49].

Figure 10 shows a heatmap of the nine intersecting genes that possess simultaneously high (log) fold changes and low probability values that were displayed in the volcano plot (Figure 9). The heatmap was constructed with the following R commands.

```r
> ii.mat <- exprs.eset[ii,]
> ii.df <- data.frame(ii.mat)
> library (‘RColorBrewer’)
> library (‘genefilter’)  
> hmcol <- colorRampPalette(brewer.pal(9, ‘Greys’))(256)
> tv.Hawaii.Resistant <- dimnames(ii.mat)[[2]][1:3]
> spcol <- ifelse (dimnames(ii.mat)[[2]] == tv.Hawaii.Resistant, ‘grey10’, ‘grey80’)
> heatmap (ii.mat, col=hmcol, ColSideColors=spcol, + margins = c (10, 15))
```

The dendrograms at the top and side of Figure 10 were constructed with an agglomerative complete linkage algorithm, using a Euclidean distance measure [47–50]. Note that the three Hawaii/Resistant and three Taiwan/Susceptible arrays cluster tightly in Figure 10, indicating that the expression level values for these arrays on the nine intersecting genes had similar expression level values. This method was found to be useful because it so clearly identified the three Hawaii/Resistant and three Taiwan/Susceptible arrays in distinctly different clusters, and performed (at least) as well as all other agglomerative methods available.

The dark gray hues indicate relatively higher gene expression levels (Hawaii/Resistant) than the light gray hues (Taiwan/Susceptible), which indicate relatively lower gene expression levels. This is consistent with the information displayed in Figure 9 (volcano plot) in which all genes of interest had positive log-fold changes—thus indicating over-expression of genes for the resistant (Hawaii) reaction. The dendrograms connect the arrays and genes with similar expression levels.

**RESULTS**

Table 1 contains information regarding the nine intersecting genes of possible biological interest determined by our analyses. It displays the Affymetrix ID along with its corresponding gene name and/or description.

Of these nine intersecting genes, only one (affyID GmaAffx.80951.1.S1_at) bears no significant similarity to sequences currently in the public GenBank database. The remaining eight have been previously
demonstrated to be involved in plant defense and stress responses. Three of the genes, HSP81-1 (2) and HSP 82 (affyIDs GmaAffx.21211.1.S1_at, GmaAffx.92386.1.S1_at and Gma.6606.1.S1_at) belong to the heat shock protein 90 family, which is composed of molecular chaperones. These proteins play an important role in the folding of newly synthesized proteins, stabilization and in refolding of denatured proteins after stress. HSP81-1 and HSP 82 have domains suggesting that they are also implicated in signal transduction [51]. Two of the other genes, Isoflavone reductase and Polyphenol oxidase (affyIDs Gma.16735.2.S1_at and Gma.7559.1.S1_s_at), are enzymes that participate in the production of defense related aromatic compounds in plants [52]. Lipase SIL1 (2) (affyIDs GmaAffx.91805.1.S1_at and GmaAffx.91805.1.S1_s_at) is an uncharacterized lipid-processing enzyme that has been shown to be induced by pathogens in another plant–microbe system [53]. Finally, an MYB transcription factor (affyID GmaAffx.92383.1.S1_at), is a homeodomain-containing DNA-binding protein that is part of the plant and animal response to abiotic stress [54] and is known to regulate flavonoid synthesis [55].

Increased production of aromatic defense compounds, which leads to thickening and strengthening of the plant cell wall, increased ability to respond to oxidative stress and enhanced signal transduction, has been shown to be an important component of the soybean response to the bacterial pathogen 

\textit{Pseudomonas syringae} [56, 57]. Three of the top nine genes identified as differentially expressed in this study participate in the production of these aromatic compounds, suggesting a role for these compounds in the immune response to soybean rust. Additional experiments will be needed to identify which specific compounds are playing a role in inhibiting fungal growth, including assays of individual protein concentrations in infected leaves, and tests for susceptibility to rust in soybean mutants with compromised aromatic synthesis.

**DISCUSSION AND SUMMARY**

This experiment was designed to identify genes that play a role in soybean resistance to a fungal pathogen. This article has described specific procedures for (i) conducting quality assessment of Affymetrix GeneChip® Soybean Genome data and (ii) performing an analysis to determine differential gene expression using the open-source R programming environment and with the open-source Bioconductor software. The analyses performed and described herein successfully identified differentially expressed genes that are already known to be involved in plant defense. This lends credibility to and supports the fact that our experimental approach and analytic procedures were useful in successfully identifying differentially expressed genes. We have shown that these procedures have successfully identified differentially expressed genes that play a role in soybean resistance to the fungal pathogen, 

\textit{P. pachyrhizi}.

**Key Points**

- Quality assessment methods of Affymetrix GeneChip® arrays provide the analyst with tools that may assist him/her in making decisions as to which arrays (if any) should be removed due to poor quality. Only high quality chips should be used for downstream analyses.

- To conduct analyses for the soybean genome, it is necessary to extract from the Affymetrix GeneChip® soybean chip only those affyIDs that pertain specifically to the soybean genome, specifically, the ‘Glycine max’ species.

- The open-source statistical R programming environment, in conjunction with the open-source Bioconductor software, provides a rich environment for performing quality control assessment, graphical displays and a multitude of functions and procedures for performing statistical and differential gene expression analyses.

**Acknowledgements**

This project has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract #N02-CO-12401. This work was supported in part by the United Soybean Board Project 4217 to Reid D. Frederick. The authors wish to thank Dr Jenny Dmievich.
(University of Illinois, Urbana-Champaign) for many useful suggestions in connection with procedures for extracting the ‘Glycine max’ species (those that specifically relate to the soybean genome) from the generic Affymetrix GeneChip® ‘soybean’ chip. The authors also thank Dr Rob Leighty (DMS, NCI-Frederick) for his assistance in the preparation of the final article. The authors also thank three anonymous reviewers for their helpful comments that improved the quality of this article.

SUPPLEMENTARY MATERIALS
Supplementary materials are available at Briefings in Bioinformatics Online.

References
Differential gene expression in the soybean genome


52. Wu Q, VanEtten HD. Introduction of plant and fungal genes into pea (Pisum sativum L.) hairy roots reduces their ability to produce pisatin and affects their response to a fungal pathogen. Mol Plant Microbe Inter 2004;17:798–804.


