# DNA Array Analysis in a Microsoft® Windows® Environment

*BioTechniques 32:110-119 (January 2002)*

**T. Conway, B. Kraus, D.L. Tucker, D.J. Smalley, A.F. Dorman, and L. McKibben**
The University of Oklahoma, Norman, OK, USA

**ABSTRACT**

*Microsoft® Windows®-based computers have evolved to the point that they provide sufficient computational and visualization power for robust analysis of DNA array data. In fact, smaller laboratories might prefer to carry out some or all of their analyses and visualization in a Windows environment, rather than alternative platforms such as UNIX. We have developed a series of manually executed macros written in Visual Basic for Microsoft Excel® spreadsheets, that allows for rapid and comprehensive gene expression data analysis. The first macro assigns gene names to spots on the DNA array and normalizes individual hybridizations by expressing the signal intensity for each gene as a percentage of the sum of all gene intensities. The second macro streamlines statistical consideration of the confidence in individual gene measurements for sets of experimental replicates by calculating probability values with the Student's* t *test. The third macro introduces a threshold value, calculates expression ratios between experimental conditions, and calculates the standard deviation of the mean of the log ratio values. Selected columns of data are copied by a fourth macro to create a processed data set suitable for entry into a Microsoft Access® database. An Access database structure is described that allows simple queries across multiple experiments and export of data into third-party data visualization software packages. These analysis tools can be used in their present form by others working with commercial* E. coli *membrane arrays, or they may be adapted for use with other systems. The Excel spreadsheets with embedded Visual Basic macros and detailed instructions for their use are available at http://www.ou.edu/microarray.*

## INTRODUCTION

With the publication of the first genome-wide expression data six years ago (13), the power of the DNA array was immediately evident. Since then, reviewers have predicted a revolution in the way biologists conduct their research (3–5,12). Today the trend toward widespread use of DNA arrays continues. Vast amounts of DNA array data are accumulating, and the need for standardized array annotation and data representation is being addressed (http://www.ebi.ac.uk/microarray/MGED/). Relational databases are being developed to handle microarray data storage in a format that facilitates data processing and visualization, allowing researchers to analyze and interpret their experiments and disseminate the data (16). Generally, the database management software used is Oracle, Web interfaces are written in JAVA or XML, and scripts used for data processing, retrieval, etc. are written in Perl, C, or other programming languages. However, for the common microarray user, the implementation of such a system may be beyond their budget or exceed their actual needs. For these users, a central archive is an attractive alternative for data storage. An example of this, the EcoReg Consortium (http://gobi.lbl.gov/~ecoreg/index.html), is being established as a public database for storage and manipulation of *E. coli* microarray and proteome data. Still, it will be necessary for individual users of consortium databases to process, analyze, and format the data for submission.

DNA arrays have been used to examine the genetics and physiology of the comprehensive biological model, *E. coli*. Some research groups have used commercial membrane-based DNA arrays (2,17,18), while others have employed DNA microarrays (6,11,19,20) or oligonucleotide arrays (15). Several practical issues regarding the use of whole-genome arrays have been addressed, and the power of this technology as a means for deducing the physiological state of the bacterial cell is now well established.

In our laboratory, we routinely use membrane-based DNA arrays for *E. coli* gene expression profiling. Multiple replicates of each experimental condition are processed and analyzed with manually executed macros written in Visual Basic and run in Microsoft® Excel® spreadsheets on Microsoft Windows®-based computers. These macros, statistical analysis, and data processing protocols are described here and are freely available to the scientific community (http://www.ou.edu/microarray).

## MATERIALS AND METHODS

### Software and System Requirements

The analysis tools described here are written for Microsoft Excel and Microsoft Access®: Microsoft Office® 97 or higher is required. Optimal system

requirements include a Windows-compatible computer with a PII processor and 128 Mb RAM or higher. Macintosh® computers running Microsoft Office 98 for Macintosh can also be used to run the macros in Excel.

### Example *E. coli* Data Set

In this study, we compare a sample data set of *E. coli* MG1655 grown at pH 7.4 and pH 5.5 under otherwise identical conditions. Cultures were grown aerobically in 50 mL MOPS (pH 7.4) or MES (pH 5.5) minimal glucose (0.2%) medium (8) in 250-mL fleakers (Corning, Acton, MA, USA) at 37°C with 300 rpm agitation and harvested in mid-logarithmic growth phase. RNA isolation, radioactive labeling during cDNA synthesis, and hybridization to DNA array membranes were described previously (17).

### Description of Raw Data

We routinely use Panorama *E. coli* Gene Arrays™ (Sigma-Genosys, The Woodlands, TX, USA) for gene expression profiling. Phosphorimaging of a hybridized membrane array produces a TIFF image file that must be further processed for data analysis. The image analysis software (ArrayVision™ version 5.1; Imaging Research, St. Catharines, Ontario, Canada) makes use of a customized template to accommodate three grid layers ($3 \times 1$; $16 \times 24$; $4 \times 4$) according to the design of the Panorama *E. coli* gene arrays. The spot labeling protocol was edited such that each spot is named by its unique array coordinate, allowing the spot intensity measurements to be easily associated with the correct gene identifiers in subsequent processing steps. The customized ArrayVision template file for analysis of Panorama *E. coli* gene arrays, and detailed instructions for its use, are available at http://www.ou.edu/microarray. The spot intensities are represented in a row-column format and are exported into Excel spreadsheets for further analysis.

### Data Processing

The macros and sample analyses can be downloaded from http://www.ou.edu/microarray. Follow the links to "Macroarray", "Data Analysis", and then "Spot-Finding and Image Quantitation" or "*E. coli* Data Analysis (software downloads)"; alternatively, the macroarray section of the site can be accessed directly at http://www.ou.edu/microarray/macroarray.htm. Raw DNA array data, exported from ArrayVision, are processed in a series of three Excel workbooks that are used to manually execute four macros written in Visual Basic (Table 1). These macros are designed to filter the data and calculate statistics to allow for further data analysis and interpretation. Detailed, step-by-step instructions for use of these analysis tools are provided on our Web site.

## RESULTS AND DISCUSSION

The workbooks, macros, and subroutines used for DNA array data analysis are outlined in Table 1. The subroutines can be run in order individually, or the macro containing all relevant subroutines can be run once to execute all subroutines. In the following section,we provide an overview of important statistical considerations, the specific processing steps, and outcomes.

### Statistical Sgnificance

Various approaches for attaching significance to DNA array data have been published, including a simple "rule-of-thumb" criterion for the value of the expression ratio (2,20). Some researchers have used the standard deviation from the mean of the expression ratios as an indicator of confidence (6,17). Arfin et al. (1) applied the Student's *t* test to experimental replicates and considered the *P* value to be the most important indicator of significance. Richmond et al. (11) combined a confidence interval obtained with the Student's *t* test and a rule-of-thumb criterion for the expression ratio. Others prefer to consider the significance of a single experimental condition based on the coefficient of variation (18). A precedent for statistical analysis of array data has yet to be firmly established, and a standard is clearly needed. Whatever statistical approach is adopted, it is essential that DNA array experiments are properly replicated and the uncertainty that lies behind individual gene measurements be considered to attach significance to data sets.

We advocate the use of at least two replicates of each experimental condition. Membrane arrays typically have duplicate spots for each gene, and each spot is considered to be a separate determination, providing a total of four determinations for the two replicates. Because membrane arrays are hybridized with a single labeled target mRNA and normalized independently (in effect a one-color experiment), the variation in the measurement is at the level of the raw data, not the measurement of the ratio (unlike the statistical approach that is popular with two-color microarrays and involves internal normalization of the measurements). The uncertainty that lies behind individual gene measurements can be variously calculated as the standard deviation of the determinations or the coefficient of variation. Since we are usually interested in the statistical significance of differences between an experimental condition and a control, we prefer the Student's *t* test as a means for calculating this probability, based on the uncertainty of the replicate measurements in both conditions. The Student's *t* test is best applied to natural log-transformed normalized data sets (7). Generally, a *P* value of less than 0.05 is chosen to indicate a 95% probability that the difference in gene expression between conditions is significant. However, it has been pointed out that with very large data sets (e.g., a bacterium with 5000 genes) choosing a value of $P < 0.05$ means that there could be up to 250 false positives in the data set (1). Thus, the researcher is left with two choices: to lower the *P* value to a level where no false positives are expected ($P < 0.0002$) or to consider a second statistical metric that, when combined with a reasonable *P* value ($P < 0.05$), is an excellent indicator of significance of a ratio value.

We use the standard deviation of the mean of the log ratios—within the context of the *P* value—to indicate significant up- or down-regulation of gene expression. This approach is meaningful where the expression level of the majority of genes does not change significantly between conditions and where the researcher is interested in genes that show

**Table 1. Useful Purpose of Workbooks, Visual Basic Macros, and Subroutines Used in This Study**

| Workbook | Macro | Subroutine | Purpose |
|---|---|---|---|
| Image Data Cruncher | | | |
| | AllDataCrunched6 | | |
| | | ArvAllSort1 | associates array coordinate with spot number |
| | | Nameall2 | associates spot number with unique identifier for gene |
| | | CalcPercentage3 | normalizes data by expressing each spot as percentage of sum of all spot intensities |
| | | Cleanup4 | reorganizes data and calculates average values for duplicate spots |
| | | Statistics5 | calculates averages of genomic DNA controls and blank spots |
| 2-Replicate-Stats | | | |
| | AllAnalysis8 | | |
| | | OrganizebySpotNo1 | sorts each of the four data sets individually by spot number |
| | | CalculateAverages2 | calculates averages of volumes and percentage values for the control and test replicates |
| | | copyvaluesintoPRaw3 | copies and pastes percentage values into a separate spreadsheet for calculation of *P* values |
| | | CalculateLn4 | copies percentage values into a separate spreadsheet and natural log transforms data |
| | | CalculatePRaw5 | calculates the *P* value for the raw data by application of the Student's *t* test |
| | | CalculatePLn6 | calculates the *P* value for the log transformed data by application of the Student's *t* test |
| | | CopyAllValues7 | copies and pastes data used for ratio calculations into separate spreadsheet |
| Data Analysis | | | |
| | AllAnalysis6 | | |
| | | SpotSort1 | sorts control and test data sets by spot number |
| | | DataSort2 | copies data set into spreadsheet used for ratio calculations, sorts by total percentage value |
| | | ThresholdRatios3 | calculates ratio of Test/Control using threshold of tot percentage value for 500th lowest gene |
| | | Cleanup4 | reorganizes data and calculates log (10) of ratio |
| | | Stats5 | calculates standard deviation of log ratio values and correlation between Test and Control |
| | (Manual Step) | | copy and paste special values Praw and PLn values from 2-Replicate-Stats to Data Analysis |
| | MakeDBsheet | | reorganizes data and copies into separate spreadsheet for entry into Access database |

substantially different expression. The standard deviation for the log ratios is calculated, and only those genes that differ by more than three standard deviations (99.9% confidence in each tail) from the mean of the log ratio (usually zero, or no change) are considered. In practice, emphasis is placed on those genes that have expression ratios greater than three standard deviations from the mean and have a reasonable probability of being significantly different between the conditions, based on a *P* value of less than 0.05. Where there are four or more determinations for each gene, the *P* value can be lowered to less than 0.005 with little change seen in the number of genes that are considered to vary significantly between conditions.

There may also be situations when the researcher is interested in changes in gene expression that are not a full three standard deviations from the mean but are still significant (i.e., where the differences in gene expression between conditions are subtle yet meaningful). In this case the Student's *t* test can be used as the sole measure of significance, but the *P* value must be adjusted to ensure that false positives are avoided. One approach for this is to apply the Bonferroni correction that describes a *P* value for significance in a large data set (7). Various strategies for implementing this correction factor have been described, and its proper use is somewhat controversial (14). The Bonferroni correction effectively lowers the *P* value to a point where false positives are avoided and consideration is given only to those genes for which there is a high degree of confidence in

the ratio value. In this light, the Bonferroni correction seems to be a reasonable statistical tool but may be too stringent for some considerations.

### "Image Data Cruncher" Workbook

To begin the data analysis process, the raw data from an experimental replicate is copied and pasted into a blank spreadsheet named "arvdata" in an Excel workbook named "Image Data Cruncher" containing the macro named "AllDataCrunched6" (Figure 1). The first subroutine in the macro, "ArvAllSort1", copies and pastes subsets from the "arvdata" spreadsheet into a second spreadsheet named "allfields" that contains information provided by the membrane manufacturer for associating the array location with a spot number that is unique to each target on the array. The second subroutine, "Nameall2", copies and pastes the data in the "allfields" spreadsheet into a third spreadsheet that associates the spot number with a unique identifier and associated genome annotation information for each gene.

Differences in spot intensities between replicate experiments arise from normal experimental variation, such as differences in growth conditions, radioactive nucleotide incorporation efficiency, hybridization conditions, or image acquisition. To compare experimental replicates (separate cultures) or technical replicates (same culture and same RNA sample), the data from each array must be normalized. The third subroutine, "CalcPercentage3", normalizes arrays; if the data are not normalized, then the values for replicate experiments, when plotted, will not pass through zero or be directly proportional (Figure 2). Array experiments can be normalized by expressing each gene-specific spot relative to an internal control, if a suitable set of control spots is present on the arrays. Unfortunately, the intensities of the genomic DNA control spots on the Panorama *E. coli* membranes vary significantly (data not shown) and are not reliable for normalization. In the absence of an internal standard, the preferred approach for normalization is to express each gene-specific spot as a fraction of the sum of all gene spots ($n = 4290$), a strategy that at once considers all variables that lie behind the array image (17,18). The "CalcPercentage3" subroutine normalizes the entire data set, expressing each spot as a percentage of the sum of all spots on the array.

The fourth subroutine, "Cleanup4", reorganizes the data to facilitate subsequent data processing steps. The fifth subroutine, "Statistics5", calculates the averages and statistics for the blank spots (empty spots between genomic DNA control spots) and null spots (empty spots within array). However, we have not found these values to offer a reliable means for establishing background on the array or for empirical determination of a threshold value and therefore do not use them. The "AllDataCrunched6" macro in "Image Data Cruncher" runs all five subroutines at once and results in a file that contains the raw data, normalized data, and associated genome annotation information for each gene on the array.

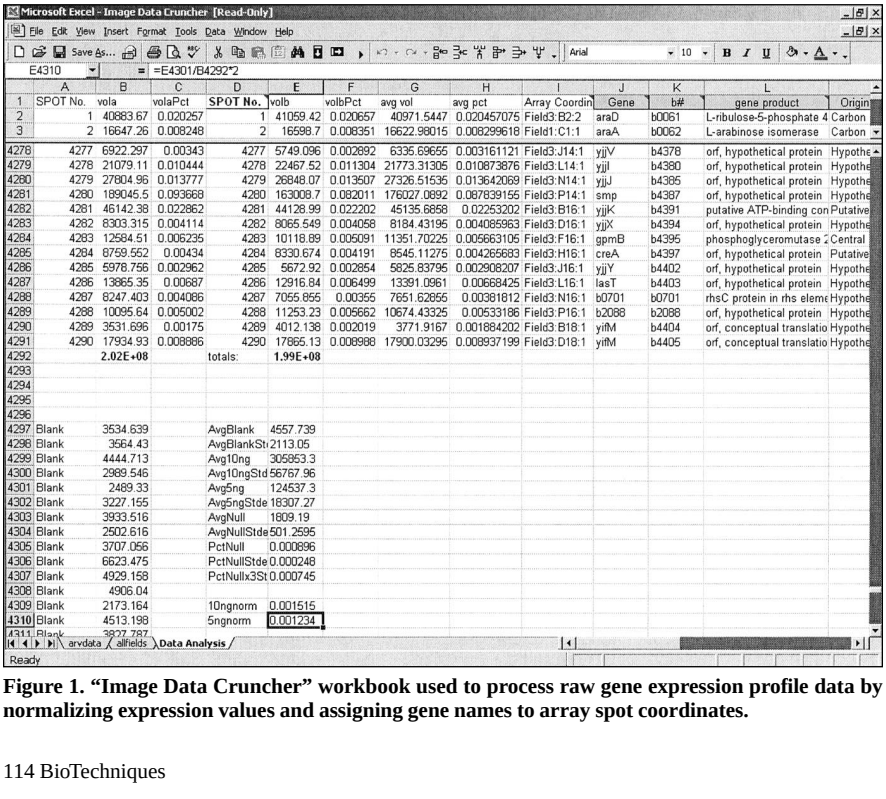



**Figure 1. "Image Data Cruncher" workbook used to process raw gene expression profile data by normalizing expression values and assigning gene names to array spot coordinates.**

### "2-Replicate-Stats" Workbook

The second workbook, "2-Replicate-Stats", contains a macro named "AllAnalysis8" that is used to calculate the probability that the average of the experimental (test) replicates is significantly different from the average of the control replicates (Figure 3). The four replicate data sets are sequentially copied from the crunched data files, beginning with the first and then the second replicate of the control, followed by the first and then the second replicate of the experimental, and pasted into the "Enter Data (2 Replicates)" worksheet. The first subroutine, "OrganizebySpotNo1", sorts the four data sets by spot number, which aligns the gene-specific data in rows. The second subroutine, "CalculateAverages2", calculates the mean of the normalized (percentage) values for the four spot intensities from each experimental condition (two spots for each gene per membrane). The subroutine, "copyvaluesintoPRaw3", copies the percentage values into a separate spreadsheet, and the "CalculateLn4" subroutine transforms the raw percentage values by the natural log while copying them to an additional spreadsheet.

The subroutines "CalculatePRaw5" and "CalculatePLn6" are used to calculate the *P* values for the raw and log transformed data, respectively, by application of the Student's *t* test to the four determinations for each of the control and experimental conditions. The last subroutine, "CopyAllValues7", reorganizes the data by pasting the data

columns to be used in subsequent steps into a separate spreadsheet named "All Values". Once the "AllAnalysis8" macro is executed, the "2-Replicate-Stats" workbook serves as an archive for the raw and normalized data from the replicates being compared and contains the *P* values that are associated with the ratio calculations in the "Data Analysis" workbook.

**"Data Analysis" Workbook**

The third workbook, "Data Analysis", contains two macros. The first macro, "AllAnalysis6", is used to calculate the log ratio of the expression levels in the experimental versus the control condition (Figure 4). The average normalized data from the "All Values" spreadsheet in the "2-Replicate-Stats" workbook are pasted into the "crunched data" spreadsheet in the "Data Analysis" workbook. The first subroutine, "SpotSort1", sorts the data by spot number such that the gene specific data are aligned in rows. The second subroutine, "DataSort2", copies the data and pastes them into the "Data-Analysis" spreadsheet to be used for ratio calculations.

The third subroutine, "ThresholdRatios3", is used to determine a threshold value for ratio calculation and calculates the absolute value of the ratio of the Test/Control such that genes that are more highly expressed in the test condition are given a positive value and genes that are more highly expressed in the control are given a negative value. The threshold value is chosen to represent the limit of detection of an expressed gene (i.e., the signal intensity at which spots are considered to be significantly higher than the array background). Any spot intensity that falls below the threshold value is raised to that value to obtain a reasonable ratio in cases where a gene is expressed below the threshold value in at least one of the two experimental conditions. Ideally, the threshold would be determined independently for each gene, based on the local spot background and the known cross-hybridization to other expressed genes in the sample, but this is not possible because of the dense packing of some membrane arrays and the lack of prior knowledge as to the number and extent of gene expression in a given growth condition. These factors make determination of the threshold value difficult. We have chosen a conservative approximation of the threshold value corresponding to the 500th lowest expressed gene based on the average of the normalized expression levels in the two conditions. This threshold value is similar to that obtained by visual inspection of array images to determine the faintest of gene-specific spots and is reasonable in light of the predicted number of expressed genes based on the number of mRNA species in the *E. coli* cell and the arrangement of genes in operons (1380 mRNAs × average 2.5 genes per operon = 3450 expressed genes) (9). Others have calculated threshold values as three standard deviations above the mean of "blank" spots (10), which corresponds to approximately 214 in the example data set used in this study. If desired, researchers can write alternative approaches for threshold determination into the "ThresholdRatios3" subroutine, or the threshold level can be edited as described on the Web site.

The fourth subroutine, "Cleanup4", reorganizes the data and calculates the log (base 10) of the expression ratio. The fifth subroutine, "Stats5", calculates the standard deviation of the mean of the log ratios. The "AllAnalysis6" subroutine executes all five subroutines and results in a "DataAnalysis" spreadsheet that contains the averaged raw and normalized data, corresponding genome annotation information for each gene on the array, and the ratios. The "AllAnalysis6" macro concludes with creation of two empty columns that are used for manual pasting of the *P* value associated with each ratio calculation from the "2-Replicate-Stats" workbook. Finally, the "MakeDB" macro in the "Data Analysis" workbook can be executed to reorganize and paste key data columns into a separate spreadsheet that can be used for data entry into a suitable database (Figure 5).

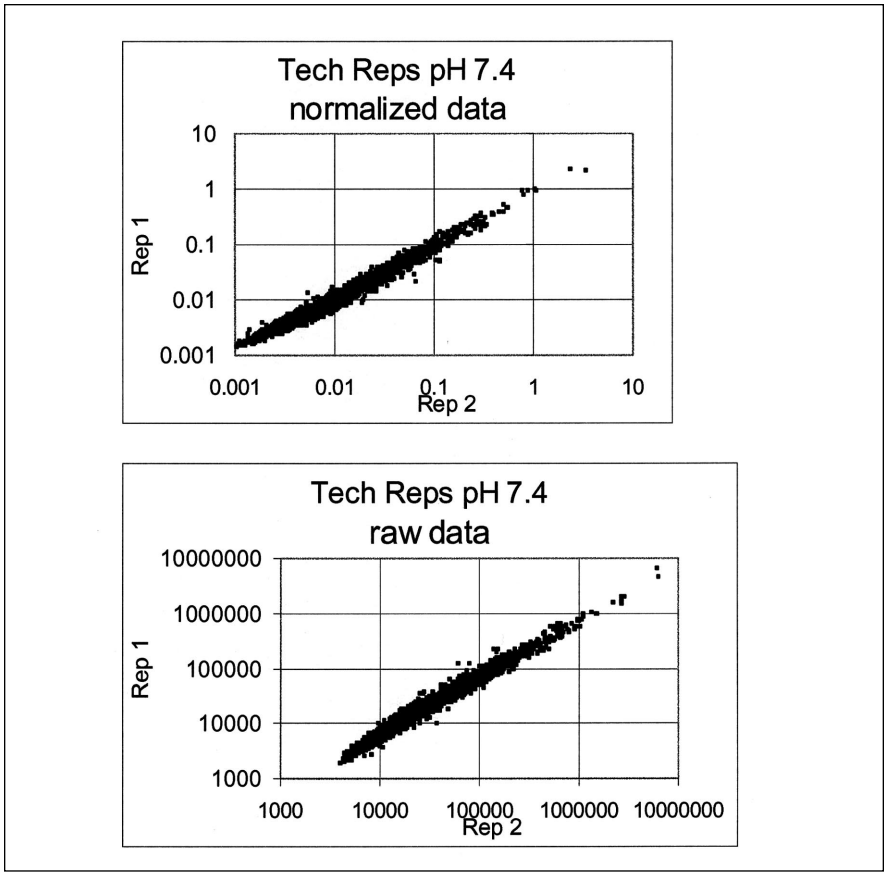
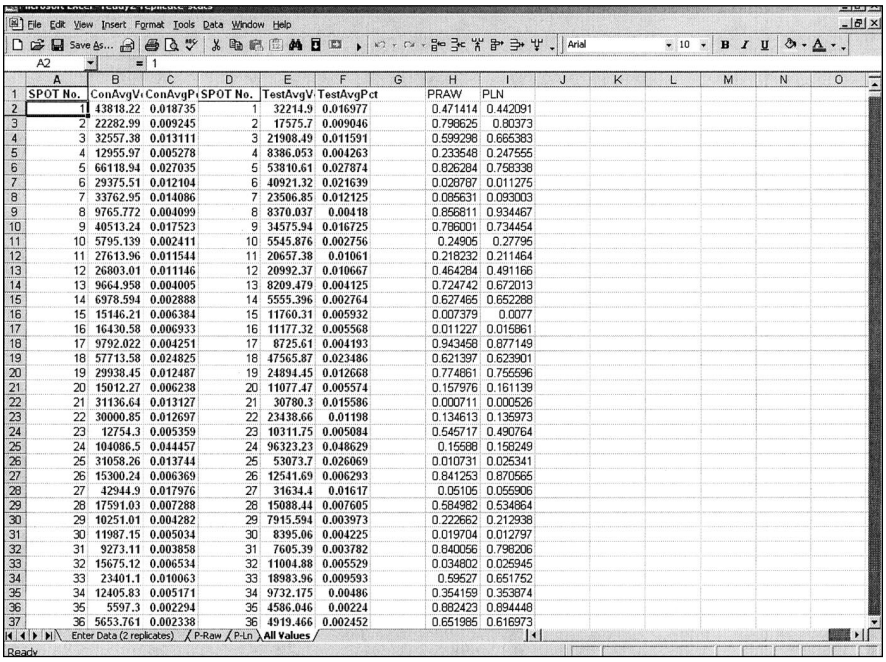



**Figure 2. Scatter plot view of normalized (top) and raw (bottom) data showing linearity of experimental replicates for normalized data.**

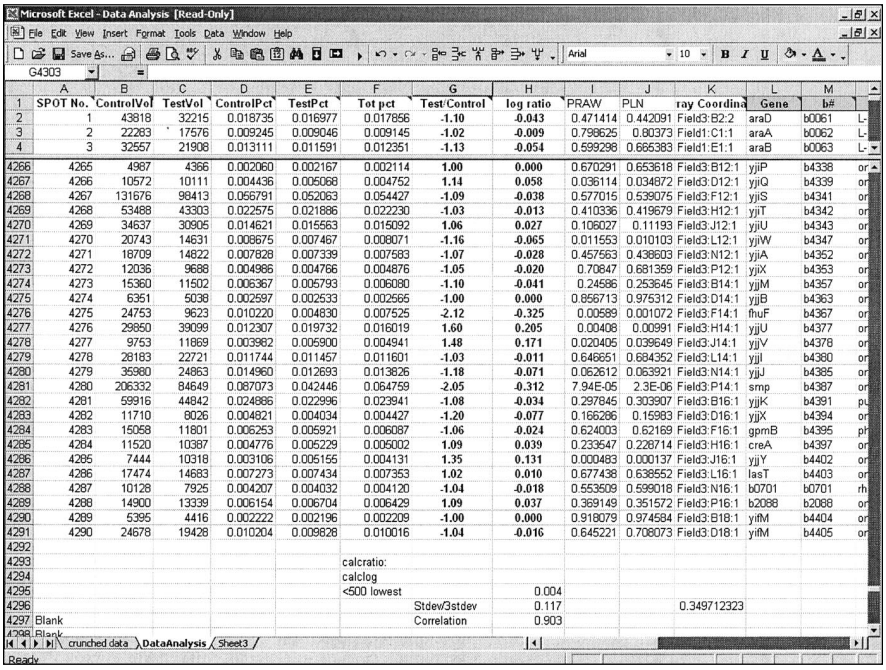
### Database Considerations

It is useful to build a database when multiple time points or multiple conditions are being compared. A sample Access database called "sampleDB" can be downloaded from our Web site. There are two tables in the sample database, one populated with annotation information such as b# (unique identifier), array coordinate, gene, gene product, functional groupings, and accession numbers, and the other table populated with data generated by the "MakeDB" macro in the "Data Analysis" workbook. Additional tables can be created and populated by copying the table (structure only), editing the design as necessary, and pasting the appropriate data from the spreadsheets created by "MakeDB". Queries are designed by linking tables by b# and can be run by manipulation of the parameters in query design. The query data can be copied or exported from Access into third-party presentation software packages. There are limitations on the total number of tables and data elements supported by Access that make it impossible to query more than 20 experiments. The macros and Excel workbooks described here are also suitable for processing data for entry into more robust database programs such as MYSQL or Oracle.



| SPOT No. | ConAvgV | ConAvgP | SPOT No. | TestAvgV | TestAvgPct | PRAW | PLN |
|---|---|---|---|---|---|---|---|
| 1 | 43818.22 | 0.018735 | 1 | 32214.9 | 0.016977 | 0.471414 | 0.442091 |
| 2 | 22282.99 | 0.009245 | 2 | 17575.7 | 0.009046 | 0.798625 | 0.80373 |
| 3 | 32557.38 | 0.013111 | 3 | 21908.49 | 0.011591 | 0.599298 | 0.665383 |
| 4 | 12955.97 | 0.005278 | 4 | 8386.053 | 0.004263 | 0.233548 | 0.247555 |
| 5 | 66118.94 | 0.027035 | 5 | 53810.61 | 0.027874 | 0.826284 | 0.758338 |
| 6 | 29375.51 | 0.012104 | 6 | 40921.32 | 0.021639 | 0.028787 | 0.011275 |
| 7 | 33762.95 | 0.014086 | 7 | 23506.85 | 0.012125 | 0.085631 | 0.093003 |
| 8 | 9765.772 | 0.004099 | 8 | 8370.037 | 0.00418 | 0.856811 | 0.934467 |
| 9 | 40513.24 | 0.017523 | 9 | 34575.94 | 0.016725 | 0.786001 | 0.734454 |
| 10 | 5795.139 | 0.002411 | 10 | 5545.876 | 0.002756 | 0.24905 | 0.27795 |
| 11 | 27613.96 | 0.011544 | 11 | 20657.38 | 0.01061 | 0.218232 | 0.211464 |
| 12 | 26803.01 | 0.011146 | 12 | 20992.37 | 0.010667 | 0.464284 | 0.491166 |
| 13 | 9664.958 | 0.004005 | 13 | 8209.479 | 0.004125 | 0.724742 | 0.672013 |
| 14 | 6978.594 | 0.002888 | 14 | 5555.396 | 0.002764 | 0.627465 | 0.652288 |
| 15 | 15146.21 | 0.006384 | 15 | 11760.31 | 0.005932 | 0.007379 | 0.0077 |
| 16 | 16430.58 | 0.006933 | 16 | 11177.32 | 0.005568 | 0.011227 | 0.015861 |
| 17 | 9792.022 | 0.004251 | 17 | 8725.61 | 0.004193 | 0.943458 | 0.877149 |
| 18 | 57713.58 | 0.024825 | 18 | 47565.87 | 0.023486 | 0.621397 | 0.623901 |
| 19 | 29938.45 | 0.012487 | 19 | 24894.45 | 0.012668 | 0.774861 | 0.755596 |
| 20 | 15012.27 | 0.006238 | 20 | 11077.47 | 0.005574 | 0.157976 | 0.161139 |
| 21 | 31136.64 | 0.013127 | 21 | 30780.3 | 0.015586 | 0.000711 | 0.000526 |
| 22 | 30000.85 | 0.012697 | 22 | 23438.66 | 0.01198 | 0.134613 | 0.135973 |
| 23 | 12754.3 | 0.005359 | 23 | 10311.75 | 0.005084 | 0.545717 | 0.490764 |
| 24 | 104086.5 | 0.044457 | 24 | 96323.23 | 0.048629 | 0.15688 | 0.158249 |
| 25 | 31058.26 | 0.013744 | 25 | 53073.7 | 0.026069 | 0.010731 | 0.025341 |
| 26 | 15300.24 | 0.006369 | 26 | 12541.69 | 0.006293 | 0.841253 | 0.870565 |
| 27 | 42944.9 | 0.017976 | 27 | 31634.4 | 0.01617 | 0.05106 | 0.055906 |
| 28 | 17591.03 | 0.007288 | 28 | 15088.44 | 0.007605 | 0.584982 | 0.534864 |
| 29 | 10251.01 | 0.004282 | 29 | 7915.594 | 0.003973 | 0.222662 | 0.212938 |
| 30 | 11987.15 | 0.005034 | 30 | 8395.06 | 0.004225 | 0.019704 | 0.012797 |
| 31 | 9273.11 | 0.003858 | 31 | 7605.39 | 0.003782 | 0.840056 | 0.798206 |
| 32 | 15675.12 | 0.006534 | 32 | 11004.88 | 0.005529 | 0.034802 | 0.025945 |
| 33 | 23401.1 | 0.010063 | 33 | 18983.96 | 0.009593 | 0.59527 | 0.651752 |
| 34 | 12405.83 | 0.005171 | 34 | 9732.175 | 0.00486 | 0.354159 | 0.353874 |
| 35 | 5597.3 | 0.002294 | 35 | 4586.046 | 0.00224 | 0.882423 | 0.894448 |
| 36 | 5653.761 | 0.002338 | 36 | 4919.466 | 0.002452 | 0.651985 | 0.616973 |

**Figure 3. "2-Replicate-Stats" workbook used to calculate probability by application of the Student's *t* test, based on uncertainty of measurements determined from replicate experiments, that gene expression is significantly different between the conditions analyzed.**



| SPOT No. | ControlVol | TestVol | ControlPct | TestPct | Tot pct | Test/Control | log ratio | PRAW | PLN | ray Coordina | Gene | b# |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 43818 | 32215 | 0.018735 | 0.016977 | 0.017856 | -1.10 | -0.043 | 0.471414 | 0.442091 | Field3:B2:2 | araD | b0061 |
| 2 | 22283 | 17576 | 0.009245 | 0.009046 | 0.009145 | -1.02 | -0.009 | 0.798625 | 0.80373 | Field1:C1:1 | araA | b0062 |
| 3 | 32557 | 21908 | 0.013111 | 0.011591 | 0.012351 | -1.13 | -0.054 | 0.599298 | 0.665383 | Field1:E1:1 | araB | b0063 |
| 4265 | 4987 | 4366 | 0.002060 | 0.002167 | 0.002114 | 1.00 | 0.000 | 0.670291 | 0.653618 | Field3:B12:1 | yjjP | b4338 |
| 4266 | 10572 | 10111 | 0.004436 | 0.005068 | 0.004752 | 1.14 | 0.058 | 0.036114 | 0.034872 | Field3:D12:1 | yjjQ | b4339 |
| 4267 | 131676 | 98413 | 0.056791 | 0.052063 | 0.054427 | -1.09 | -0.038 | 0.577015 | 0.539075 | Field3:F12:1 | yjjS | b4341 |
| 4268 | 53488 | 43303 | 0.022575 | 0.021886 | 0.022230 | -1.03 | -0.013 | 0.410336 | 0.419679 | Field3:H12:1 | yjjT | b4342 |
| 4269 | 34637 | 30905 | 0.014621 | 0.015563 | 0.015092 | 1.06 | 0.027 | 0.106027 | 0.11193 | Field3:J12:1 | yjjU | b4343 |
| 4270 | 20743 | 14631 | 0.008675 | 0.007467 | 0.008071 | -1.16 | -0.065 | 0.011553 | 0.010103 | Field3:L12:1 | yjiW | b4347 |
| 4271 | 16709 | 14822 | 0.007828 | 0.007339 | 0.007583 | -1.07 | -0.028 | 0.457563 | 0.438603 | Field3:N12:1 | yjiA | b4352 |
| 4272 | 12036 | 9688 | 0.004986 | 0.004766 | 0.004876 | -1.05 | -0.020 | 0.70847 | 0.681359 | Field3:P12:1 | yjiX | b4353 |
| 4273 | 15360 | 11502 | 0.006367 | 0.005793 | 0.006080 | -1.10 | -0.041 | 0.24586 | 0.253645 | Field3:B14:1 | yjjM | b4357 |
| 4274 | 6351 | 5038 | 0.002597 | 0.002533 | 0.002565 | -1.00 | 0.000 | 0.856713 | 0.975312 | Field3:D14:1 | yjjB | b4363 |
| 4275 | 24753 | 9623 | 0.010220 | 0.004830 | 0.007525 | -2.12 | -0.325 | 0.00589 | 0.001072 | Field3:F14:1 | fhuF | b4367 |
| 4276 | 29850 | 39099 | 0.012307 | 0.019732 | 0.016019 | 1.60 | 0.205 | 0.00408 | 0.00991 | Field3:H14:1 | yjjU | b4377 |
| 4277 | 9753 | 11869 | 0.003982 | 0.005900 | 0.004941 | 1.48 | 0.171 | 0.020405 | 0.039649 | Field3:J14:1 | yjjV | b4378 |
| 4278 | 28183 | 22721 | 0.011744 | 0.011457 | 0.011601 | -1.03 | -0.011 | 0.646651 | 0.684352 | Field3:L14:1 | yjjI | b4380 |
| 4279 | 35980 | 24863 | 0.014960 | 0.012693 | 0.013826 | -1.18 | -0.071 | 0.062612 | 0.063921 | Field3:N14:1 | yjjJ | b4385 |
| 4280 | 206332 | 84649 | 0.087073 | 0.042446 | 0.064759 | -2.05 | -0.312 | 7.94E-05 | 2.3E-06 | Field3:P14:1 | smp | b4387 |
| 4281 | 59916 | 44842 | 0.024886 | 0.022996 | 0.023941 | -1.08 | -0.034 | 0.297845 | 0.303907 | Field3:B16:1 | yjjK | b4391 |
| 4282 | 11710 | 8026 | 0.004821 | 0.004034 | 0.004427 | -1.20 | -0.077 | 0.166286 | 0.15983 | Field3:D16:1 | yjjX | b4394 |
| 4283 | 15058 | 11801 | 0.006253 | 0.005921 | 0.006087 | -1.06 | -0.024 | 0.624003 | 0.62169 | Field3:F16:1 | gpmB | b4395 |
| 4284 | 11520 | 10387 | 0.004776 | 0.005229 | 0.005002 | 1.09 | 0.039 | 0.233547 | 0.228714 | Field3:H16:1 | creA | b4397 |
| 4285 | 7444 | 10318 | 0.003106 | 0.005155 | 0.004131 | 1.35 | 0.131 | 0.000483 | 0.000137 | Field3:J16:1 | yjjY | b4402 |
| 4286 | 17474 | 14683 | 0.007273 | 0.007434 | 0.007353 | 1.02 | 0.010 | 0.677438 | 0.638552 | Field3:L16:1 | lasT | b4403 |
| 4287 | 10128 | 7925 | 0.004207 | 0.004032 | 0.004120 | -1.04 | -0.018 | 0.553509 | 0.599018 | Field3:N16:1 | b0701 | b0701 |
| 4288 | 14900 | 13339 | 0.006154 | 0.006704 | 0.006429 | 1.09 | 0.037 | 0.369149 | 0.351572 | Field3:P16:1 | b2088 | b2088 |
| 4289 | 5395 | 4416 | 0.002222 | 0.002196 | 0.002209 | -1.00 | 0.000 | 0.918079 | 0.974584 | Field3:B18:1 | yifM | b4404 |
| 4290 | 24678 | 19428 | 0.010204 | 0.009828 | 0.010016 | -1.04 | -0.016 | 0.645221 | 0.708073 | Field3:D18:1 | yifM | b4405 |
| | | | | | calcratio: | | | | | | | |
| | | | | | calclog | | | | | | | |
| | | | | | <500 lowest | | 0.004 | | | | | |
| | | | | | | Stdev3stdev | 0.117 | | | 0.349712323 | | |
| Blank | | | | | | Correlation | 0.903 | | | | | |

**Figure 4. "Data Analysis" workbook used to calculate gene expression ratios between conditions.**

### Modification and Adaptation of Visual Basic Macros

If desired, the macros described above can be modified in any number of ways to streamline the process or to tweak the statistical parameters. This is most easily accomplished using the Microsoft Visual Basic Editor within the Excel workbook environment. For example, the threshold value can be adjusted by simply changing the "cell" location that is used for the ratio calculation, or the four macros described here could be combined in a single workbook containing several spreadsheets. (We have chosen not to do so because this would create a very large, complicated workbook.) More importantly, these macros can be adapted for use with membrane systems other than the one described here. An example of this is the creation of an analogous set of macros for processing of data from Atlas™ Mouse 1.2 arrays (BD Biosciences Clontech, Palo Alto, CA, USA). These macros and associated Excel workbooks are also publicly available on our Web site.

### Web Site and Public Access to Analysis Tools

A Web site has been created for users to download these analysis tools and protocols: http://www.ou.edu/microarray. The protocols on the site describe, in a simplistic way, the step-by-step implementation of the analysis tools.

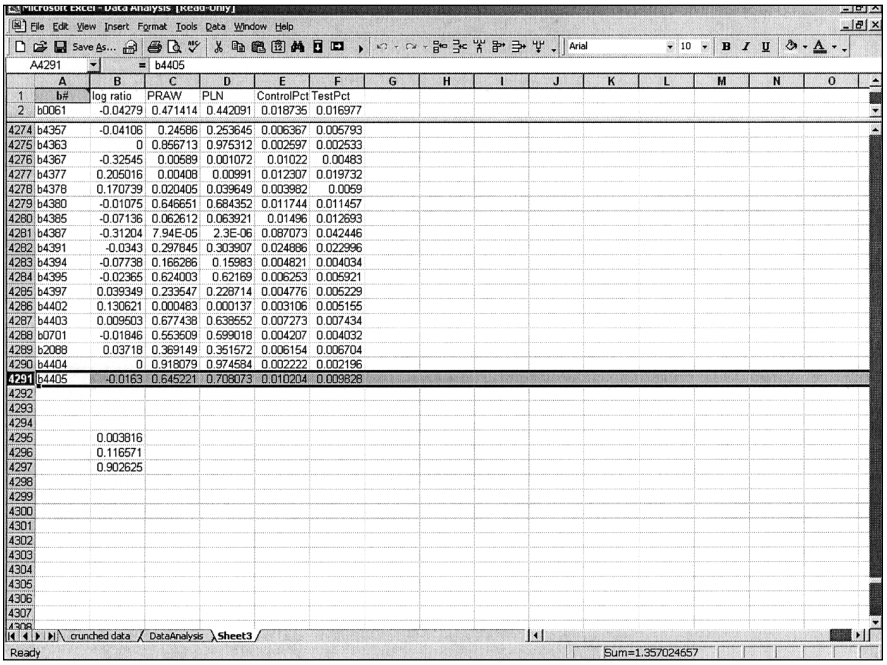



**Figure 5. "MakeDB" worksheet containing selected data columns ready for entry into gene expression experiments database.**

## Benefits of Semi-Automated Data Processing

Our earliest protocol for membrane array data acquisition and analysis required more than 10 man-hours for each experimental replicate (17). Thus, it was important to automate the data analysis process to the greatest extent possible. By taking advantage of improvements in commercially available software and by writing a series of macros in Visual Basic for semi-automated data processing in Microsoft Excel, the time was shortened to 5 min per experimental replicate. In addition, it is critical that the possibility of human error be eliminated when carrying out a large number of manipulations of massive data sets such as those involved in gene expression profiling. The macros described here effectively minimize the level of manual data processing by automating much of the process. The integration of four macros in three workbooks provides for a significant level of human supervision of an otherwise error-free, automated process.

## ACKNOWLEDGMENTS

The authors wish to thank Bill Cuevas and Simon Sims for information before publication and helpful discussions. This work was supported by National Institutes of Health grant nos. to RO1 AI48945-01 and RR-01-005 to T.C.

## REFERENCES

1. **Arfin, S.M., A.D. Long, E.T. Ito, L. Tolleri, M.M. Riehle, E.S. Paegle, and G.W. Hatfield.** 2000. Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor. J. Biol. Chem. *275*:29672-29684.
2. **Barbosa, T.M. and S.B. Levy.** 2000. Differential expression of over 60 chromosomal genes in *Escherichia coli* by constitutive expression of MarA. J. Bacteriol. *182*:3467-3474.
3. **Brown, P.O. and D. Botstein.** 1999. Exploring the new world of the genome with DNA microarrays. Nat. Genet. *21(1 Suppl)*:33-37.
4. **DeRisi, J.L. and V.R. Iyer.** 1999. Genomics and array technology. Curr. Opin. Oncol. *11*:76-79.
5. **Ferea, T.L. and P.O. Brown.** 1999. Observing the living genome. Curr. Opin. Genet. Dev. *9*:715-722.
6. **Khodursky, A.B., B.J. Peter, N.R. Cozzarelli, D. Botstein, P.O. Brown, and C. Yanofsky.** 2000. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. Proc. Natl. Acad. Sci. USA *97*:12170-12175.
7. **Long, A.D., H.J. Mangalam, B.Y. Chan, L. Tolleri, G.W. Hatfield, and P. Baldi.** 2001. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. J. Biol. Chem. *276*:19937-19944.
8. **Neidhardt, F.C., P.L. Bloch, and D.F. Smith.** 1974. Culture medium for enterobacteria. J. Bacteriol. *119*:736-747.
9. **Neidhardt, F.C., J.L. Ingraham, and M. Schaechter.** 1990. Physiology of the Bacterial Cell: A Molecular Approach. Sinauer Associates, Sunderland, MA.
10. **Pomposiello, P.J., M.H. Bennik, and B. Demple.** 2001. Genome-wide transcriptional profiling of the *Escherichia coli* responses to superoxide stress and sodium salicylate. J. Bacteriol. *183*:3890-3902.
11. **Richmond, C.S., J.D. Glasner, R. Mau, H. Jin, and F.R. Blattner.** 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. *27*:3821-3835.
12. **Schena, M., R.A. Heller, T.P. Theriault, K. Konrad, E. Lachenmeier, and R.W. Davis.** 1998. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol. *16*:301-306.
13. **Schena, M., D. Shalon, R.W. Davis, and P.O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*:467-470.
14. **Scherf, U., D.T. Ross, M. Waltham, L.H. Smith, J.K. Lee, L. Tanabe, K.W. Kohn, W.C. Reinhold et al.** 2000. A gene expression database for the molecular pharmacology of cancer. Nat. Genet. *24*:236-244.
15. **Selinger, D.W., K.J. Cheung, R. Mei, E.M. Johansson, C.S. Richmond, F.R. Blattner, D.J. Lockhart, and G.M. Church.** 2000. RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. Nat. Biotechnol. *18*:1262-1268.
16. **Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J.C. Matese, S.S. Dwight, M. Kaloper, S. Weng et al.** 2001. The Stanford Microarray Database. Nucleic Acids Res. *29*:152-155.
17. **Tao, H., C. Bausch, C. Richmond, F.R. Blattner, and T. Conway.** 1999. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J. Bacteriol. *181*:6425-6440.
18. **Tao, H., R. Gonzalez, A. Martinez, M. Rodriguez, L.O. Ingram, J.F. Preston, and K.T. Shanmugam.** 2001. Engineering a homoethanol pathway in *Escherichia coli*: increased glycolytic flux and levels of expression of glycolytic genes during xylose fermentation. J. Bacteriol. *183*:2979-2988.
19. **Wei, Y., J.M. Lee, C. Richmond, F.R. Blattner, J.A. Rafalski, and R.A. LaRossa.** 2001. High-density microarray-mediated gene expression profiling of *Escherichia coli*. J. Bacteriol. *183*:545-556.
20. **Zimmer, D.P., E. Soupene, H.L. Lee, V.F. Wendisch, A.B. Khodursky, B.J. Peter, R.A. Bender, and S. Kustu.** 2000. Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation. Proc. Natl. Acad. Sci. USA *97*:14674-14679.

Received 29 June 2001; accepted 30 August 2001.

**Address correspondence to:**

Dr. Tyrrell Conway
*Department of Botany and Microbiology*
*The University of Oklahoma*
*Norman, OK 73069-0245, USA*
*e-mail: tconway@ou.edu*