[41] Functional Genomics: High-Density Oligonucleotide Arrays

By SASHWATI ROY, SAVITA KHANNA, KIMBERLY BENTLEY, PHIL BEFFREY, and CHANDAN K. SEN

Introduction

The term *functional genomics* can be referred to as the "development and application of a global (genome-wide or system-wide) experimental approach to assess gene function by making use of the information and reagents provided by structural genomics."[1] It is characterized by high-throughput or large-scale experimental methodologies combined with statistical and computational analysis of the results. The fundamental strategy in a functional genomics approach is to expand the scope of biological investigation from studying single genes or proteins to studying all genes or proteins at once in a systematic fashion. Functional genomics promises to rapidly narrow the gap between sequence and function and to yield new insights into the behavior of biological systems.

As the Human Genome Project and related efforts identify and determine the DNA sequences of human genes, it is important that highly reliable and efficient mechanisms be found to assess individual genetic variation. Three methods for obtaining genome-wide mRNA expression data—oligonucleotide "chips,"[2] serial analysis of gene expression (SAGE),[3] and DNA microarrays[4,5]—are particularly powerful in the context of knowing the entire genome sequence (and thus all genes).[6]

Types of DNA Hybridization Arrays

Current array formats can be categorized into the following four groups.

1. *Macroarrays:* Macroarrays rely on robotically spotted probes that have been immobilized on a membrane-based matrix. The probe density is generally

[1] P. Hieter and M. Boguski, *Science* **278,** 601 (1997).

[2] S. P. Fodor, R. P. Rava, X. C. Huang, A. C. Pease, C. P. Holmes, and C. L. Adams, *Nature (London)* **364,** 555 (1993).

[3] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, *Science* **270,** 484 (1995).

[4] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* **93,** 10614 (1996).

[5] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, *Science* **270,** 467 (1995).

[6] V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler, *Cell* **88,** 243 (1997).

0076-6879/02 $35.00

lower on these arrays compared with those of the other three groups. These arrays mostly use radioactive probe labeling. In some cases chemiluminescent labeling has also been described.

2. *Microarrays:* Microarrays use a glass or plastic slide as matrix. These arrays have a higher density of probes compared with macroarrays and use fluorescent labeling-based detection.

3. *High-density oligonucleotide arrays (gene chips):* The probe is generated *in situ* on the surface of the matrix. The leader in these arrays is Affymetrix (Santa Clara, CA) and their combinatorial synthesis method.

4. *Microelectronic arrays:* Microelectronic arrays represent one of the more recent formats of hybridization arrays currently under development by Nanogen (San Diego, CA). Instead of a membrane or a glass slide platform, these arrays consist of a set of electrodes covered by a thin layer of agarose coupled with affinity moiety (permitting biotin–avidin immobilization of probes). Selection and adjustment of proper physical parameters enable rapid DNA transport, site-selective concentration, and accelerated hybridization reactions to be carried out on active microelectronic arrays. These physical parameters include DC current, voltage, solution conductivity, and buffer species. Generally, at any given current and voltage level, the transport or mobility of DNA is inversely proportional to electrolyte or buffer conductivity. The incorporation of controllable electric fields gives a new degree of control over probe deposition and target hybridization.[7,8]

High-Density Oligonucleotide Arrays

The leading arrays in the category of high-density oligonucleotide arrays are manufactured by Affymetrix and utilize the combinatorial synthesis principle.[9] The arrays are designed by using a light-directed chemical synthesis process that creates a series of photolithographic masks to define chip exposure sites, followed by specific chemical synthesis steps. This process constructs high-density arrays of oligonucleotides. Approximately 20 different probe pairs represent each gene on a chip. Each probe pair consists of a perfect match (PM) oligonucleotide probe and a single-base mismatch (MM) oligonucleotide (Fig. 1). The arrays are designed for gene expression as well as single-nucleotide polymorphism (SNP) detection and they cover a large range of different species. The sequence data that Affymetrix uses to build the arrays are downloaded from public databases such as UniGene and GenBank.

[7] C. F. Edman, D. E. Raymond, D. J. Wu, E. Tu, R. G. Sosnowski, W. F. Butler, M. Nerenberg, and M. J. Heller, *Nucleic Acids Res.* **25,** 4907 (1997).

[8] W. M. Freeman, D. J. Robertson, and K. E. Vrana, *Biotechniques* **29,** 1042 (2000).

[9] G. C. Kennedy, *EXS* **89,** 1 (2000).

FIG. 1. Approximately 16–20 different probe pairs represent each gene on a chip. Each probe pair consists of a perfect match (PM) oligonucleotide probe and a single-base mismatch (MM) oligonucleotide.

For the chips to work properly, a sample must be prepared according to Affymetrix protocols. A brief description of the procedures involved in assessing a gene expression profile, using Affymetix GeneChip arrays, is provided.

Sample Preparation

The oligonucleotides on the chip or microarray are called the *probes* and the sample (total RNA or mRNA) that is put on to interrogate is called the *target*. The process is inverted from a traditional Northern analysis.

Total RNA Isolation

RNA is extracted from cells with an RNeasy total RNA isolation kit (Qiagen, Chatsworth, CA). For tissues, RNA is first extracted with TRIzol (Invitrogen, Carlsbad, CA) RNA extraction reagent and then cleaned up with an RNA isolation kit (Qiagen).

cDNA Synthesis

The first strand is synthesized by reverse transcribing the RNA, using the Superscript Choice system (Invitrogen) and oligo(dT)24-anchored T7 primer [high-performance liquid chromatography (HPLC) purified] at 42° for 60 min and then at 70° for 15 min. The second strand is synthesized by using the first-strand synthesis reaction, 5× second-strand buffer, *Escherichia coli* DNA polymerase, and T4 DNA polymerase. The cDNA is isolated according to the Phase Lock gel extraction (Eppendorf, Hamburg, Germany) procedure.

In Vitro Transcription, cRNA Clean-Up, and Fragmentation

Biotinylated RNA is synthesized with an RNA transcript labeling kit (BioArray HighYield: Enzo Diagnostics, Farmingdale, NY). A detailed protocol is provided with the kit.

In Vitro Transcription Clean-Up. Qiagen RNeasy minicolumns are used to clean up the *in vitro* transcription (IVT) cRNA. After the clean-up, cRNA is fragmented with 5× fragmentation buffer (200 m*M* Tris–acetate, pH 8.1; 500 m*M* potassium acetate; 150 m*M* magnesium acetate).

GeneChip: Hybridization, Washing, and Scanning

Further sample processing is mostly automated. The hybridization oven 640 automates the hybridization process for GeneChip probe arrays. The oven provides precise temperature control to ensure successful hybridization, and cartridge rotation to provide continuous mixing. Up to 64 arrays can be processed at one time. The GeneChip fluidics station automates the introduction of the nucleic acid target to the probe array cartridge and controls the delivery of reagents and the timing and temperature for hybridization of nucleic acid target to the probe array. Each fluidics station can independently process four arrays at one time. The probe array nucleic acid target is simply loaded on the fluidics station. Information about the type of array to be analyzed is punched in and the software automatically selects the appropriate protocol. Once processing is complete, messages displayed on the PC and the fluidics station indicate that the probe array is ready for scanning. The GeneArray scanner is from Agilent (Palo Alto, CA) and utilizes a charge-coupled device (CCD) camera and an argon ion laser to excite fluorescent molecules incorporated into the nucleic acid target to generate a quantitative hybridization signal (Fig. 2). With precise optics, the GeneArray scanner focuses the laser on 3-μm spots within each of the thousands of probe features contained on the GeneChip probe array. A high-resolution image of the probe array is displayed in real time during scanning, and fluorescence intensity data are automatically stored in a raw file.

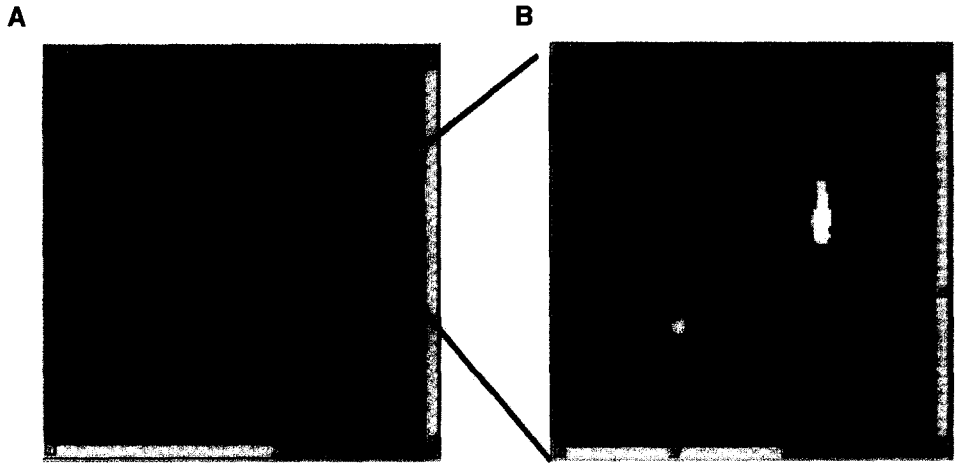

FIG. 2. A representative high-resolution image (A) and a zoomed area of the image (B) of hybridization signals generated by the GeneArray scanner.

Notes

Because chips can be used only once, the target must be tested to ensure that it is of high-enough quality to go onto the expression chip. The test chip serves as one control for the experiment, and the other controls are discussed below. It is critical that test chips be used before the target is put on the gene expression chip (sections below contain example data).

Test Chips: Target Labeling and Hybridization Efficiency Analysis

Once the two targets are prepared, they are each put on a test chip to determine whether they are of high-enough quality to go onto the expression chips. When looking at the data, the control gene names always begin with an AFFX prefix. These are AFFX-Murine BetaActin and AFFX-Murine GAPDH (glyceraldehyde-3-phosphate dehydrogenase) on the murine arrays. Ideally, when comparing 5′ signal with 3′ signal, a 1 : 1 ratio will be seen. Empirically, targets having 5′ : 3′ ratios between 0 and 3 generally give good results, those with ratios between 3 and 4 give marginally good results, and those with ratios >4 give poor results. Thus test chips are used to determine target-labeling quality. Also included are control probe sets that interrogate phage sequences. BioB, BioC, BioD, and Cre are such sequences and are used as hybridization controls. These probe sets are designed to detect the prelabeled oligonucleotides that are contained in the eukaryotic hybridization control kit. This is to ensure that the hybridization, washing, staining, and scanning steps are capable of detecting a broad linear range of labeled cRNA with a high level of sensitivity. Each chip must first be analyzed, using the above-described controls as criteria, before the chips can be compared with one another. After basic analysis, scale factors must be examined between chips and should not vary by 3-fold.

Affymetrix has included controls on the chips so that the data can be quantified and also reproduced. The controls are also used to test different parts of the procedure so that troubleshooting can be performed if necessary. All the genes also act as their own control, with the perfect match and mismatch sequences that are used on the arrays, and the difference in hybridized signal between the probe sets is used to identify nonspecific hybridization and background signal. Affymetrix calls this value the *average difference* (see Data Analysis, below) and it is commonly used as the expression level for the probe set.

For the pilot study, RNA was isolated from normal and treated mice. The Affymetrix protocol was used to prepare the targets. The basic or absolute analysis of chip 1 had a value of 1.9 for the 5′ : 3′ ratio of GAPDH. Furthermore, the control probe sets (BioB, BioC, BioD, and Cre) were also present on the chip. The basic or absolute analysis of chip 2 (RNA from treated mouse) had a value of 2.0 for the 5′ : 3′ ratio of GAPDH. Control probe sets BioB, BioC, BioD, and Cre were also

present on the chip. Both chips passed the specification of the controls, and the last criterion that they must pass is the fold change of the scale factor between them. This is a crucial point because if the criteria are met then the chips can be compared with one another and the differences in gene expression will be analyzed. For this study, the two chips had a fold change of 1.5 for the scale factor, which means that they can be compared with one another.

The second round of analysis is called *comparative analysis*. This analysis uses a control chip that is compared with an experimental chip. In the pilot study, the control chip would be chip 1 and the experimental chip would be chip 3. There are more than 12,000 genes and expressed sequence tags (ESTs) on one murine chip. The first step is to reduce the number of genes, so that the data show only those that have a significant change in either the control chip or the experimental chip. The software suite provided by Affymetrix does this analysis by using different algorithms that compare the two chips. It is important to note that the algorithms tend to be on the conservative side when determining whether a gene is present or not. In the pilot study, the reduced data, once run through the algorithms, showed that 283 genes had a significant difference between the control and experimental chips (Fig. 3). Each gene in a comparison analysis has five potential difference call outcomes: Increase, Marginally Increase, Decrease, Marginally Decrease, and No Change. To reduce the data, the No Change calls are removed, as are those with a fold change of less than 2. The fold change indicates the relative change in the expression levels between the experiment and control targets. Genes that indicate a marginal increase or decrease must be looked at individually to determine whether a definite call can be made. The reduced data can then be easily exported into a Microsoft Excel file for further manipulation.

Data Analysis

To analyze massive amounts of genome-wide data generated by microarray experiments is a challenging task. Gene expression data are useless unless biologically meaningful information can be extracted and presented in some readily comprehensible fashion.[10] The production of this information, involving many facets of image processing, statistical analyses, and data visualization, is possible only with computers powered by sophisticated software. The choice of data analysis strategy should be influenced by the purpose of the microarray experiment and the user's knowledge of the biology of the system under investigation.

Data Mining

The discovery of patterns in gene expression relationships is part of the realm of data mining. Known collectively as *clustering,* these multivariate statistical

[10] J. Quackenbush, *Nat. Rev. Genet.* **2,** 418 (2001).

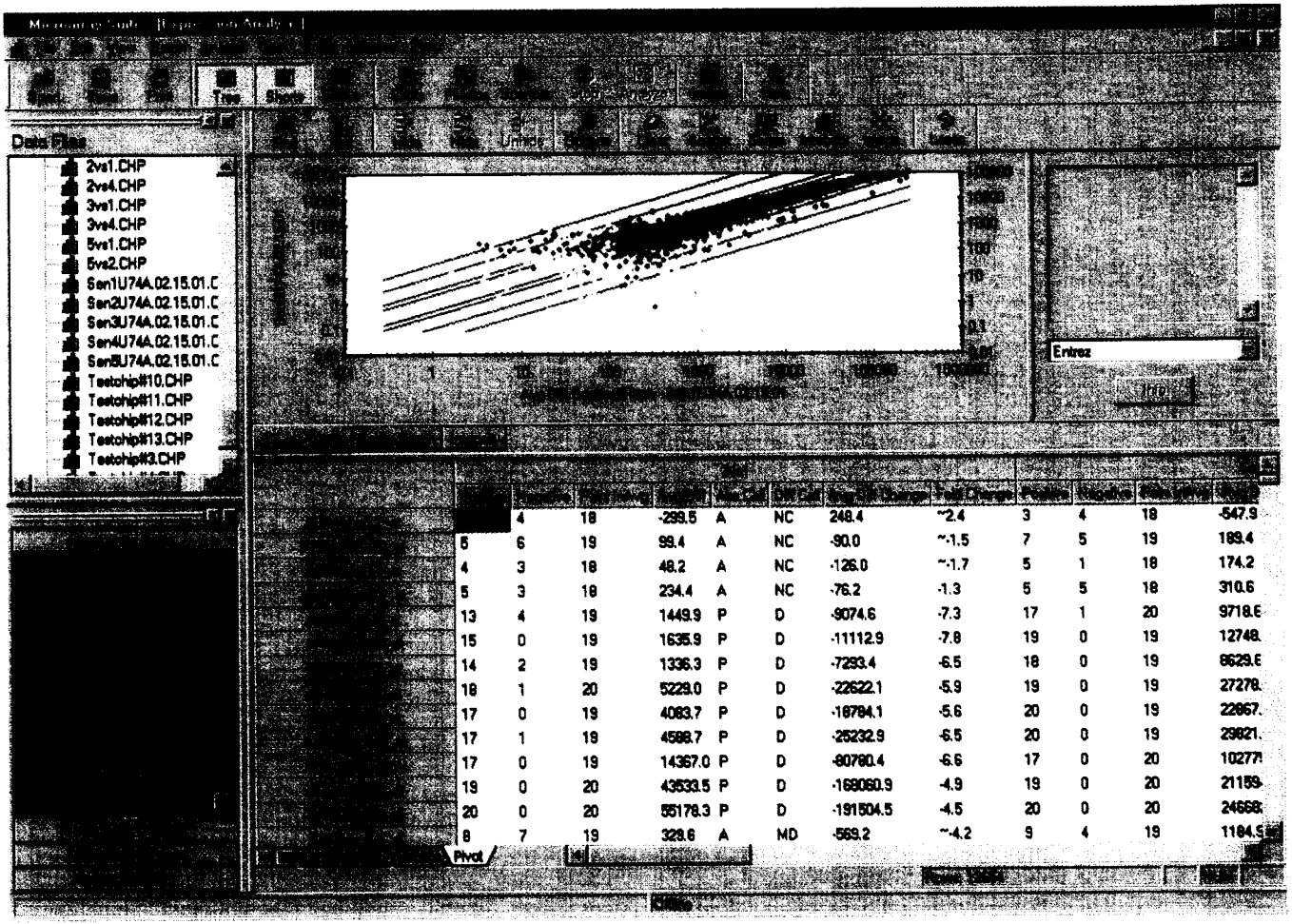

FIG. 3. A representative average difference scatter plot displayed by Microarray Suite (Expression analysis).

methods have become the essential tools for the elucidation of gene expression patterns in microarray data. A number of different methods, for example, *k*-means, self-organizing maps (SOMs), hierarchical clustering, support vector machines, and Bayesian statistics, are employed for clustering analysis.[10] Another useful data-mining method, *principal component analysis* (PCA), is a data reduction technique used to identify uniquely expressed genes. Bioinformatics and data storage are the culmination of the microarray analysis process. Many of the software analysis packages offer immediate access to many public or institutional genetic databases via the Internet.

For the present study, data were analyzed as described below, using Xalysis-Lite (XPROTEIN; Bioinformatics, San Rafael, CA). Three samples, each from a different control animal, were assigned to chips in control group A; likewise, three samples from different treated animals were assigned to chips in experimental group B. The goal to identify candidate genes that vary according to treatment should not be hindered by changes due to a single individual's traits. The analysis of individual probe sets and how they vary between the control group and the experimental group must take this into account. The approach described here is different compared with the "pooling" technique, in which samples are taken from a population of individuals, mixed, and analyzed on a single chip, thereby averaging their expression characteristics. In the present study, rather than mix the samples, a different chip is used for each. Assigning individual samples each to a different chip reduces biological variation because multiple different animals are used, and it reduces variation due to the measuring technology because multiple chips are used. As a result this technique can mask small, but real, expression differences because the source of variation is confounded. It may be coming from the individual animals, or the individual chips, or both. It is advisable, therefore, to design the overall experiment to use as many replicate chips measuring the same sample as time and budget allow.

After processing each chip as described earlier, the overall expression mean and standard deviation of each chip is calculated from each probe set's *average difference* value, one of the standard values output by the Affymetrix software. The highest 2% and lowest 2% of the values are considered outliers and not used in this calculation.

The statistics of experimental groups vary considerably, but because each comes from a different treated individual, all are accepted for further analysis (Table I). More stringent criteria for acceptance can always be adopted later.

Probe Set Analysis Protocol

The average difference value of each probe set is first clamped to zero and normalized by using the mean of its chip (i.e., from Table I). It is then scaled by using the mean of the group means. This calculation smooths out differences between individuals, but retains the overall expression level suggested by the group.

TABLE I
GROUP DATA

Group	Mean	Standard deviation
Control		
A1	29.16	61.92
A2	27.60	59.92
A3	29.74	63.91
Experimental		
B1	18.63	41.15
B2	38.66	81.23
B3	90.25	176.0

The consistency of each probe set's values across all experiments must be carefully evaluated before an effect can be considered real. To that end a two-sample independent t test between the control and experimental groups is applied to each probe set's values (after normalization and scaling) as follows:

$$t = \frac{\bar{A} - \bar{B}}{\left[\frac{\sum A_i^2 - \frac{(\sum A_i)^2}{n_A} + \sum B_i^2 - \frac{(\sum B_i)^2}{n_B}}{(n_A - 1) + (n_B - 1)}\right]^{1/2} \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

In this equation, A_i refers to the individual values of one probe set from the control group and B_i refers to the individual values of the same probe set from the experimental group; n_A and n_B are the total number of experiments in each group; and $\bar{A}$ and $\bar{B}$ are the means. The t value can be used to determine the probability that values from group A belong to the same statistical distribution as values from group B. The probability value implied by t depends on the degrees of freedom of the data:

$$\mathrm{df} = (n_A - 1) + (n_B - 1)$$

Given t and df, the probability can be determined from a standard table of t values. If the probability is small, it is reasonable to assume that the treatment has altered the expression of the probe set in some way. Only those probe sets that pass this test are considered for further analysis (Table II). Regarding Table II, the first probe set illustrates data consistent among the experiments from each group; the second shows data that are highly inconsistent. Because the probability value of the first set is below the chosen threshold of 1% it is considered for future analysis; the second set, with a probability of more than 17%, is rejected.

It is reasonable to rely on this test to avoid the confounding problem because it rejects probe sets that vary in a nonspecific way between experimental groups, whether it be due to biological noise or noise from the technology. It is possible, however, that a real effect is being masked by technology noise and is, therefore,

TABLE II
STATISTICAL ANALYSIS

	Transcript A		Transcript B	
Probe set	Avg. diff.	Normalized	Avg. diff.	Normalized
A1	45.90	45.39	97.20	96.12
A2	39.00	40.75	58.60	61.22
A3	44.70	43.33	78.40	76.00
Mean of A		43.16		77.78
B1	50.20	132.50	37.60	99.24
B2	89.90	114.38	213.70	271.89
B3	207.40	113.02	237.20	129.25
Mean of B		119.96		166.80
t value		−11.96		−1.64
Probability		0.03		17.59
Fold change		2.80		2.1

Abbreviations/symbols: Avg. diff., Average difference.

rejected by this test. That is why it is advisable to run replicate experiments using the exact same sample when possible.

Scoring Candidate Probe Sets

Applying the above described test identified about 300 genes worthy of further examination. Which should be examined first? A quick way to decide is to score the candidate probe sets that passed the *t* test according to their fold change in expression weighted toward higher overall expression values. The mean value of the probe set within each experiment group is used from here on as its expression value within that group.

$$\text{Score} = \frac{\bar{B} + x}{\bar{A} + x} \quad \text{if } \bar{B} \text{ is greater than or equal to } \bar{A}, \text{ or}$$

$$\text{Score} = -\frac{\bar{A} + x}{\bar{B} + x} \quad \text{if } \bar{A} \text{ is greater than } \bar{B}$$

The value of variable x is set equal to the expression level deemed significant by the researcher. The list of candidates is then ordered according to this score. Probe sets near the top of the list are overexpressed in group B compared with group A; probe sets near the bottom of the list are underexpressed in group B compared with group A. Although the *t* value alone can be used to order the data in a similar fashion, the score calculated with this method is similar to the fold change value

for the expression range of interest. Thus, the topmost and bottommost entries in this sorted list provide a convenient starting point for evaluation and further experimentation when examined by the critical eye of the researcher.

Acknowledgments

Supported by NIH GM 27345, the Surgery Wound Healing Research Program, and U.S. Surgical, Tyco Healthcare Group. The Laboratory of Molecular Medicine is the research wing of the Center for Minimally Invasive Surgery.

[42] Reporter Transgenes for Study of Oxidant Stress in *Caenorhabditis elegans*

By CHRISTOPHER D. LINK and CAROLYN J. JOHNSON

Introduction

For many studies of the effects of oxidant stress on cells it can be advantageous to visualize the transcriptional response of the cell *in vivo* in real time. In optically transparent model systems, gene expression can be directly visualized by the construction of reporter transgenes expressing green fluorescent protein (GFP), as originally demonstrated by Chalfie and colleagues.[1] We describe both the general considerations involved in the construction of GFP reporter transgenes responsive to oxidative stress and the specific details of constructing a representative transgenic reporter in the model nematode worm *Caenorhabditis elegans*. Although the details of the representative reporter transgene apply specifically to *C. elegans*, the general approach should be applicable to many model systems.

Identification of Oxidant Stress-Responsive Genes

Construction of oxidative stress-responsive reporter transgenes first requires identification of oxidative stress-responsive genes. Candidate responsive genes can be identified by extrapolation from studies of other systems [e.g., a *C. elegans* GFP reporter transgene based on the small *C. elegans* heat shock protein 16 (HSP16) was found to be responsive to oxidative stress,[2] an unsurprising result considering previous studies of mammalian small heat shock proteins[3]] or from direct gene

[1] M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher, *Science* **263,** 802 (1994).
[2] C. D. Link, J. R. Cypser, C. J. Johnson, and T. E. Johnson, *Cell Stress Chaperones* **4,** 235 (1999).
[3] X. Preville, H. Shultz, U. Knauf, M. Gaestel, and A. P. Arrigo, *J. Cell Biochem.* **69,** 436 (1998).

0076-6879/02 $35.00