CHAPTER ONE

Analysis of Gene Function Using DNA Microarrays

Andrew P. Capaldi

Contents

Abstract

This chapter provides a guide to analyzing gene function using DNA microarrays. First, I discuss the design and interpretation of experiments where gene expression levels in mutant and wild-type strains are compared. I then provide a detailed description of the protocols for isolating mRNA from yeast cells, converting the RNA into dye-labeled cDNA, and hybridizing these samples to a microarray. Finally, I discuss methods for washing, scanning, and analyzing the arrays. Emphasis is placed on describing approaches and techniques that help to minimize the artifacts and noise that so often plague microarray data.

Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, USA

Methods in Enzymology, Volume 470
ISSN 0076-6879, DOI: 10.1016/S0076-6879(10)70001-X

1. Introduction and Experimental Design

DNA microarrays are a powerful tool for studying the function of signaling proteins, transcription factors, and the networks that they comprise. The basic premise of the approach is simple; by analyzing the global gene expression profile of mutant and wild-type (WT) strains it should be possible to deduce the function of any protein or genomic element perturbed. In practice, however, the design, execution, and interpretation of these experiments require strict attention to detail. This is particularly true if the data is to be interpreted at a quantitative level.

1.1. Single-mutant analysis

Perhaps the most straightforward approach to interrogating gene function using microarrays is to compare the mRNA levels in a strain with a gene deleted to those in the WT parental strain. This approach has now been used to analyze the function of hundreds of *Saccharomyces cerevisiae* genes leading to genome-wide maps of signaling pathways, cell cycle control, and the DNA damage response (Hughes *et al.*, 2000; Ideker *et al.*, 2001; Roberts *et al.*, 2000; Workman *et al.*, 2006). This approach has also been applied to other yeast species such as *Candida albicans* and *Saccharomyces pombe*, leading to interesting insights into the evolution of signaling systems and the molecular determinants of pathogenicity (Enjalbert *et al.*, 2006; Smith *et al.*, 2002; Tsong *et al.*, 2006; Tuch *et al.*, 2008). The difficulty with such experiments, however, is that gene deletion often leads to secondary changes that make the microarray results difficult to interpret. For example, if removing the gene for one transcription factor also leads to changes in the expression of several other transcription factors, it will not be possible to draw simple conclusions about the genes regulated by the factor that is knocked out. Secondary effects can also be created through downstream changes in metabolite concentrations or the activity of signaling molecules, and are therefore a potential problem in almost all mutant strains. Computational approaches can be used to dissect out the direct and indirect effects of such deletions (Workman *et al.*, 2006) but, in general, this deconvolution is difficult to achieve.

To avoid confounding secondary effects, microarray experiments should therefore be designed around the conditional removal of gene function. In many cases, this can be achieved simply by growing the cells in conditions where the gene product is inactive and then probing mRNA levels shortly after activation by appropriate stimuli. For example, in a recent study of the Hog1 network, array analysis showed that deletion of network components had little or no effect on gene expression in standard growth conditions (Capaldi *et al.*, 2008). However, when the same strains were examined

shortly after osmotic stress (10–20 min later), dramatic changes were observed in the expression of hundreds of genes. Correlation with genome-wide transcription factor binding data from chromatin immunoprecipitation (ChIP; described in detail in Chapter 4 of this volume) showed that these effects are direct. By contrast, when transcription factor binding data was compared to expression data collected an hour after exposure to osmotic stress, the correlations were weak, demonstrating a buildup of secondary effects.

In those instances where proteins are constitutively active, alternative approaches can be taken to avoid or minimize secondary effects. For example, several studies have taken advantage of analog-sensitive kinase alleles to conditionally block kinase activity (Carroll *et al.*, 2001; Papa *et al.*, 2003; Zaman *et al.*, 2009). Other approaches to conditional perturbation include utilization of inducible promoters, temperature-sensitive alleles, or drug-regulated protein degradation. In cases where conditional perturbations are not possible, secondary effects need to be kept in mind and analysis limited to global effects and correlations.

A highly related approach to examining gene function using DNA microarrays is to measure gene expression in strains with mutations that eliminate sites of posttranslational modification or modify DNA binding sites. Such studies have been used to build detailed models of regulatory mechanisms (Leber *et al.*, 2004; Springer *et al.*, 2003; Wang *et al.*, 2004), but the same caveats apply. If the mutation leads to constitutive changes or the expression changes are examined long after activating conditions are applied, secondary effects can complicate the interpretation.

In either gene deletion or gene mutation experiments, the resulting data can be examined at several levels. First, it is usually possible to gain insight into the function of the element perturbed by examining the biological role of the genes up- or downregulated in the mutant; for example, by looking for gene ontology (GO) terms enriched in the regulated gene-set (Chu *et al.*, 1998; DeRisi *et al.*, 1997). Second, it is often possible to identify the pathway(s) and/or proteins that are affected by your mutation by looking for correlations with previous datasets. As proteins in the same or interacting pathways regulate highly overlapping gene-sets, comparison with previously acquired KO data (e.g., the Hughes compendium; Hughes *et al.*, 2000) can be highly informative (Marion *et al.*, 2004; Segal *et al.*, 2003). It is also possible to indentify transcription factors regulated by, or that interact with, the gene under study by looking for significant overlap with target genes identified using genome-wide ChIP analysis (Bar-Joseph *et al.*, 2003). Finally, the transcription factors regulated directly or indirectly by the genetic element under investigation can be identified using motif analysis (Beer and Tavazoie, 2004; Roth *et al.*, 1998; Wang *et al.*, 2005). Further details of the computational methods used to perform such analyses are provided in Chapter 2 of this volume.

1.2. Double-mutant analysis

While much can be gained by analyzing a strain with a single mutation/deletion under a single condition, the full power of microarray analysis is only realized when multiple related experiments are compared. By examining expression in a range of conditions it is possible to determine when a protein or pathway of interest is activated and how its function depends on signal level and/or type. Furthermore, by examining expression in a variety of strains, a quantitative genome-wide interaction map can be constructed. This approach is particularly powerful when double mutants are examined, as described below.

A direct or indirect interaction between two factors, A and B, can be inferred when microarray data reveal a significant overlap in the gene-sets that A and B regulate. However, such data cannot be used to determine if the factors act independently, cooperatively, or partially cooperatively to regulate these genes (Fig. 1.1A). To complicate matters further the interaction type may vary from gene to gene. Therefore, to distinguish between these mechanisms, gene expression data in the double-mutant strain must also be examined. If the factors A and B act cooperatively to regulate a gene, the expression defect in the single- (AΔ and BΔ) and double (AΔBΔ)-mutant strains will be identical (Fig. 1.1B, middle panel). By contrast, if the factors act independently, the defect in the double mutant will be the sum of the defects in the single mutants (Fig. 1.1B, top panel). In cases where the interaction is partially cooperative, the expression defect in the double mutant will be somewhere in-between the values expected for a fully independent or fully cooperative interaction (Fig. 1.1B, bottom panel).

The strength of double-mutant analysis is the ability to determine how the interaction of any two proteins (or mutations) affects each and every gene in the genome. This analysis not only makes it possible to build detailed network or circuit diagrams but also provides clues as to the precise mechanism of the identified interaction. For example, where all or most of the genes regulated by factors A and B are influenced by a cooperative interaction, it is highly likely that the interaction occurs at the signaling level. By contrast, where only a subset of genes depends on a cooperative interaction, it is more likely that the interaction occurs at the level of transcription factor activation or through another downstream affector. Importantly, however, the ability to build such detailed models is limited by the noise in the microarray analysis; noise that is compounded when single and double mutants are compared. Therefore, multiple measurements need to be made and statistics applied to distinguish between the possible regulatory mechanisms (Fig. 1.1) at each gene. To allow such error analysis, it is critical to first break down the data describing the interaction at each gene into its fundamental components. Following the example in Fig. 1.1,

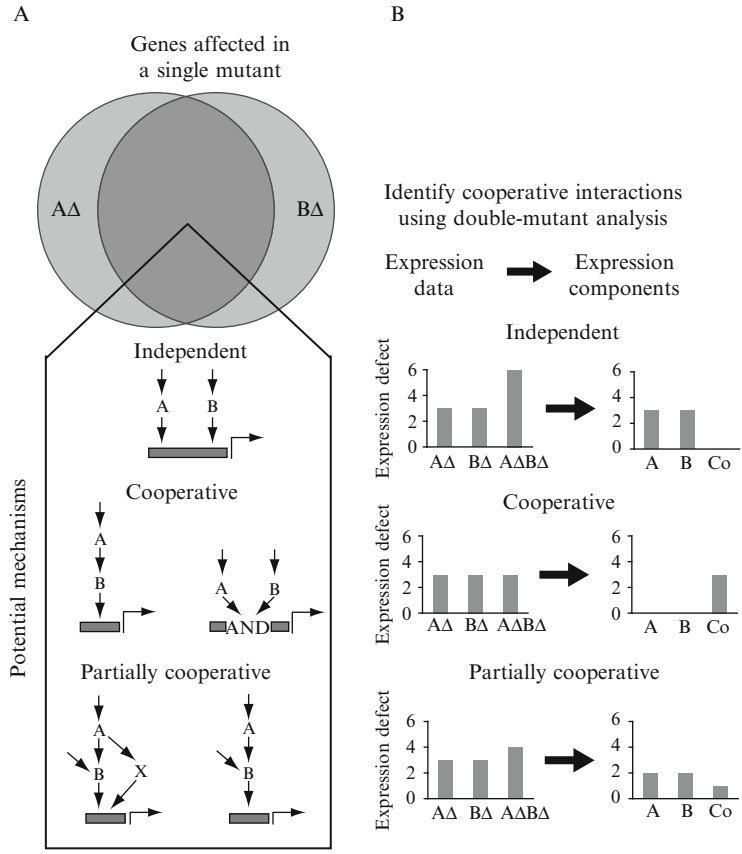

Figure 1.1 Single- and double-mutant analysis of gene expression. (A) Venn diagram summarizing the overlap in genes with a significant defect in gene expression due to deletion of gene A (AΔ) or gene B (BΔ). The wiring diagrams indicate the possible ways factors A and B can interact to regulate expression of overlapping sets of genes. (B) Schematic illustrating the application of the double-mutant approach to analyzing transcriptional network structure and function. The bar graphs on the left show the defects expected in AΔ, BΔ, and AΔBΔ strains for each of the three sample mechanisms. The bar graphs on the right show the values of the three expression components (A, B, and Co) determined by fitting the expression data for the AΔ, BΔ, and AΔBΔ strains (see the text for details).

these components are the induction/repression from A alone, the induction/repression from B alone, and the influence that the interaction between A and B has on expression (or the cooperative component; Fig. 1.1B). The values of these components can be determined simply by comparing the expression defects in microarrays examining the single and double mutants. The array comparing expression in AΔ to the WT reports the value of A + Co for each gene, the array comparing expression in BΔ

to the WT reports the value of B + Co for each gene, while the data for the double mutant AΔBΔ compared to the WT reports the value of A + B + Co. In this case, the error for each expression component for each gene can be estimated by propagating the errors through the calculation and using a simple *t*-test (using log expression values). Once this is done it is possible to cluster data based on the type of interaction (pattern of significant A, B, and Co components) and identify groups of genes regulated by a common mechanism. The double-mutant approach, its application, and the analysis of the associated errors are described in more detail in Capaldi *et al.* (2008). Statistical analysis using these or related methods can be carried out using the free software package R (http://www.r-project.org) or MATLAB (http://www.mathworks.com/).

2. Methods

Once an experiment or series of experiments examining gene function is outlined, it must be translated into a detailed procedure designed to measure mRNA levels while limiting noise (biological or otherwise) in the data. In all cases this means using two color DNA microarrays. The outline of the procedure (developed by DeRisi *et al.*, 1997) is as follows: Cells are grown under the appropriate conditions and mRNA is extracted and purified. The mRNA is then converted into cDNA using reverse transcription and labeled with one of two fluorophores (Cy3 or Cy5). Two cDNA samples, an experimental sample labeled with Cy5 and a control labeled with Cy3, are then hybridized to an array consisting of thousands of different DNA fragments spotted onto a glass slide, where each fragment is complimentary to a single gene. These arrays are then washed to remove any cDNA that binds nonspecifically and analyzed using a laser scanner to measure the Cy5 and Cy3 fluorescence. Finally, after data normalization, the ratio of Cy5 to Cy3 fluorescence at each spot is calculated and used to determine the difference in the mRNA expression levels in the two samples.

2.1. Experimental design

As each microarray compares the expression levels in the two samples, the first step in designing a microarray procedure is deciding which samples will be compared on a given array. The idea is to accurately measure the parameters of interest using the smallest total number of arrays. For large experiments, identifying the best experimental design can be complicated (Kerr and Churchill, 2001), but for more limited experiments some simple principles apply. In most cases where a mutant is being analyzed, it is best to directly compare the WT and mutant strains, grown under the conditions of

interest, on the same array. This way the influence that the mutant has on gene expression is determined with the errors from a single array. An alternative approach is to measure gene induction after stimulus in the WT strain on one array (e.g., WT + stress versus WT in no stress) and the gene induction after stimulus in the mutant strain (e.g., AΔ + stress versus AΔ in no stress) on a separate array. In this commonly used scheme the influence that a mutant has on gene expression can only be calculated by dividing the values from the two arrays and thus the errors from each individual array are multiplied. However, in the case where mutant effects are measured on a single array, it is still important to examine expression in the WT strain (e.g., WT + stress versus WT in no stress) to ensure that the genes influenced by mutation are also regulated in the WT background. Where double mutants are examined it is best to compare the expression levels directly to the single mutant(s) on the same array. This reduces the magnitude of the change in gene expression seen on the array and thus the overall noise as its major component is proportional to signal.

2.2. Cell growth

Once the experiments are designed, cells need to be grown and harvested carefully to limit unwanted sample-to-sample variation. First, strains that are going to be compared on a single array should be grown at the same time and in identical medium as variation in nutrient levels, temperature, and other parameters can introduce substantial noise. Second, the strains should grow for at least two doublings to wash out any differences in the overnight cultures. Third, cells should be harvested at the same optical density (OD) and this OD should be selected so that the cells are approximately one doubling away from any transition induced by nutrient depletion. For example, if cells are studied in log growth phase they should be harvested at a density substantially below that of the diaxuic shift to ensure that small variations in cell number do not affect the gene expression pattern (for *S. cerevisiae* $OD_{600} = 0.6$ works well). Finally, it is best to harvest cells by filtration using a 0.2 μm 90 mm filters (Millipore). This filter is then rolled, placed in a 50 ml conical tube and submerged in liquid nitrogen. This ensures rapid harvesting (<1 min) and little sample-to-sample variation. For the protocol described below it is best to harvest between 75 and 150 OD_{600}/ml units. Once harvested the cells can be stored at −80 °C for several weeks before the RNA is extracted.

2.3. Total RNA isolation and purification

To extract the mRNA from the frozen cells, first add 12 ml of AE buffer (50 m*M* sodium acetate (pH 5.2) and 10 m*M* EDTA) to the 50 ml conical tube and then rotate/shake the tube to remove the cells from the filter.

When this is done for each of the tubes in the set (no more than eight at a time, but always prepare samples to be compared on the same array at the same time), transfer the cells and buffer to a 50 ml centrifuge tube and add 800 μl of 25% (w/v) SDS to each of the samples. Finally, add 12 ml of 65 °C acid phenol (Fluka #77608, note that this is not standard buffer-saturated phenol) to each tube and incubate for 10 min in a 65 °C water bath, vortexing every 30 s or so. When this procedure is finished, cool the samples on ice for 5 min and then centrifuge for 20 min at 12,000×*g* and 4 °C. Carefully decant the phenol–buffer mix from these tubes into prespun (5 min at 1500 rpm) phase lock tubes (5-Prime) and then add 13 ml of chloroform to each tube, mix vigorously, and spin at 3000 rpm for 10 min. After centrifugation the RNA will be in the 10–12 ml aqueous layer at the top of the tube, separated from the organic phase by the gel in the phase lock tubes. Pour this solution into a clean centrifuge tube and add 1 ml of 3 *M* sodium acetate (pH 5.2) and 10 ml of isopropanol; mix by inverting the tube several times and spin for 45 min at 17,000×*g* and 10 °C. At this point the total RNA will be present in an approximately 1 cm white pellet. Carefully decant the isopropanol and add 10 ml of 70% ethanol to each tube, without disturbing the pellet, and spin again at 17,000×*g* and 10 °C but this time for 20 min. Decant as much of the ethanol as possible and then spin the tube at 17,000×*g* again for 1 min to collect any remaining liquid at the bottom of the tube and remove it carefully with a pipette. Finally, let these pellets dry on the bench until they are translucent (30–60 min).

Once the total RNA pellets are dry, resuspend them in 800 μl of RNase-free water and measure the absorbance at 260 and 280 nm. The sample should have >2 mg of RNA and an $A_{260}/A_{280} > 2.0$. It is also useful, especially in the first few RNA preparations or if problems are encountered, to check the integrity of the RNA sample using an Agilent Bioanalyzer or a similar device (agarose gels are only useful for detecting severe degradation). Here a good quality sample should have distinct rRNA and tRNA bands with well-defined edges. If the sample fails in any of the quality controls it should be discarded.

2.4. Purification of poly-A RNA

To isolate mRNA from the total RNA sample, cellulose resin with a poly-deoxythimidine oligomer attached (oligo-dT cellulose) is used to purify transcripts with a poly-A tail. This purification should be done on the same day as the total RNA purification to limit degradation. First, wash 60 mg of cellulose resin three times with 750 μl of NETS buffer (0.6 *M* NaCl, 10 m*M* EDTA, 10 m*M* Tris–HCl (pH 8.0), 2% (w/v) SDS) in a 2 ml screw cap tube. Here the resin should be spun at 3000 rpm for 1 min on a benchtop centrifuge between washes and the buffer removed by aspiration. At this stage, incubate 750 μl of 2–4 mg total RNA at 65 °C for 10 min, and then

add to a 2 ml tube along with 750 μl of 2× NETS buffer also at 65 °C. Each tube should then be left to mix on a rotator for 1 h at room temperature. After incubation, apply this sample to a disposable column (BioRad #732-6008) that has been washed once with NETS buffer. Once the resin has settled in the column, wash it three times with 750 μl NETS buffer and then elute the mRNA with 650 μl of 65 °C ETS buffer (NETS buffer without the NaCl) by injecting it directly into the column bed. Finally, add 65 μl of 3 *M* sodium acetate and 650 μl of isopropanol to these tubes, mix well by inversion, and incubate at −20 °C overnight. The next morning spin the sample at full speed in a benchtop centrifuge for 1 h at 4 °C. When the spin is complete, remove the isopropanol buffer mix from the tube and add 250 μl of 70% ethanol to the sample, taking care not to disturb the pellet, and spin at full speed for 20 min at room temperature. Carefully remove the ethanol from these samples and allow them to air dry completely (residual ethanol will inhibit the reverse transcription in the next step). When dry, the pellets will become white and powdery. Resuspend these pellets in 20 μl of RNase-free water and then spin for 1 min at full speed to remove the cellulose fragments that were not trapped by the column. Remove the supernatant and measure the absorbance and A_{260}/A_{280} ratio. This is best done on a nanodrop spectrophotometer (Thermo Scientific) due to the small volume. The yield should be between around 20 μg and the A_{260}/A_{280} ratio again greater than 2.0.

2.5. Reverse transcription and dye labeling

On the same day as the mRNA is purified it should be converted into cDNA by reverse transcription. Combine four micrograms of poly-A RNA with 5 μg of an oligo-dT primer (T_{20}) and 5 μg of a random primer (N_9) in a total volume of 15.5 μl and incubate at 70 °C for 8 min before cooling on ice. Once cool, perform cDNA synthesis using AffinityScript reverse transcriptase (Stratagene) by adding the RNA and primer mix to 2 μl of enzyme, 3 μl of AffinityScript buffer, 3 μl of 100 m*M* DTT, 5.9 μl of water, and 0.6 μl of 50× aa-dNTP mix and incubate at 42 °C for 2 h. Here the aa-dNTP mix is made up of 1 mg of aminoallyl-dUTP, 20 μl of water, 30 μl of 100 m*M* dTTP, and 50 μl of 100 m*M* A, C, and GTP (store this nucleotide mix at −20 °C in single use aliquots). This reaction will result in a cDNA library where approximately 1/10 bases have a free amine group that can be labeled by Cy5 or Cy3. To degrade the RNA template, add 4 μl of 1 *M* NaOH and 8 μl of 50 m*M* EDTA and incubate at 65 °C for 10 min. Finally, neutralize the solution using 40 μl of 1 *M* HEPES, pH 7.0. The cDNA can then be purified using a Clean and Concentrator-5 Kit (Zymo Research) following the manufacturer's instructions except that the cDNA–HEPES buffer mix should be mixed with 1 ml of binding buffer before applying it to the column. Elute the DNA from the column in 12 μl

of water and then determine the yield and A_{260}/A_{280} ratio using the nanodrop spectrophotometer; they should be ≥ 2 μg and 1.8, respectively. The cDNA samples can be stored at -20 °C for many weeks before labeling and hybridization.

Once the cDNA is synthesized it needs to be labeled with either Cy5 (typically the sample) or Cy3 (typically the control). To do this, add 1 μl of 1 *M* sodium bicarbonate buffer (pH 9.0) to 10 μl of the cDNA solution and add 1 μl of *N*-hydroxysuccinimidyl ester Cy5 or Cy3 (GE Biosciences) that has been resuspended in DMSO and incubate at room temperature for 4 h. Each Cy dye pack has enough dye to label 4–8 samples. After labeling, purify the samples using the Clean and Concentrator-5 kit, following the manufacturer's instructions, and again measure the concentration and A_{260}/A_{280} ratio using the nanodrop spectrophotometer. The purified and labeled cDNA should have visible color after labeling. If necessary, these samples can be snap frozen in liquid nitrogen and stored at -80 °C.

2.6. Hybridization

The labeled cDNA samples are now ready to hybridize to a DNA microarray. Many types of microarrays are available for gene expression analysis and each has its advantages and disadvantages. The most common of these are the printed arrays where PCR products or DNA oligomers are spotted onto a polylysine-coated slide and commercial arrays from companies such as Agilent and Roche Nimblegen where oligomers are synthesized directly on the slide. The advantage of printed arrays are that, when thousands of arrays are used, the cost can be as low as $20 per array. However, printed arrays have several distinct disadvantages. First, there tends to be significant variation in the quality and size of the DNA spots. This makes it difficult to accurately determining the expression ratio for some genes and contributes substantially to replicate variation. Second, these arrays have to be post-processed to neutralize the otherwise highly charged lysine surface. This postprocessing leads to imperfections on the surface of the slide and substantial background variation. Finally, the polylysine coated surface is delicate and is often damaged in the printing and hybridization procedure leading to further background noise. By contrast, the commercial arrays are printed with high accuracy and on more stable substrates, resulting in very low background noise. Moreover, the stability of the slides surface means that the sample applied to these arrays can be vigorously mixed during hybridization. This dramatically increases the signal-to-noise ratio and means that far less cDNA needs to be applied to the array during hybridization. The cost of such arrays is presently greater than $100 a piece, but this continues to fall. Here, I will describe hybridization to Agilent arrays (see the manufactures website for further details), but the same samples can be hybridized to arrays from other companies following

the manufacturer's instructions, or to printed arrays using protocols from Derisi Lab (DeRisi *et al.*, 1997).

The eight array format from Agilent (eight, 15,000 spot arrays per slide) works well when 100 ng of Cy5- and 100 ng of Cy3-labeled cDNA is hybridized to each array. To do this, bring 125 ng of Cy5- and 125 ng of Cy3-labeled sample to 25 μl total volume. Heat this sample to 98 °C for 2 min. Cool the sample by centrifugation for 1 min and then add 25 μl of 2× Hi-RPM hybridization buffer (Agilent) to the sample. Carefully apply 40 μl of the sample to the gasket slide positioned in the base of SureHyb chamber, repeat seven more times with different samples to fill each position, and place the microarray face down onto this gasket slide. Add the top to the hybridization chamber, tighten the screw, and then rotate the chamber slowly to ensure that the large bubble in the chamber moves freely (this bubble is critical for mixing during the hybridization) and that no smaller bubbles are stuck to the array surface. If any bubbles remain fixed to the array, gently tap the chamber on a hard surface until they are dislodged and then place in the rotating oven (Agilent) at 65 °C for 17 h.

2.7. Microarray washing

Once the sample is hybridized to the array, excess sample and cDNA that is bound nonspecifically must be washed off and the array dried. Listed here is a protocol that works well for Agilent arrays; those working with other types of arrays should select the appropriate alternative protocol.

Fill two slide staining dishes with wash buffer I (6× SSPE, 0.005% *N*-Lauryl sarcosine where SSPE is 150 m*M* NaCl, 10 m*M* sodium phosphate, and 1 m*M* ETDA, pH 7.4), a third chamber with wash buffer II (0.06× SSPE, 0.005% *N*-Lauryl sarcosine), and a fourth chamber with ozone protection and drying solution (Agilent). Place a slide rack in the second chamber and set chambers 2–4 up on stir plates at a medium setting, ensuring that the stir bars used are small enough to remain below the bottom of the slide rack. Submerge the first array in chamber 1 and pry it open using plastic tweezers so that the gasket slide falls away. Gently move the array from side to side while submerged in the chamber to remove any bubbles from the surface and then place it in the rack in chamber 2 (only handle the label on the array). Repeat this process until all the slides are in chamber 2 and then leave the arrays in this low-stringency wash for 1 min. At this stage transfer the entire slide rack to chamber 3, ensuring that the slides spend minimal time exposed to the air and that little of the wash buffer I is transferred into this new chamber. The slides should then be allowed to sit in this high-stringency wash for exactly 1 min before transferring the rack to the drying solution for 30 s. At this time slowly lift the rack out of the chamber, ensuring that droplets do not form on the surface of the array. Now the array is dry and ready to scan.

2.8. Array scanning

To quantify the amount of Cy5- and Cy3-labeled DNA hybridized to each probe (spot on the array), the washed array is analyzed with a laser scanner that excites both Cy5 (at 635 nm) and Cy3 (at 532 nm) and then measures the emitted light at the appropriate wavelength (570 and 670 nm, respectively) using a photomultiplier. It is best to perform the scan immediately after washing the arrays, but arrays can be stored in a dry ozone-free area for a day or more before scanning if necessary. The most common scanners are the Axon 4000B from Molecular Devices and the DNA microarray Scanner from Agilent.

To get the best quality data from the arrays, the scanner needs to be set up appropriately. First, the lasers should be focused onto the surface of the slide that is spotted with the probes. This focal plane is easily identified as the position where a scan gives the highest overall signal. Next, keeping the laser power at 100%, the voltage applied to the photomultiplier tube should be set to ensure that the highest Cy5 and Cy3 signals measured are just below the maximum of the digitization range (65,536 for a 16-bit A to D converter). This ensures the highest possible signal-to-noise ratio without losing data at some probes due to saturation. Often it is necessary to scan arrays several times to find such settings. This does not present a problem as the Cy dyes are photostable.

The resulting image should show clear spots, each with little pixel-to-pixel variation in the Cy5 and Cy3 ratio, surrounded by a uniform background with low signal (40–50 units on 65,000 unit scale, Fig. 1.2A). Poor quality array images generally indicate a problem with sample labeling or with the wash and hybridization steps. If the signal/noise is uniformly low in a single color, the problem is likely to be poor labeling. This is often due to degraded dye. If the signal/noise is poor in both colors (Fig. 1.2B and C), the problem likely lies in the wash (low signal; overly stringent washing, high background; poor washing). However, poor labeling in both channels can lead to similar problems. Large regions without signal can be caused by bubbles on the surface of the array or by leakage during hybridization (Fig. 1.2D). Finally, speckles on the array are often caused by dust or precipitate in the hybridization or wash buffers. Array images should be saved as a high-resolution tif image file for further analysis

2.9. Gridding and normalization

Once a microarray image has been collected the precise position and identity of each probe (or "spot") must be identified. This is done through a process known as gridding. First a grid file must be assembled; this establishes the identity and expected location of each spot on the array. The grid is then overlaid onto the array and any variation in spot location or

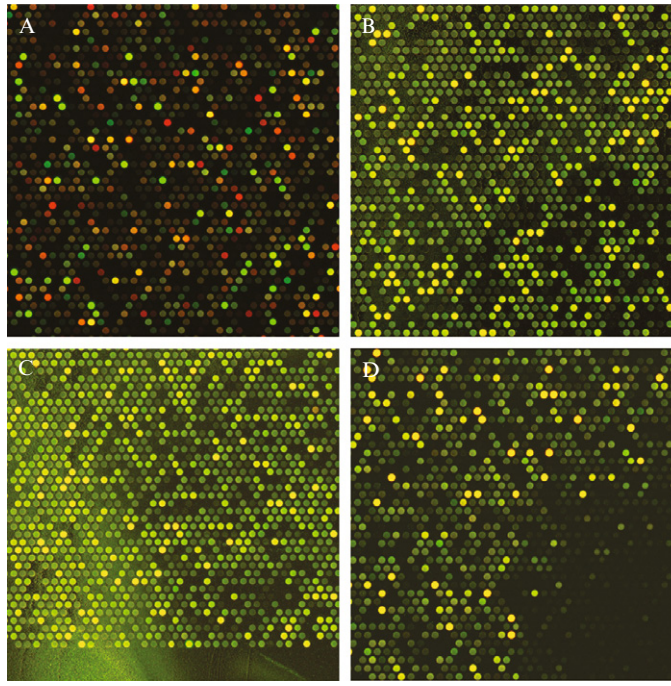

Figure 1.2 High-resolution DNA microarray images showing common hybridization artifacts. (A) A high-quality array as defined by a high signal-to-noise ratio, a low (background level) signal at the negative control spots, and an absence of washing or hybridization artifacts. (B) A poorly washed array showing the characteristic high background signal. (C) An array with variable background signal caused by cDNA precipitation and poor washing. Such precipitation is often caused by loading too much cDNA onto the chip. (D) An array with nonuniform hybridization due to leakage from the hybridization chamber.

size adjusted for through a probe-by-probe alignment. The gridding software that comes with most scanners does this automatically but in many cases some manual adjustments need to be made. At the end of the process, spots that overlap with artifacts on the array, or are highly irregular, should be flagged to ensure that they do not affect the downstream analysis. At this stage, the Cy5 and Cy3 signal intensity at each probe is determined within each spot using the same software. While this data represents the gene expression changes measured on the array, normalization is required before detailed analysis can be performed. First, any systematic difference between the Cy5 and Cy3 signals, due to differences in the quantum yield and the amount of cDNA loaded on the array, must be corrected. For printed arrays this can be accomplished by multiplying the Cy5 and Cy3 signals by a constant so that their average, across all spots, is the same. Care must be taken, however, to ensure that this is an appropriate normalization.

For example, when a strain with the gene for a global repressor deleted is compared to the WT strain, the average Cy5 and Cy3 signals are expected to differ systematically. In this case a set of spike-in control RNAs, or an appropriate subset of genes, can be used to normalize the Cy5 and Cy3 signals. For commercial arrays, where the DNA in each spot is at high-density and precisely aligned, the Cy3-to-Cy5 ratio tends to be nonlinear as a function of signal intensity, due to dye quenching and other effects, and thus the more sophisticated locally weighted scatterplot smoothing (LOWESS) normalization procedure should be used. Finally, the data for the spots with weak intensity need to be thrown out or at least weighted appropriately. One simple way to do this is to eliminate all data where both the Cy3 and Cy5 signals are less than 1.5-fold above the background. For printed arrays the background is determined by the signal around each spot, while in commercial arrays it is determined by the signal at negative control spots printed at various positions on the array. The former method is critical for subtracting away a variable background signal, while the later method is better where the surface chemistry inside and outside the spot are different and background variation is negligible. As an alternative to throwing out data with low signal, pixel-to-pixel variation in the background and spot intensity can be used to calculate an error range for each spot and these error values propagated through the analysis. Such data filtering and normalization can be carried out in wide number of databases (e.g., the Stanford Microarray Database or Rosetta Resolver). Such databases are also extremely useful for storing microarray results and images and building tables of data from multiple arrays. These tables can then be fed into one or more of the wide range of microarray analysis packages available, or programs such as R and MATLAB (The MathWorks), for detailed analysis as described further in Chapter 2 of this volume.

REFERENCES

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21,** 1337–1342.

Beer, M. A., and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell* **117,** 185–198.

Capaldi, A. P., Kaplan, T., Liu, Y., Habib, N., Regev, A., Friedman, N., and O'Shea, E. K. (2008). Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat. Genet.* **40,** 1300–1306.

Carroll, A. S., Bishop, A. C., DeRisi, J. L., Shokat, K. M., and O'Shea, E. K. (2001). Chemical inhibition of the Pho85 cyclin-dependent kinase reveals a role in the environmental stress response. *Proc. Natl. Acad. Sci. USA* **98,** 12578–12583.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680–686.

Enjalbert, B., Smith, D. A., Cornell, M. J., Alam, I., Nicholls, S., Brown, A. J., and Quinn, J. (2006). Role of the Hog1 stress-activated protein kinase in the global transcriptional response to stress in the fungal pathogen *Candida albicans*. *Mol. Biol. Cell* **17,** 1018–1032.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., *et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934.

Kerr, M. K., and Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2,** 183–201.

Leber, J. H., Bernales, S., and Walter, P. (2004). IRE1-independent gain control of the unfolded protein response. *PLoS Biol.* **2,** E235.

Marion, R. M., Regev, A., Segal, E., Barash, Y., Koller, D., Friedman, N., and O'Shea, E. K. (2004). Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc. Natl. Acad. Sci. USA* **101,** 14315–14322.

Papa, F. R., Zhang, C., Shokat, K., and Walter, P. (2003). Bypassing a kinase activity with an ATP-competitive drug. *Science* **302,** 1533–1537.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., *et al.* (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287,** 873–880.

Roth, F. P., Hughes, J. D., Estep, P. W., and Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16,** 939–945.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34,** 166–176.

Smith, D. A., Toone, W. M., Chen, D., Bahler, J., Jones, N., Morgan, B. A., and Quinn, J. (2002). The Srk1 protein kinase is a target for the Sty1 stress-activated MAPK in fission yeast. *J. Biol. Chem.* **277,** 33411–33421.

Springer, M., Wykoff, D. D., Miller, N., and O'Shea, E. K. (2003). Partially phosphorylated Pho4 activates transcription of a subset of phosphate-responsive genes. *PLoS Biol.* **1,** E28.

Tsong, A. E., Tuch, B. B., Li, H., and Johnson, A. D. (2006). Evolution of alternative transcriptional circuits with identical logic. *Nature* **443,** 415–420.

Tuch, B. B., Li, H., and Johnson, A. D. (2008). Evolution of eukaryotic transcription circuits. *Science* **319,** 1797–1799.

Wang, Y., Pierce, M., Schneper, L., Guldal, C. G., Zhang, X., Tavazoie, S., and Broach, J. R. (2004). Ras and Gpa2 mediate one branch of a redundant glucose signaling pathway in yeast. *PLoS Biol.* **2,** E128.

Wang, W., Cherry, J. M., Nochomovitz, Y., Jolly, E., Botstein, D., and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: A case study of sporulation. *Proc. Natl. Acad. Sci. USA* **102,** 1998–2003.

Workman, C. T., Mak, H. C., McCuine, S., Tagne, J. B., Agarwal, M., Ozier, O., Begley, T. J., Samson, L. D., and Ideker, T. (2006). A systems approach to mapping DNA damage response pathways. *Science* **312,** 1054–1059.

Zaman, S., Lippman, S. I., Schneper, L., Slonim, N., and Broach, J. R. (2009). Glucose regulates transcription in yeast through a network of signaling pathways. *Mol. Syst. Biol.* **5,** 245.