quantitate the level of expression. Remember that expression is measured comparatively, so the key value is the red/green (R/G) ratio. If a red fluorescent tag is used for condition 2 and green for condition 1, an R/G ratio of 2.0 for a particular gene means that there is twice as much mRNA for this gene under condition 2 than under condition 1. A value of 0.5 means there is twice as much mRNA for this gene under condition 1 than under condition 2.

Analyzing microarray data is more complicated than it sounds. "Noise" in the data resulting from nonuniform spots, problems with repeatability, and other technical issues make it difficult to obtain precise, quantitative data from a microarray experiment. Information derived from a microarray must therefore be used with caution, and careful statistical analysis of repeated experiments with good controls is essential. Microarrays, however, provide an excellent means of identifying potentially interesting genes that are induced or repressed under a specific condition and thus determining starting points for future experimentation.

In this chapter, you will start by using a freely available software application called MAGIC Tool to analyze microarray data stored in public databases in the guided *Web Exploration*. In the *On-Your-Own Project*, you will develop your own computational solutions to the problems of processing and analyzing microarray data.

▶ 9.2.1 Objectives

- Understand the importance of measuring gene expression and the advantage of global analysis using microarrays.
- Understand how microarrays work and how they measure gene expression.
- Work with publicly available data and analysis tools to investigate changes in gene expression under experimental conditions.
- Understand what data manipulations are necessary to obtain usable information from raw microarray data.
- Learn how the large amount of data generated by a microarray can be mined to answer different questions.
- Develop programs to process microarray data, analyze gene expression, and examine the functions of genes with a particular expression pattern.

9.3 Guided Project: Analyzing Microarray Data

9.3.1 Web Exploration: Using MAGIC Tool to Analyze Two-Channel Microarray Data

Cells encounter many kinds of stresses in their daily lives. Although the systems of the human body maintain constant temperature, pH, salt concentration, and so on, those cells exposed to the outside are subjected to temperature extremes, and diges-

tive-system cells are exposed to countless molecules—some potentially toxic—found in the foods we eat. We may also stress our cells by becoming dehydrated, consuming alcohol, and so on. The problem is even worse for a single-celled organism such as yeast (*Saccharomyces cerevisiae*) that is directly exposed to its environment. Heavy metals are one example of an environmental stress that cells commonly encounter; because lead, zinc, cobalt, and other metals are toxic at high concentrations, we carefully monitor our drinking water, for example, to ensure that the levels of these metals do not exceed safe limits. Most cells have developed mechanisms to help them resist toxic heavy metals, and microarray analysis is one way to better understand those mechanisms.

For this project, we explore gene expression in yeast in response to zinc using a freely available software application called MAGIC Tool (MicroArray Genome Imaging and Clustering Tool) to analyze microarray data. Developed by Laurie Heyer and her team of undergraduate students at Davidson College (see References and Supplemental Readings), MAGIC Tool is an open-source program that works on all major platforms and is particularly useful in academic settings because of its ease of use. This software was developed as part of the Genome Consortium for Active Teaching (GCAT), a project that promotes participation of undergraduates in actual microarray experiments.

Microarray experiments produce data for the expression of thousands or tens of thousands of genes. A researcher conducting an experiment can "mine" his or her data to answer a particular question: *In which genes does expression increase the most when these cells are starved? Do muscle cells express a particular set of signaling genes? Which genes are turned off when these cells are infected by a virus?* Later, however, another researcher could potentially use the same microarray data to ask a different question. Thus, public databases (similar in principle to the GenBank database you have already worked with) have been established to store microarray data for later use. In this project, you will analyze data retrieved from the Stanford MicroArray Database.

The project consists of five steps: (1) installing and loading MAGIC Tool, (2) downloading and formatting the microarray data, (3) converting array data to numbers representing gene expression and building an expression file, (4) transforming the data to a more useful format, and (5) exploring the expression file to get information on gene expression. If time is limited, an option is to skip Steps 2 and 3 and to use the prebuilt expression file provided on the companion website to analyze expression of the genes. See the *Note* at the beginning of Step 4 (p. 287).

Step 1: Installing and Loading MAGIC Tool

MAGIC Tool can be found on the GCAT website at http://www.bio.davidson.edu/projects/GCAT/gcat.html or by using your favorite search engine to find the GCAT homepage. Click on the link `MAGIC Tool analysis software plus practice tiff files`. Alternatively, you can access MAGIC Tool directly at http://www.bio.davidson.edu/projects/MAGIC/MAGIC.html. Download and unzip the MAGIC Tool

software into the directory of your choice (make sure you do not include dots or spaces in the name of your folder); no installation process is required. You will need to have at least 512 MB of RAM (1 GB is preferred). Note that while this project will guide you through the use of MAGIC Tool to do some basic analysis, an extensive user's manual, online Flash tutorials, and considerable information about the inner workings of MAGIC Tool can be found on the GCAT website.

Find and double-click the `MAGIC-launch` icon in the appropriate MAGIC Tool directory (Mac or DOS). This will run MAGIC Tool. Make sure you do not close the DOS window that pops up! You now have MAGIC Tool loaded.

Step 2: Downloading and Formatting Microarray Data

(Note: If time is a concern, you can skip Steps 2 and 3. However, you may wish to read through these steps to gain an understanding of the process.)

The Stanford MicroArray Database (SMD) is currently located at http://smd.stanford.edu, or you can find it using a search engine (search for `stanford microarray`). Once loaded, click on the `Search` link to search through the experiments in the database, then choose `Search by Datasets`. On the next page, pick `Experiments` as the Results Type (*not* Experiment Sets). In the box where you are asked to choose an organism, select `Saccharomyces cerevisiae`, then in the Data Identifier box, choose `Metals`. This will limit your search to microarray data for yeast involving the use of heavy metals. Click the `Display Data` button to see the results of your search.

We will use microarray data comparing the gene expression of yeast grown with two different concentrations of zinc: 3 mM (high) and 61 nM (low). The microarray contains all the genes from the yeast genome, so we'll be doing a global analysis of the effect of an increased concentration of zinc on gene expression. The experiment ID for this project is 819; you should see it in the first column of the results. This experiment was conducted by Audrey Gasch in the David Eide laboratory (see References and Supplemental Readings). In this experiment, the green-labeled cDNA represents mRNA from the cytoplasm of yeast cells grown at the low concentration of zinc (61 nM), whereas the red-labeled cDNA represents mRNA from cells grown at high concentration (3 mM). Notice that there are various kinds of files you can download (icons near the middle of the page) to see the results of this experiment and various kinds of analysis. In order to do your own analysis, follow these steps to download the raw data and use MAGIC Tool to reach your own conclusions.

1. Create a folder called `GuidedProject` within your MAGIC Tool folder to hold the downloaded files for this project. Within this folder, create another folder called `DataFiles`.
2. Click the `OriData` icon in the `Options` column on the SMD page to download the microarray data: intensity of red and green fluorescence for every pixel of every spot on the microarray. It may take a few minutes to retrieve the files. Download into your `GuidedProject` folder the two TIFF files labeled `819GENEPIX0`, repre-

senting scans of the microarray for red and for green fluorescence, and unzip them into your `DataFiles` folder.

3. Besides the actual fluorescence data, you need a list to tell MAGIC Tool which gene is represented by which spot on the microarray. Go back to the SMD search results page and click on the `Raw Data` icon in the `Options` column. Download file `819.xls.gz` into your `GuidedProject` folder and unzip this file into your `DataFiles` folder.

If you had done this microarray experiment yourself, you would have a gene list file listing only the identities of each spot. Here, some of the analysis has already been done, and the `819.xls` file includes information that MAGIC Tool does not need, such as the calculated expression ratios for each gene. So, you need to use this file to construct a gene list—a file from which MAGIC Tool can link gene information to the expression data for each microarray spot. Our new gene list file must contain the name of the ORF (gene) assigned to each spot. (Note that some spots are blanks, standards, duplicates, etc.)

4. Open the gene information file `819.xls` in Excel (or another spreadsheet program). Notice that there are a number of rows of text information describing the file. Delete these rows until the top row in the spreadsheet is the row of headings for the data.
5. Although we are only interested in the gene names for the spots, it is important to point out that the order of these names is critical when analyzing the microarray data. MAGIC Tool will assign these names to the spots on the microarray based on parameters you will be providing. You can see how these names correspond to the spots on the microarray by looking at the data in columns `Q`, `R`, and `S` (`X Grid Coordinate`, `Y Grid Coordinate`, and `Sector`, respectively). In this case, you can see that the data consists of 16 blocks or "grids" of data, each read as a 24 × 24 matrix of spots, and that the spots are set up row by row rather than column by column. Armed with this information, you can discard everything from this file except the actual list showing, in order, what is in each spot; this is the contents of column `O` (`Name`). Notice that there are a number of different data terms used in this column, such as `EMPTY`, `GENOMIC`, and ORF names (`YAL001C`, `YAL003W`, etc.). Delete all the columns from the spreadsheet except column `O`.
6. Delete the header row. Your spreadsheet should now consist of only one column of data. Save the file as a tab delimited text file (use `SaveAs` and change the file type before saving). Name this file `819genelist.txt`.

Step 3: Building an Expression File

You are now ready to use MAGIC Tool, which has two major functions: (1) converting raw microarray data to an expression file containing quantitative data representing gene expression levels, and (2) analyzing those quantitative data to learn

about gene expression. You need to create a new project in MAGIC Tool and then build the expression file to analyze.

1. Start by creating a new project: in MAGIC Tool, choose `Project | New Project` from the menu. Call this project `guided9` and save it in your `GuidedProject` folder. (In the future, you can go back and work on this project by choosing `Project | Load Project` and loading `guided9.gprj`.)
2. You now need to associate the data files (TIFF files and gene list) you downloaded with this project. `Choose Project | Add Directory` and select your `DataFiles` folder. This will merge the data files from this folder into appropriate MAGIC Tool folders within your project. A window will pop up asking about the `819.xls` file. Skip past this file as prompted, since MAGIC Tool does not use this file. (You can also load the three files you need manually by choosing `Project | Add File` and loading `819_ch1.tif`, `819_ch2.tif`, and `819genelist.txt` separately.)
3. It is now time to tell MAGIC Tool which TIFF file represents the red channel and which represents green. Choose `Build Expression File | Load Image Pair | Red (none)`. Select the `819_ch2.tiff` file (red channel). Repeat with `Expression File | Load Image Pair | Green (none)` and load the `819_ch1.tiff` file.
4. Load the gene list so MAGIC Tool knows which spot is which. Choose `Build Expression File | Load Gene List` and select the `819genelist.txt` file.

You are now ready to "grid" your data. The **gridding** step identifies the spots on the microarray for MAGIC Tool by drawing a grid and defining the area within which the florescence data for each spot should be calculated. In this step, you can adjust the grid to compensate for larger or smaller spots, blocks of spots that are off-center, and so on.

5. From the menu, choose `Build Expression File | Addressing/Gridding | Create/Edit Grid`.
6. When you are asked whether you understand the layout of the grid, click `OK`. When you are prompted for the number of grids, enter `16`. You may leave all the other options at their default settings.

A new window will appear that contains the microarray grid. On the left, a number of text fields appear that you fill in to identify the grid layout. You need to determine the coordinates for the upper-leftmost spot, upper-rightmost spot, and a bottom-row spot for each grid. Although this step is a bit tedious, with practice it becomes much faster.

7. First zoom in so you can see the spots on your grid more clearly. On the upper-left window is a `Zoom In` button. Click on this button and then click anywhere on your grid a few times to zoom in.
8. Unfortunately, the spots are still too dark to see. At the top of your window is a slide bar that changes the contrast. Position this somewhere between 500 and

1000. You should now be able to better distinguish the spots on the microarray. Red, green, yellow, and black spots are visible, and it should be clear to you that this image represents a computerized merge of the red and green TIFF files.

9. To grid the data, MAGIC Tool needs the upper-left spot, upper-right spot, and a bottom-row spot's coordinates. An easy way to enter this information is to click on the appropriate button on the left of the window and then move the cursor over the *center* of the corresponding spot on the microarray image and click. MAGIC Tool will then insert the coordinates for you. Repeat these steps for the top-right spot and a spot on the bottom row. (Although it is difficult to see the top-right spot, try to position your cursor as close to the center of where this spot would be as you can.)
10. Check the coordinates in the fields on the left; they should be close to the following (you can also enter these numbers directly if you have trouble): `left (137, 104)`, `right (545, 104)`, `bottom (341, 490)`. (Note that the *x*-coordinate for the bottom row can vary depending on which spot in that row you choose.)
11. Enter the total number of rows (23) and columns (24) in the appropriate text field and click `Update`. A grid will appear around your spots.
12. Adjust the grid as needed so that each spot will fit cleanly within a square of the grid (see Figure 9.5). Use the `rotate grid` buttons at the bottom of the left window, reposition the grid by dragging it, change the space within each

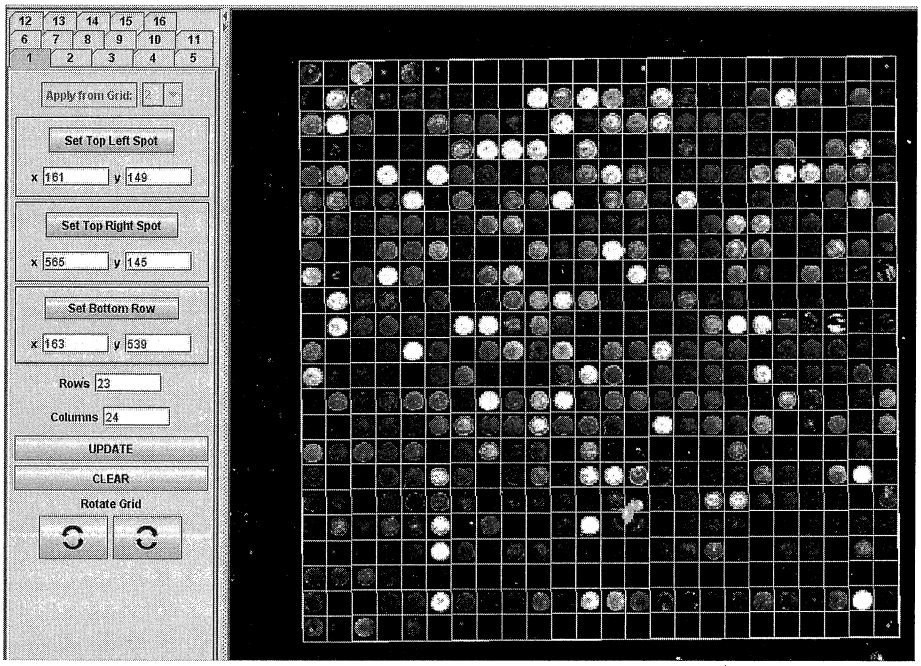

Figure 9.5 • Gridding the microarray data in MAGIC Tool.

square by clicking and dragging the tiny gray circles that appear in the four corners of the grid, or by changing the coordinates in the boxes and clicking `Update`.

13. Once you are happy with the layout of the grid, repeat this process for the remaining 15 grids using an easy shortcut. You can copy the grid layout you just created. Make sure your grid is currently selected (yellow grid), hold the `CTRL` key down, and click on the top-left spot on the second grid. This will copy the currently selected grid (yellow) to the new grid. In some cases, you may find it difficult to determine where your copied grid should be positioned: For some grids, it is difficult to see the first spot even with the contrast way up. If you mouse over any spot on your grid, you will see the gene name appear for that spot in the lower-left portion of your window. All yeast genes begin with the letter Y. A simple way to ensure you are positioning the grid correctly is to mouse over various spots on your grid. A valid yeast gene name should display for any brightly colored spots, while `EMPTY` should appear as the gene name for the completely black squares, particularly blanks in the top and bottom rows. Adjust the grid as needed and then move on to the next grid.
14. As you continue to grid, make sure you move from left to right and top to bottom in your image (grid 1 is the upper-leftmost grid, grid 2 is immediately to the right of grid 1, and so on until grid 16 is in the lower-right of the image) and make sure the correct corresponding tab number is highlighted in the left window for each grid. You may have difficulty aligning some of the grids, but align them as well as you can—remember, you are working with real data, and real data are often messy.
15. Once you have finished all 16 grids, click `Done`. You will be prompted with a series of questions, eventually typing in the name for your grid file. Enter `guided9grid` as the file name. You have just gridded your data.

You probably noticed that some of the grids contained **noise**, or obviously unusable data. For example, some spots expanded into other spots, invalidating the data both in the offending spot and the surrounding ones it touched. Some microarrays may even contain streaks blacking out data from several spots or bright smudges that do not look like spots. Before you move on you may want to use **flagging**—an option in MAGIC Tool to review the gridded data and "flag" spots containing noisy data that should not be used in the analysis.

16. Choose the `Spot Flagging` option under `Addressing | Gridding`. You will see an image of your grid.
17. Increase the percent contrast and use the `zoom` buttons to identify noisy areas.
18. Click off the `zoom` buttons and your cursor will become a flagger. Click on any spot you wish to exclude from your data analysis. An X will appear over the

spot indicating that its data will be excluded from the expression file. A good example spot to flag for this microarray appears in grid 2, col 19, row 22. If you look at this spot, you will see that it bled into the surrounding area, distorting the data in those spots. Go ahead and flag this spot along with its neighboring spots (left, right, and bottom spots). Flagging can also be used to exclude specific genes you do not wish to analyze. (Notice that as you mouse over a spot, the name of the gene appears at the bottom of the window.)

Once you are happy with the gridded and flagged data, go ahead and click on the `Done` button and follow the directions. You can now move on to **segmentation**—a process in MAGIC Tool that will find the actual red and green signals from the microarray data and calculate red/green ratios for each, a numeric representation of the relative expression of each gene at 3 mM zinc (red) as compared with 61 nM (green).

19. Under `Build Expression File`, choose `Segmentation`.
20. You will see a grayscale representation of the red image and green image for each spot. In the left panel, under `Spot Number`, use the `Next` button to move through the spots in the first grid until you get one that is fairly bright. Spot number 80 is a good choice.
21. You can choose various segmentation methods to calculate the intensity of red and green in each spot. The default option is `Fixed Circle`, which uses the same size circle to calculate ratios for each spot. However, we will use the `Adaptive Circle` option. This option will change the size and location of the area used to determine the signal ratio based on where the program thinks the spot is. It should be more accurate than the fixed circle method (though slower) where the spot sizes are not perfectly uniform. If you move through the spots, you can see the location of the circle changing when using this option.
22. The ratio method describes how the program calculates the red/green ratio. Use the default method `Total Signal`, which uses every pixel of the spot when determining ratio.
23. Click on the `Create Expression File` button at the bottom of the left panel. Enter `ch9guidedexp` for the file name and `zn3mM` for the column name. You now have an expression file to analyze.

Step 4: Transforming the Expression Data

(Note: If time is a concern or for consistency of results, you may decide to use the expression file already created for you. This file can be found on the companion website and is named `ch9guidedexp.exp`. *You can download this file and save it in your* `Guided Project` *folder. If you have not already created a project in MAGIC Tool, create a new project by choosing* `Project | New Project` *from the menu. Call this*

project guided9 and save it in your GuidedProject folder. (In the future, you can go back and work on this project by choosing Project | Load Project and loading guided9.gprj.) You now need to associate the expression file you downloaded with this project. Choose Project | Add File *and choose* ch9guidedexp.exp. *You can now proceed to the remaining steps to transform your expression file.)*

1. Choose Expression | Working Expression File and choose ch9guidedexp.exp.
2. Choose Expression | View Data and the expression data file will open. You should have two columns of data. The names of the genes are in the first column. Actual yeast genes look like YPR150W, while spots with names like 1_rep11 are controls.

You now need to log transform your data. Suppose gene A is expressed four times as much when exposed to high zinc as when the cells are grown under normal conditions. Gene B, on the other hand, is expressed only one-fourth as much under high zinc conditions. The red/green ratio (which is what you see in the data file) would be 4.0 for gene A and 0.25 for gene B, even though both showed a four-fold change in expression (in different directions). If you take the log (base 2) of these numbers, however, then you get 2.0 for gene A and −2.0 for gene B, which helps you see the real picture of their changes in expression more clearly.

3. In order to log transform your data, click on Expression | Manipulate Data | Transform. A small window should pop up.
4. Make sure the log option is chosen and that 2 appears in the field for the log value. Click OK.
5. You cannot take the log of zero, so if some of the spots are completely black, the program will you ask what you want to do about it. It's OK to accept the default and set these to a very small number.
6. Save the new file that contains the log transformation data as ch9guidedexp_tlog2.exp.

Finally, you may have noticed that there are replicates in the data, indicated, for example, by YDR238C_rep1. MAGIC Tool can automatically average those replicates, giving us better data. If you are using the downloaded expression file, you can again skip this step.

7. Choose Expression | Average Replicates.
8. From the drop-down in the dialog box, choose the log-transformed expression file you just created.
9. Save the new file containing the averaged data as ch9guidedexp_tlog2_avg.exp.
10. If you view the data, you will see that there is now one list of all the genes with the log of the red/green ratio shown. Most of the control data are now out of

the way, too. Some genes will be labeled "missing" if a spot were flagged as unreadable.

Step 5: Exploring Gene Expression

At last, you are ready to begin exploring your data. This means you can actually find out the effect of zinc on gene expression in yeast! Suppose we want to identify genes whose expression increases significantly in the presence of 3 mM zinc, as well as those whose expression decreases significantly.

1. Choose `Expression | Explore` from the MAGIC Tool main menu.
2. Click `Find Genes Matching Criteria` in order to select genes of interest from the expression file.
3. In the dialog that opens, you can choose any of a variety of criteria to define what genes you would like to look at. Start with those that are induced in the presence of zinc, and only look at those that show at least a fourfold induction (log-transformed red/green ratio > 2.0). Set the option `Value in column labeled...` to `> 2` and click `OK`.
4. You should see that approximately 15 genes met these criteria (the exact results may vary slightly based on how you gridded and segmented the data). To get a visual picture of the results, click `Plot Selected Group`. You will see a graph with a dot to represent each gene whose expression was at least four times higher in the presence of 3 mM zinc.
5. Click on a point on the graph, and you will see the name of the gene it represents in the lower-right corner. (If you get a control spot, just click another point.)

At this point, it would certainly be helpful to know more about the genes you have identified. Some undoubtedly have known functions, and in some cases, induction in response to zinc might make sense with what is known about the gene, whereas in other cases this might be an unexpected finding that could point you toward a new discovery. Furthermore, the functions of about a quarter of yeast genes remain unknown; finding that some of those respond to zinc could be a first step to understanding what they do. MAGIC Tool can display basic information about each gene (such as what chromosome it is on and any known functions) if we give it a file to work from.

6. Download the file `yeastgenes.info` from the companion website. This file also appears on the GCAT website. Add this file to your project.
7. From the MAGIC Tool `Expression` menu, choose `Import Gene Info`.
8. Choose the averaged expression file, then find `yeastgenes.info` and click `OK`.
9. Repeat the process of selecting genes whose expression increases by at least fourfold and plotting them. Choose an interesting gene from the plot. Then click the small arrow near the top of the plot frame to reveal details about that gene.

10. For still more information about the gene, use the Saccharomyces Genome Database (http://www.yeastgenome.org). Type the name of the gene in the search box and click Search. You should see a wealth of information about the gene you have selected, especially if it happens to be a well-characterized one.

Web Exploration Questions

1. How many different yeast genes are represented in this microarray? What is the total number of genes in the yeast genome?
2. List two yeast genes whose expression in 3 mM zinc is at least eight times as high as in 61 nM zinc and which have characterized functions.
3. For each gene you listed, briefly describe what is known about its function. Can you suggest why the cell might "want" to induce these genes in response to zinc?
4. List one yeast gene whose expression in 3 mM zinc is at least eight times as high as in 61 nM zinc and whose function is unknown. *(Hint: You may want to select all the genes in the plot and take a look at what is in the gene information file before deciding what criteria to use.)*
5. How many genes of unknown function whose expression increases by more than eightfold are there altogether in this microarray?
6. List two yeast genes whose expression *decreases* by at least fourfold in the presence of 3 mM zinc. What is known about these genes?

Putting Your Skills into Practice

Now that you are somewhat familiar with using MAGIC Tool, the following exercises ask you to use data from two microarrays to learn more. MAGIC Tool can put data from multiple microarrays into a single table or graph, allowing researchers to look at expression over various time points or to compare multiple conditions. Each time point or condition tested requires a microarray with both red and green fluorescence data, just as you have seen so far.

The yeast protein ZAP1 is a zinc-regulated transcription factor. In the presence of high concentrations of zinc, it binds to the promoter regions of genes and induces their transcription. Thus, many of the genes whose expression increased when zinc was added in the first experiment we looked at could have been induced because ZAP1 bound to sites in their promoter regions and stimulated transcription. We can use microarray analysis to investigate this further and find the answers to two interesting questions: (1) What genes does ZAP1 regulate in response to zinc? and (2) Are there genes whose expression is induced by zinc that are *not* regulated by ZAP1?

In order to look at these questions, we need to look at gene expression under low (61 nM) and high (3 mM) zinc conditions again. This time, however, the experiment is done in a mutant yeast strain that has a nonfunctional *zap1* gene. In