2

Linking Image Quantitation and Data Analysis

Gregory C. Bloom, Peter Gieser, and Emmanuel N. Lazaridis

1. A Shifting Paradigm

Until recently, image-based experimentation in molecular biology has been primarily concerned with qualitative results produced as a result of such experiments as Northern blots, immunoblotting, and gel electrophoresis. These experiments result in a relatively small number of bands on an autorad or other imaging medium. These bands or spots would be visually inspected to determine their "presence" or "absence," or visually compared with other spots on the medium to determine their relative intensities. Sometimes, comparisons would be enhanced using quantities derived from densitometry analysis. Such comparisons were often performed to provide a numerical summary of a clearly visible difference. This summary may have been required for publication of the experimental results. This approach seemed to serve the investigator well because there existed no real need for accurate image quantitation or data analysis and a simple qualitative result would suffice.

However, many recent advances in molecular biology, coupled with the increasing knowledge of the human genome, have made possible the ability to simultaneously test the expression level of several thousand individual genes, as in the case of microarray analysis (*see* Chapter 3 by Gieser, et al.), or hundreds of expressed proteins, as with two-dimensional (2-D) gel electrophoresis (*see* Chapter 4 by Seillier-Moiseiwitsch, et al.). While this ability is essential to further molecular biology research and is a giant leap forward from more traditional approaches, it has raised several questions about the use of the "old" paradigm of image quantitation and data analysis and whether that paradigm can be successfully applied to these new image types. Several characteristics of modern molecular biology experiments—including the need to investigate and understand subtle changes in molecular quantities and the

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

increasing sensitivity of quantitation to the imaging process—suggest that the old paradigm must be modified. In this chapter, we suggest a new approach that allows investigators to better handle the needs of image-based experimentation.

To demonstrate why a new paradigm linking image quantitation and data analysis is needed, and to better understand the scope of the problems faced when analyzing a laboratory image, we briefly describe some of the new technologies and the image types they produce.

Microarray analysis (*see* Chapter 3 Gieser, et al.) is a procedure that allows an investigator to simultaneously visualize the expression levels of thousands of genes whose complementary sequence or a portion thereof has been arrayed on a class slide or chip. The measurement of mRNA levels in, for instance, a normal tissue or cell line to its paired experimental sample can elucidate which genes and, indirectly, which proteins are present or absent, and their relative expression levels in one condition as compared to another. This gives the investigator a starting point to determine which genes or groups of genes are important in a particular experimental context. Regardless of the type of question(s) being asked, this experiment invariably results in a large image or set of images with thousands of features, each of which needs to be geometrically defined into a region of interest (ROI) and subsequently quantitated. The microscopic scale on which this kind of experiment is performed plays an important role in determining the sensitivity of analytic results to the imaging process. Ratios of quantities across images are frequently needed to compare the relative expression across conditions.

The second type of modern biological image-based experimentation is termed *proteomics*. This is the science that deals with gene products, namely, proteins, and concerns itself with the collection of proteins (the "proteome") produced by a particular cell or organism. Important information can be derived from experiments seeking to establish whether specific proteins are made in higher or lower concentrations in response to disease, drug treatment, or exposure to toxicants. The most commonly used approach to protein identification and quantitation is 2-D gel electrophoresis, which combines a first dimension separation by isoelectric focusing (IEF) with a second dimension separation by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Whereas microarray experiments result in images with features whose geometry is determined by the physical assembly of samples on a substrate, proteomics 2-D gel images consist of many spots whose location and shape cannot be prespecified easily. As with the microarray, the 2-D gel image consists of several hundred to several thousand features of varying intensity that need to be characterized. Each feature may or may not be important in the context of a given experiment.

In both microarray and proteomics 2-D gel contexts, effects due to image background, signal-to-noise ratio, feature imaging response and saturation, and experimental design and execution must be accounted for and factored into the overall image quantitation procedure. Any and all of these factors can have far-reaching effects on the subsequent data analysis. Under the old paradigm, determination of the effects of variation in these factors on subsequent data analysis is impossible, if for no other reason than that image quantitation would typically proceed under a single set of conditions, in a step that would never be revisited. If subtle differences in the performance of image quantitation may substantially affect the subsequent data analysis, then the old paradigm simply no longer serves, as it allows only a single best "guess" at what imaging parameters are optimal and allows for no testing to see if the guess was correct.

A key point of the discussion of our new paradigm for treating images from biological experimentation is that image quantitation can have a potentially large effect on the data that are being obtained and these effects would feed through the subsequent data analysis. In any imaging experiment there exists an infinitely large number of ways in which an image can be quantitated, all of which may be "correct" in that they all lie in some reasonable envelope of imaging procedures. Among these methods, and even across subtle variations of a single method, substantial variability in quantitation may result. This is particularly important when one considers searching for subtle trends or effects in a data set. For the newer types of image-based biological experimentation, such subtle differences in how image quantitation is performed can completely alter the data analysis outcome. A method is needed for linking the imaging and data analysis processes so that the one can feed into the other, enabling the investigator to understand the effect of choices made in image quantitation on the resulting data analysis. The reverse situation is also important, as the results of data analysis can drive the choice of procedures for image quantitation. For example, an analysis of data derived from a particular procedure for image quantitation using a specific background cutoff value in a given image may demonstrate that the imaging procedure eliminated too many features of the image from consideration, necessitating that the image quantitation process be revisited. The idea of using the results of one of the two steps in this process to drive the other process is central to the new paradigm.

Such an approach is important for the analysis of the newer types of biological images produced today because of their sheer complexity and the large number of features contained within each image. In hindsight it seems that the traditional segregation of image analysis from the data analytic process may have been sub-optimal for analysis of more traditional types of biological experiments as well.

In this chapter, we introduce a new paradigm with an accompanying schema for the treatment of experimentation involving images and their subsequent

data analysis, and point out the benefits of this new approach. The new approach encourages cooperation between image quantitation and data analysis. Ideally, this implies that the two processes should be performed by a single software application. While not necessary, integration of imaging and statistical software tools can make application of the new paradigm easier, as we will describe and illustrate in detail later in the chapter.

2. Conceptualization of the New Paradigm

The first action with any imaging experiment is to produce the medium with the features or items to be imaged. The medium can be a microarray chip or slide, a proteomics 2-D gel, or any number of other experimental media. The second step is image acquisition. This can be as simple as scanning a piece of exposed film or as complex as scanning a 2-D SDS-PAGE gel in a proteomics experiment. To illustrate the use of our novel paradigm for the image quantitation and data analysis step of an imaging experiment, a workflow diagram is shown in **Fig. 1**.

The first step in this process is image quantitation. Image quantitation consists of translating the underlying pixel information in the image into useful data through the use of imaging methods. The set of imaging methods and their associated parameters constitute an *imaging envelope*. Methods in the imaging envelope may differ in how they treat background signal information, identify signal in the presence of noise, characterize feature geometry, and identify features with labels. The parameters that are required by a particular method to perform quantitation can include numerical summaries of background signal, expected signal-to-noise ratios, or signal thresholds for the image. The methods and parameter values defining the imaging envelope are what determine the values of the resulting *data sets* (shown below the imaging envelope in **Fig. 1**). It is important to note here that even subtle changes in the imaging envelope can lead to large changes in the acquired data set(s). These changes will, in turn, alter the *inferences* obtained by application of the *statistical algorithm*. It is therefore very important to incorporate a reality check after data analysis and a subsequent feedback mechanism for improving specification of the imaging envelope. Modifying the envelope in turn will necessarily alter the inferences. Note that in some situations, particularly when formal analytic protocols must be consistent over multiple analyses, feedback may be undesirable beyond an exploratory stage.

After the data sets are obtained, a single statistical algorithm is applied to each individually. The type of statistical algorithm used is not critical to the paradigm and may be anything from a *t*-test to linear regression. In one possible path of workflow, the inferences are grouped into a *inference set*, representing the individual values obtained from application of the statistical algorithm. At

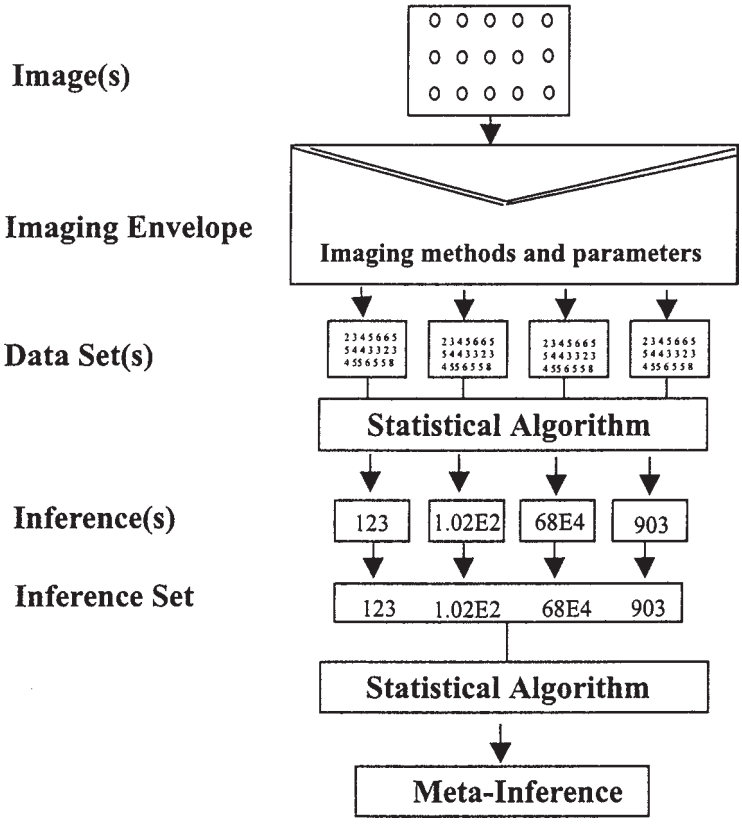

Fig. 1. Work flow for an imaging experiment using the new paradigm.

this point a meta-analysis of the inference set, using analysis of variance (ANOVA), for example, is performed to arrive at a summary description or *meta-inference*. This summary result incorporates not only the final outcome of the data analysis, but also a measure of the variability or potential error introduced by the imaging process.

The other possible path through the work flow diagram summarizes the data sets obtained as a result of image quantitation into a single meta-data set before application of a statistical algorithm. This treatment leads to a single inference at this point in the flow and no further analysis is necessary. This approach has an advantage in that it is more amenable to specification of distributions for parameters characterizing the imaging envelope. For example, when one is interested in integrating out the effect of a particular parameter from a specific imaging algorithm, one can place a prior distribution on that parameter and calculate an inferential posterior distribution using a Bayesian approach. In

addition, the loss of information resulting from the application of the statistical algorithm occurs at only one point in the flow, making it easier to evaluate goodness-of-fit. Disadvantages of this approach include the fact that the image quantitation procedures must be "compatible" across the imaging envelope so they can be combined in the context of conducting a single statistical operation. In the alternative approach, only the intermediate inferences, and not their underlying data sets, need be combined for the subsequent statistical analysis leading to the meta-inference. Therefore, different statistical algorithms may be applied to each individual data set as long as the inferences can be meta-analyzed, making this approach more flexible.

When the process illustrated in **Fig. 1** is integrated in a single software platform, models of the experiment that account for use of different imaging parameters and quantitation procedures can be more readily explored, reducing the potential for imaging-related biases in the analytic results. The sensitivity of any given analysis to changes in quantitation procedure can also be rapidly investigated, thereby increasing the quality of information derived even from simple statistical models. The next section describes a novel application that allows this conceptual solution to be practiced in a real-world environment.

3. Application of the Paradigm: The Midas Key Project

While it is easy to conceptually cycle through several rounds of quantitation and data analysis using the approach described in **Subheading 2.**, it is much more difficult to perform this task in a real-world environment. This is especially true if the processes of image quantitation and data analysis are physically separated. In fact, this is the situation that currently exists. Many systems are available for image analysis, including home-grown and commercial, general and special-purpose packages such as Optimas (Media Cybernetics, Inc.; general purpose imaging), SpotFinder (TIGR; microarray slide imaging), and CAROL (Free University of Berlin; proteomics 2-D gel imaging). Indeed, many vendors of biological equipment produce and distribute their own software, which they bundle with their equipment. While some of the available packages may provide sophisticated image-analysis tools, little sophistication is available in the included mathematical and statistical methods for analysis of the resulting data. Conversely, popular analysis packages such as SAS, SPSS, and S-Plus, while providing sophisticated models for data analysis, lack any facility for image quantitation. Thus, the typical scientific segregation of the analytic role from the process by which image-related data are obtained is also reflected in available software. While such software may suffice to conduct the kinds of traditional biological experimentation that relied primarily on qualitative examination of images, it was recognized that use of such software in the context of the new biological experimentation would be suboptimal.

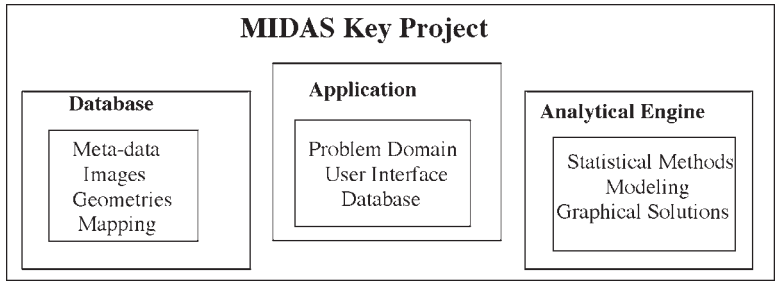

Fig. 2. Diagram of the three components constituting the MIDAS Key Project and their elements.

The paradigm of marrying imaging and mathematical modeling and statistical tools to analyze the results of modern biological experimentation could be implemented using the disparate software applications described previously. This approach has several limitations, however, foremost of which is the ability to quickly incorporate the results derived from either of the two analytic domains into the other. One would need to go back and forth between the imaging and data analysis exercises hundreds or thousands of times. A platform that would allow imaging and data analyses to proceed in tandem would substantially enhance the analytic exercise. Thus, we have been developing an application that incorporates all aspects of image and data analysis along with data storage into a single unit. We describe the design and merits of this application in the paragraphs that follow.

The goal of the MIDAS Key Project is to build an integrated imaging and modeling analytic environment over a sophisticated database backbone. By borrowing and uniting technologies from multiple fields, we seek to empower researchers in basic and clinical imaging studies with a sophisticated analytic toolbox.

Figure 2 illustrates the three major components that constitute the MIDAS Key Project. A description of each of the elements contained in each of the components is also given. The top-level box is the Java *application*. This is the central component of the key project and ties the other components together. The Problem Domain of the application contains the objects that define the underlying data structures used for the project. The Java application also controls interactions with the database; this is done in the Database area. The third area is the User Interface. This package is responsible for all aspects of interaction with the application, including the menu-driven frame-based interface and image display. The use of Java allows us to maintain cross-platform independence; to integrate tools existing in multiple, otherwise unrelated, applications; and to easily deploy a client-server multithreaded model system.

We are currently using Java 2 as the basis for our code, supplemented by the Java Advanced Imaging (JAI) Application Programming Interface (API). The JAI API is the extensible, network-aware programming interface for creating advanced image-processing applications and applets in the Java programming language. It offers a rich set of image processing features such as tiling, deferred execution, and multiprocessor scalability. Fully compatible with the Java 2D API, developers can easily extend the image-processing capabilities and performance of standard Java 2D applications.

The current Java Development Kit (JDK) fully incorporates Swing components (which are used for windowing functions) and the 2D API, both of which are employed throughout our code. The Java Database Connectivity (JDBC) API allows developers to take advantage of the Java platform's capabilities for industrial-strength, cross-platform applications that require access to enterprise data.

The *database* component of the MIDAS Key Project contains the table spaces that hold all long-term storage needed in the application. The tables contained here include those for storing project, experiment, and image meta-data; tables to store the images and their associated geometries; and mapping tables to tie the data together. For our work we chose to employ Oracle for all data storage and management. Oracle provides many unique technical features that we leverage in the Key Project including Java integration, extensibility and scalability, and support for multimedia data types that allow for efficient integration of imaging and meta-data information.

The most important characteristic of an *analytical engine* in the Key Project is its amenability to integration with other software, including novel statistical methods. A second characteristic is the ease with which it interacts with Java applications. We chose to employ the S-Plus statistical processing system for our work in spite of the fact that it is not fully Java aware. A fully Java-aware analytic engine would allow dynamic statistical methods to be incorporated into our Java interfaces, allowing application of real-time graphical data exploration methods and interactive statistical diagnostics. In addition, we can conveniently employ the S-Plus system on desktop computers separately from our Java interfaces, assisting in rapid methods development and evaluation.

Figure 3 shows a typical application of the Key Project system, focusing on the Oracle backbone, which is used for object persistence. First, a series of image-dependent or imageless layers, upon which analysis will be performed, are loaded into the system (step 1). Memory is carefully managed at this step and throughout the process, as it is impossible to expect either client or server to simultaneously manage, say, 40 microarray images, each of which is upwards of 40 Mb long. A rendered composite image, if available, is displayed on the client according to user-adjustable preferences. We allow for imageless layers so that we might work in our analytic environment with data obtained

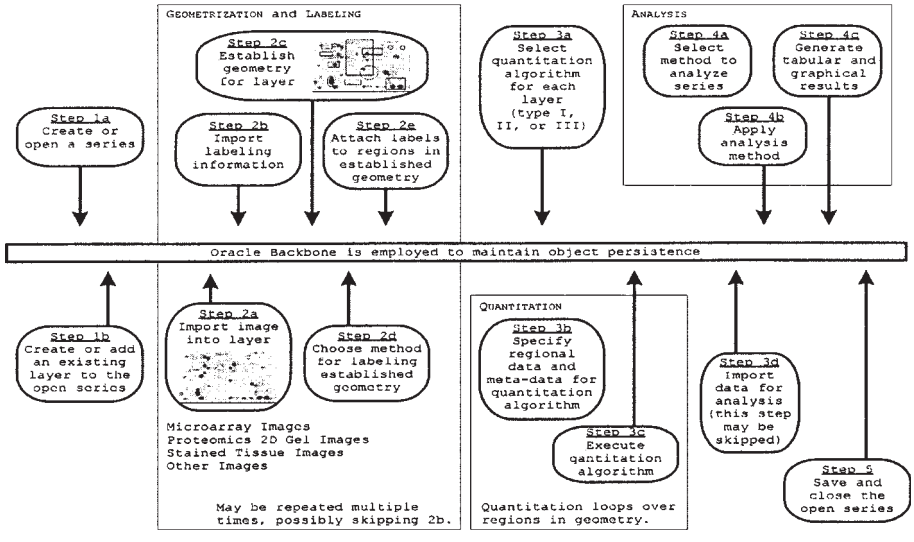

Fig. 3. Schematic of the MIDAS Key Project system showing Oracle backbone.

through sources whereby the associated images are not available. When images are available, we proceed to establish one or more geometries for each layer (step 2). By a *geometry* we mean a set of closed, possibly overlapping regions-of-interest (or shapes), each of which is not exclusively contained in any other. Geometry may be established by hand through a sketchpad interface or by application of a geometrization algorithm. The use of geometrization algorithms allows us to model in a single system images with formats that are largely fixed by the investigator, such as, for example, results from microarray studies, images with semifixed geometries such as from proteomics studies, and images with free-form geometries such as from cell or tissue microscopy. Labels are then attached by reference to one or more labeling algorithms (end of step 2). These may be relatively simple—typically, microarray labels are established by considering the spot centers—or fairly complex—protein labels on 2-D gels are established by considering the overall geometry and relative positions of shapes in that geometry. Geometries are calculated and labels established. Next, quantitation is carried out (step 3) by referencing one or more quantitation algorithms, which execute looping over shapes in the geometry. Quantitation may result in all kinds of information, including: (1) primary signal information, such as average or median intensity of the pixels in regions of interest; (2) signal variability information, such as pixel variance, kurtosis,

or direction of one or more principal components; (3) signal location information, such as coordinates of the intensity mode within a region of interest; and (4) cross-image signal comparison information, such as pixel correlation between two images (used for quality control). The design of our system allows for substantial extensibility in the application of geometrization, labeling, and quantitation algorithms. Depending on the algorithm, quantitation may be performed by server-side Java or C++ code or by the S-Plus Server system. Note that geometrization algorithms may also be employed within the quantitation step, without requiring persistent storage of the resulting geometry, as might be needed when one wishes to compare quantitative performance of two spotfinding algorithms within regions of interest in a specified geometry. External data, for which no images are available, are also retrieved at this time. Analysis of the quantitation results occurs in step 4. We employ standard methods such as simple regressions, ANOVA, and principal components analyses by referring to the methods built into the S-Plus analytic engine. Novel mathematical models are included by incorporating C++ or Fortran compiled code into the S-Plus engine or by direct reference to external code on the server. Graphical, tabular, or data-formatted results can be exported for reports or stored on the Oracle backbone for later use (step 5).

Initial exploration of multiple image-based experiments suggests that the variability associated with application of reasonable but differing imaging procedures to the same images is nontrivial. The total effect this variability will have on various statistical models is unknown at present. Without reference to our new paradigm for imaging and data analysis, it would remain largely unknowable.

4. Midas Center at USF—An Interdisciplinary Implementation of the New Paradigm

The new paradigm has changed the way researchers at our institution interact to analyze imaging-based experiments. The University of South Florida (USF) Center for Mathematical-Modeling of Image Data Across the Sciences (MIDAS) brings together faculty and student investigators from disparate fields to develop sophisticated mathematical and statistical models of data derived from images. Under the umbrella of MIDAS, we seek to address pressing analytic needs related to molecular biology experiments in many areas, including microarray, microscopy, proteomics, and flow cytometry. In each kind of experiment, an image or a set of images is typically derived by a primary investigator—say, a biologist or pathologist—in an experimental context. To get from the images to informative research conclusions, the steps of quantitation, analysis, and interpretation must be traversed. In today's research environment, the primary investigator usually directs quantitation of the images, sometimes in conjunction with an imaging scientist. The resulting data

may then be given to a statistician or other numerical analyst. The basic tenet of the MIDAS Center is that segregation of the analytic role in this context is suboptimal. At the present time, the MIDAS Center is integrating researchers from multiple schools and programs around USF. Investigators from programs in Biology, Bioengineering, Computer Science, Mathematics, Statistics, Medicine, Medical Imaging, Oncology, Biochemistry, Pathology, and Public Health are collaborating to address important analytic problems.

5. Example

Synchronous implementation of the new paradigm in both software and the collaborative environment allows for easy conduct of joint imaging and analysis experiments. In this section we first present a hypothetical experiment employing the new paradigm, and then illustrate application of the paradigm to an image using the MIDAS Key Project.

A hypothetical experiment using the new paradigm might be the following. Suppose we have conducted an experiment using 40 microarray slides that were assembled on two different days. We are concerned that our data analysis might be sensitive to problems we suspect with the microarrayer pins, and we have developed three combined sets of geometrization, labeling, and quantitation algorithms that we can apply to these data, each of which has some benefits and some drawbacks in terms of ability to adjust the resulting data for experimental difficulties. Each algorithm additionally has some imaging parameters that can be specified by the user, such as background pixel intensity cutoffs, complexity-cost, scale, or tolerance parameters. Suppose there are five such parameters in each algorithmic set, each having a low, medium, or high value in a reasonable range. Using the Key Project system, one could analyze the microarray slide images using each of the algorithm sets and a range of parameters to obtain, say, an analysis based on each of $3 \times 3 \times 5 = 45$ combinations of imaging methods. These analyses could then be averaged and deviant analytic results investigated using statistical meta-analysis techniques that would also be built into the system. In addition, we could consider employing Bayesian statistical methods to average-out the effect of imaging-related variability from the analysis, thereby obtaining a composite estimate that does not rely on a specific imaging protocol.

In the following example, we used the MIDAS Key Project to specify an imaging envelope around a spotted array image, having the usual red and green channels. For simplicity, rectangular areas were drawn on the image to identify 100 spots, and only two imaging choices were compared. For each of the rectangles, quantitation proceeded by setting a background threshold and computing the average pixel intensity. Two different threshold values were used, 5 and 25. These background levels are virtually indistinguishable

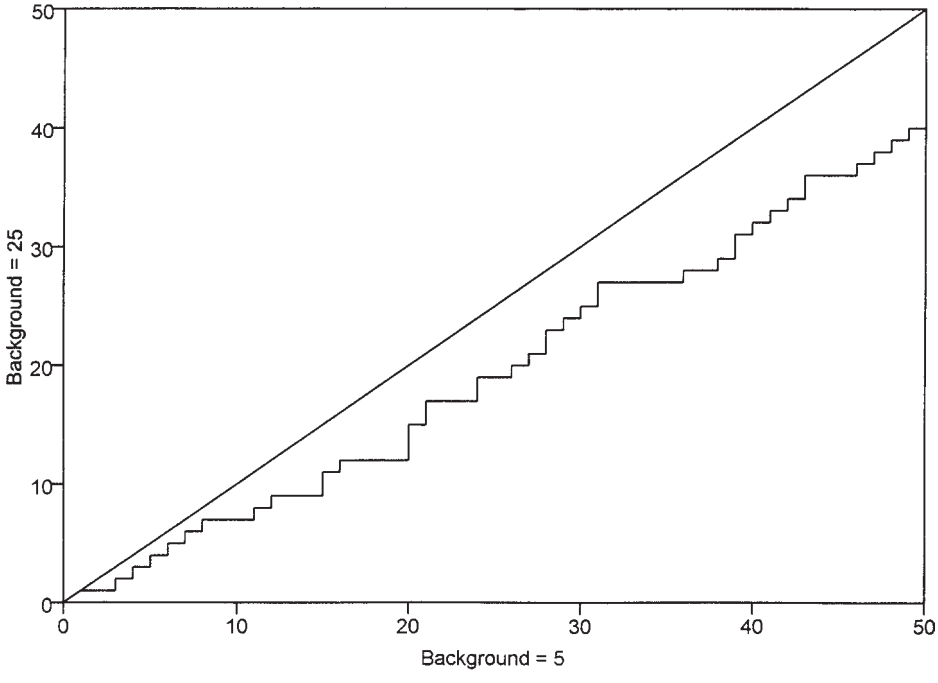

Fig. 4. Comparison of gene ranking of fold change between background levels of 5 and 25.

when visualized; visual comparison with the TIGR image suggested that either may be a reasonable choice. Thus the imaging envelope consisted of two members. The statistical analysis consisted of estimating fold change between the red and green channel and computing the corresponding rank of each gene.

Figure 4 presents a comparison of the relative ranks of the genes, across this simple imaging envelope. The height of the curve is the number of genes in the intersection of the top *x* ranked using a background value of 25 vs a background of 5. For example, at the value 10 on the horizontal axis, the height of the curve is 7, indicating that only 7 of the genes using a background of 25 overlap with the top 10 using a background of 5. Even in this simple example, inferences derived using two parameter values in a reasonable neighborhood demonstrate only 80% consistency. In more complicated situations, 30–60% or more additional and previously unrecognized variability may be captured in a reasonable imaging envelope. In this example, the inferences drawn across the imaging envelope could be meta-analyzed to form a consensus inference concerning the order of differentially expressed genes.

6. Conclusion

In every experimental context in which images are captured in the process of obtaining information, it is important to realize that the images are the data. Historically, inadequate attention has been paid to this viewpoint. As a researcher, one seeks conclusions that are resistant to the peculiarities of any particular imaging methods used in the process by which inference is obtained. The main benefit to an investigator is the ability to account for various factors within the imaging phase of the experiment. As detailed earlier, factors such as background signal, geometry characterization, and signal thresholding can and do have an effect on the resulting data, which in turn affects downstream analysis. Control and awareness of these influences allows an investigator to conduct inference that better reflects the underlying biology. We suggest that the MIDAS Key Project described in this chapter and the paradigm on which it is based provide such an approach, enhancing the completeness of data analysis and leading to better models for inference in image-based experiments.