Modeling and Analysis of Gene Expression Arrays

by

Keith Howard Duggar

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

May 10, 2004 [June 2004]

© Keith H Duggar, MMIV. All rights reserved.

The author hereby grants to MIT permission to
reproduce and distribute publicly paper and electronic copies
of this thesis document in whole or in part.

Author .... Keith H Duggar
B. S. Chemical Engineering
Georgia Institute of Technology

Certified by .... Douglas A Lauffenburger
Professor of Chemical Engineering
Thesis Supervisor

Certified by .... Peter K Sorger
Professor of Biology
Thesis Supervisor

Accepted by .... Daniel Blankschtein
Professor of Chemical Engineering
Graduate Officer

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY
JUN 16 2004
LIBRARIES

ARCHIVES

Modeling And Analysis of Gene Expression Arrays

by

Keith Howard Duggar

Submitted to the Department of Chemical Engineering on May 10, 2004,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Abstract

Gene expression arrays are a technology used to measure quantities of messenger ribonucleic acid (mRNA). Application of the technology involves a variety of physical processes beginning with the acquisition of mRNA samples and ending with the fluorescence imaging of a gene expression array. This thesis examines these physical processes, develops a mechanistic model, and derives the analysis procedure based on the model. Chief advantages of this approach are that it accounts for certain previously unexplained array phenomena and is based in a clear way on physical knowledge allowing non-arbitrary determination of both the probability that any given gene has altered expression ratio relative to a control as well as the magnitude of this induction or repression. We demonstrate its use on simulated and real array data, and show that a considerable amount of previously unrecognized information concerning gene expression differences is inherent in the array measurements.

Acknowledgements

If I were to thank all those who have contributed to building the man who wrote this thesis, you would need google in order to find mention of any particular person. So I ask that you forgive me for mentioning only the most direct contributors to my graduate research adventure though many fold more have contributed in some part along the way.

Firstly, and ironically, I would never have applied to graduate school had not a young woman, Nhung Nong, threatened to break up with me if I did not earn a Ph. D. So, I went to MIT and then we broke up. As you can see, I was left stranded here in the frozen North, in graduate school, but to do this day I owe her a debt for pushing me out of my warm tropical home in Florida.

I would never have been accepted, indeed my application would never have been reviewed, were it not for the generous recommendation and of Arnold Stancil, a mentor and finest teacher from Georgia Tech.

From my first days at MIT, Douglas Lauffenburger has been a friend, a guide, a champion, and the most patient advisor I could imagine. He gave me at once both freedom and direction, support and challenge. Truly, he is an amazing man. And I want to thank my advisor Peter Sorger who managed to revitalize my enthusiasm for science with his. Not to mention he is a heck of a funny guy.

I would like to thank my parents and family who have been supportive through my entire education. They have always encouraged me to do the best that I can and have never once been judgmental or impatient despite the many years of waiting.

And what would my days have been like without my fantastic friends Tim Finegan, Kurt Yanagimachi, Jamie Portsmouth, and Adam Capitano. And Mike Caplan whom I can never thank enough for introducing me to Sivia's book and changing my entire research path. Terry Johnson, it was a blast having you around and I'm going to keep your drawing until it sets off a Geiger counter at the border one day. Luis Alvarez, what's up, WHAT'S UP? And the beautiful ladies Lily Koo and Wendy Prudhomme. I bet you two are still in shock that I became Christian again. You both added a most appreciated touch of elegance to the lab.

Thank you members (over the years) of the Lauffenburger and Griffith labs for your many helpful discussions and questions. I've learned something valuable from each of you. Thank you Gustavo Stolovitzky and IBM for your support in these final months. I'd like to acknowledge the Whitaker Foundation for five years of funding that so greatly enhanced my freedom to grow and learn.

Finally, and especially thank you Huiqin for your love and kindness. I cannot imagine these last years without you. How different and empty they would have been. You've changed my life in ways that I never expected and now cannot do without.

Table of Contents

1 Introduction

1.1 Gene Expression

The workers, tools, and engines of life are proteins, long chains of amino acids folded into unique shapes. Organisms express or build these proteins according to plans coded in units of deoxyribonucleic acid (DNA) called genes [1-3]. A living cell's gene expression and thus its production and inventory of proteins can vary widely. It is this variation in gene expression that allows genetically identical cells to act and appear entirely different and to change over time [4]. For example, every healthy cell in a human body carries the same genetic code yet it is readily apparent that the body is built from many types of cells such as hair cells, skin cells, and blood cells. What makes these cells different are differences in gene expression and this is why measuring gene expression is so important to understanding life on Earth.

When a gene is expressed its DNA is first transcribed into ribonucleic acid (RNA) which is then translated into protein [4]. Since this RNA carries of copy of the gene from DNA to protein it is called messenger RNA (mRNA) and measuring these intermediate mRNA molecules measures one aspect of gene expression. Indeed, cellular mRNA concentration has become the current working definition of gene expression [5]. The two primary goals of gene expression profiling are to measure a sample's mRNA concentration and to determine whether these concentrations differ significantly between samples; if the concentrations differ between two samples, the gene is said to be differentially expressed or simply "delta-expressed".

1.2 Microarray Technology

1.2.1 Hybridization and Architecture

One family of profiling technologies, gene expression microarrays, measure mRNA by using a special property of nucleic acids called hybridization; for each nucleic acid strand there is a unique complimentary nucleic strand that will bind tightly together or hybridize to form a double stranded helix [4]. Microarrays employ a dense array of nucleic acid called "probe" patterned as spots on a substrate such as glass. Each spot contains nucleic acid specific to a particular gene and the field of glass surrounding the spots is treated to block unintended binding of sample or "target" nucleic acid. In this way, the spots become sticky brooms that can sweep through a mixture of target nucleic acid and collect particular genes onto particular spots. We now summarize these technologies [5-12].

Hybridization applies to all nucleic acid including DNA and RNA. Therefore an array can be printed with either DNA to bind mRNA directly or complimentary DNA (cDNA) to bind DNA reverse transcribed from sample RNA. Furthermore, although hybridization is strongest between entire strands of nucleic acid, short units of nucleic acid (oligonucleotides) will also hybridize. The various DNA microarray technologies differ in the manner of printing spot patterns, the spot size, the number of spots, whether spots contain cDNA or DNA, and whether they use entire complimentary strands or shorter oligonucleotides. This thesis applies in whole or in part to any array technology that utilizes hybridization. Arrays using oligonucleotides particularly ones that are relatively short such as Affymetrix arrays are likely to experience greater degrees of cross-hybridization (target nucleic acid binding to non-complimentary or "incorrect" probe molecules). Thus we suggest that our work should be augmented to include theory tying together the multiple oligonucleotide spots intended to bind the same gene that are usually included on short oligonucleotide arrays.

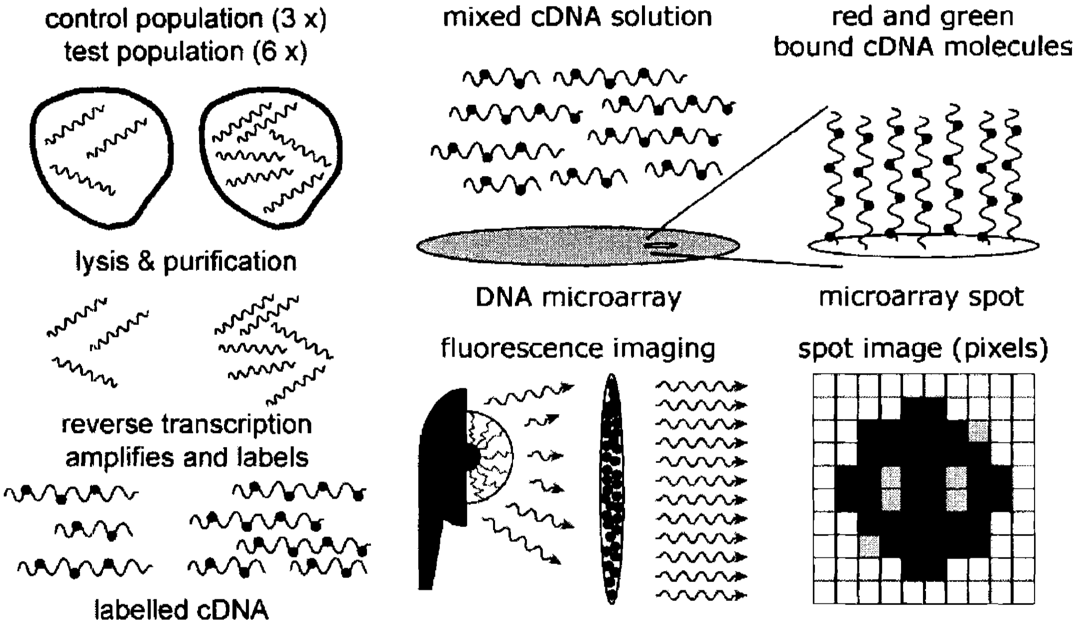


Figure 1.1: Microarray Technology

1.2.2 Co-hybridization

Aside from technological differences there are also two ways of running a microarray. If the array printing technology is tightly controlled and very uniform then we can be reasonably sure that spots for a particular gene are essentially identical from array to array. Therefore, we can hybridize different samples on different arrays and later directly compare these different arrays. One the other hand, if we are unsure that printing quality is uniform then we cannot reasonably directly compare different array measurements. In this case, to compare two samples we must co-hybridize both samples onto the same array thus eliminating the possible variability across different arrays. An extension of this method is to run different samples on different arrays yet co-hybridized with a reference sample. The remainder of this thesis focuses on co-hybridized arrays. However, much of the analysis can be readily extended and in some parts simplified to handle hybridization across multiple arrays.

1.2.3 Labeling Methods

In order measure hybridized nucleic acid the majority of microarray technologies utilize fluorescence labeling. That is, dye molecules are incorporated into or attached to the target nucleic acid. Once hybridized to the array the nucleic acid is fluorescently imaged to obtain a measure of the amount of bound target. However, there are other labeling techniques such as radioactive labeling and other detection methods such as CMOS chips. Our hybridization analysis applies to any of these methods so long as the expected signal in proportional to the number of bound target molecules. Our background analysis applies to any of the imaging methods include radioactivity but alternate detection methods would need to be evaluated on a case-by-case basis.

1.2.4 Sample Preparation

To employ the array, test and control mRNA is harvested and reverse transcribed back into DNA. During reverse transcription fluorescent molecules are incorporated into the DNA to allow quantification by fluorescence; one color is used for the test DNA and another for the control DNA. The labeled DNA samples are washed, purified, then mixed together and the mixture is washed across the array allowing the cDNA spots to collect and hybridize the DNA. Finally the array is washed and dried leaving behind molecules of DNA bound to spots matching their specific gene.

1.2.5 Array Imaging

Once dried, the array is fluorescently imaged to measure the hybridized DNA. A light source illuminates the fluorescent molecules attached to the DNA causing them to excite and emit light of a particular color; one color for test DNA and one color for control DNA. The light is separated by color and collected into detectors to record signals for both test and control DNA across the entire array gene-array-imag. The two imaging systems currently in common use are laser scanners and wide field imagers. Laser scanners scan a laser beam across the array to excite the DNA and use a photo multiplier tube (PMT) to count the photons emitted at each position of the scan. Wide field imagers excite large blocks of the array simultaneously using a lamp and use a charge coupled device (CCD) to count photons in a grid across this block. In both systems, the final result is a grid of photon counts; the grid is imposed in laser scanners by the grid of scan positions and in wide field imagers by the CCD detection grid. As a note, this grid of photon counts or image is an artifact derived from the underlying presumably continuous distribution of fluoresence and with each image pixel being an aggregate of the signal from a small portion of the array. The objective of subsequent array analysis is to process the red and green fluoresence signals to infer the sample-to-control mRNA ratio for every gene.

1.3 Microarray Analysis

Many of the steps described above are physically simple and well understood; however, each of them is a source of variability and, when taken as a whole, they form a complex chain of mechanisms that, as we shall see, has eluded effective analysis for many years. In fact, during the writing of this thesis a news feature in the journal Nature [13] reflected upon the need for effective analytical tools. vital-stats Here are a few quotes from the article:

> "How do you separate significant differences in gene expression from background fluctuation? ... there are no simple answers: interpreting microarray experiments is taxing the skills of even the most adept number-crunchers... If trying to make sense of microarray data has left you with spots circling before your eyes, you're in good company. The problem is perverse ..."
>
> -- Tilstone, Clair University of Bath

> "It's a technical and esoteric topic... There are lots of good statisticians out there, but not as many of them have been exposed to microarray data as are needed."
>
> -- Meltzer, Paul, National Human Genome Research Institute

> "If the collection, analysis and interpretation of the data are flawed then it may not only be a waste of a valuable resource, we could draw faulty conclusions and potentially risk our health and environment."
>
> -- Fisher, Nick, Statistical Society of Australia

Meltzer may be correct in pointing out that not many statisticians have been exposed to microarray data, though there currently about nine hundred papers by more than a hundred authors concerning gene expression analysis. I have personally read more than three dozen papers regarding microarray image analysis alone. Perhaps a more accurate assessment is that DNA microarray technologies utilize physics and chemistry that many statisticians may not be familiar with. Thus, they may lack key insight into the physical mechanisms at play that would otherwise guide and perhaps ease their analytical efforts. In addition, biological systems are notorious for high variability, array experiments are expensive to repeat, and certain experiments may not be repeatable at all. This high variability and scarcity of repeats breaks common statistical methods and renders many of a statisticians tools ineffective.

On the other hand, the scientist familiar with physical, chemical, and biological mechanism are faced with an equally daunting analysis problem. Probability theory treads a narrow path surrounded by subtle pitfalls. Applying it correctly and effectively takes training, experience, and time and often the resulting analytical procedures are often computationally difficult or even impossible with present technology. Add to this the additional burden of understanding contemporary statistical jargon and the problem does indeed seem, as Tilstone comments, perverse.

Finally, even when a hybrid statistician-scientist well versed in both gene expression technology and statistical analysis is available, the simple and yet essential experiments needed to gain insight and understanding my not be readily accessible. It is probably for these reasons rather than a simple lack of exposure that effective array analysis still remains a challenge.

Over the last several years, microarray use exploded with researches everywhere attempting to forge ahead with genome wide studies. Yet, the simple experiments needed to fundamentally understand, analyze, and control the results were never done. The result is that researchers now find themselves drowning in deep pools of data that at best cannot be efficiently summarized, communicated, or shared and at worst seem to make no sense at all. Nick Fisher, in his comment above, points out some of the dangers of proceeding further without first sorting out the analysis of array data.

This thesis grew in an attempt to help meet these needs, to take steps toward melding a physically based quantitative analytical approach to this complex life sciences technology; however, considerable effort is still needed before we will have an approach that we can consider to truly encompass the majority of physical processes inherent in arrays experiments. Though we believe we are taking important steps in the right direction they are only a few steps with many more still needed. What we contribute are carefully designed experiments to elucidate array mechanisms, a physical model that ties the mechanisms together and explains several observed array phenomena, and analytical tools specifically tailored to this physical model.

1.3.1 The State of the Art

Most would agree that if we could repeat an array experiment many times, then debate over analytical procedures would dwindle away and we could use common statistical methods however inefficient to obtain the reasonable estimates of the degree delta-expression and measurement confidence. However, it is also recognized that with biological systems such repetition is often prohibitive or even impossible thereby limiting or eliminating the availability of replicates. To make viable inferences, this lack of experimental information must be compensated with informative prior knowledge in the form of physical models.

We surveyed the previous attempts to bridge this knowledge gap and found a trend toward models of increasing statistical detail but little or no advances in physical understanding. Probably the first model was that of Chen *et al* who modeled the distribution of expression ratios as the ratio of two Gaussian distributions [14]. The only justification offered for this choice were arguments regarding the nimbleness of gene expression. The authors focused on modeling ratios which has a subtle but important drawback. If the choice of what we call "test" or "control" and hence "red" or "green" is simply a matter of naming then half of the possible ratios fall between zero and one while the other half lie in the significantly larger interval of one to infinity. This inherent asymmetry yield a method whose confidence intervals were likewise asymmetric in red and green; in other words, their methods gives diverse conclusions in the case of a dye-swap or simply a change in which mRNA sample the experimenter decides to call red or green. Even if the method were corrected for this asymmetry its chief disadvantage is that is makes artificial assumptions about the global distribution of expression ratios which is in fact the very quantity we are trying to determine. For samples with very few delta-expressed genes this may not be a significant problem however for general expression profiles the assumption would be problematic.

Despite its drawbacks, the use of ratio-based statistics persisted for at least a few years. Newton et al developed a more advanced ratio-based scheme that utilized replicate data an intensity error with multiplicative and additive errors stating they had arrived at this model empirically after examining many data sets, and completed the work with a maximum likelihood estimator [15]. Theilhaber *et al* used a similar error model to develop a Bayesian ratio-based estimator and accompanying algorithm complete with a fantastic acronym, PFOLD [16].

Though the asymmetric ratio model of Chen *et al* remained the only published error model for a few years after the development of microarrays, some authors were simultaneously (or at least soon after) using the log ratio metric rather than simple ratio. Though we cannot trace the exact origin of this log transformation one of the earliest mentions is in the clustering work of Eisen et al [17]. The logarithm transformation is perhaps the most natural and intuitive metric other than ratio. For example, it represents the notion of fold changes (as in two fold, four fold, etc) in a linear scale. Most importantly from a statistical standpoint, the scaling is symmetric for over-expressed and under-expressed genes. That is, regardless of which sample is called test or control or red or green, any conclusions remain physically unchanged. Within two years of Eisen *et al*, a plethora of statistical error models were emerged for the log ratio metric [18-24].

Apart from these general trends there are two notable additions. Many researchers intuitively recognized that not only ratios but also the magnitude of the total intensity should contain information relevant to the measurement confidence. Accordingly, Newton *et al* modeled average spot intensities in an attempt to capture this information content in a spirit similar to Ideker *et al*. In their lofty yet almost comedic sounding "*On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data*" [25] they utilized gamma distributions chosen solely for analytical convenience. It is therefore not surprising that the model does not agree with real array data, failing to capture the intensity covariance typical of microarray measurements. We show a plot of their model fit to real data in our results section. On the other hand, in a fascinating and by comparison rather early work, Audic *et al* derive a novel significance tested by observing the sampling characteristic of sequence tag based expression profiling [26]. In fact, our results will prove similar in some ways to these early results of Audic *et al*. Our molecular distribution work could be viewed as a generalization of the their significance model.

In a final work, Sorger *et al* examined microscopically imaged array spots discovering structure and variability at even smaller scales and believed that, like intensity magnitude, this pixel information might also be relevant to spot quality and measurement confidence [27]. They attempted to capture this by modeling pixel intensities as independent normal deviates and developed a novel measured of spot ratio variability. As far as we know, this is the first work attempting to quantify the quality of a spot based on its microscopic structure. A large portion of our work will be to extend this effort and provide a firmer footing in physical theory for the pixel-to-pixel variation with individual spots.

1.3.2 Conclusions

As you can see, even from our partial review, since the advent of microarrays in 1995 the art microarray statistical analysis has grown considerably. However, there is a crucial point we need to make. With the partial exception of Audic *et al*, none of the works we've cited offer any real physical basis for their error or significance models. That is, they propose various models but never provide a mechanistic explanation for why we should believe their model over any competing models. Therefore, we believe it is fair to classify the state of the art as *ad hoc*. To be sure, many of these model do a fair job of presenting actual data. For example the model of Ideker *et al* was arrived at after examining many data sets and from our experience it is empirically pretty good. On the other hand, the model of Newton *et al* was arrived at for the sake of convenience and fails quite spectacularly. This is the limitation of ad hoc models. While we may find models that fit a given set of data we can have little confidence that they fit a wider range of conditions. On the other hand, if we derive models from a physical theory that we believe hold for a wide range of conditions then we will have far greater confidence in our model to hold under those conditions.

Accordingly, our aim is to develop the beginnings of a microarray model based on physical theory. It will certainly be far from complete but it will be physically based, it will explain some interesting array phenomena, and most importantly it will be testable. That is,, since it will be based on physical hypotheses we will be able to and we shall physically test those hypotheses. Finally, it will gives us a common physical foundation upon which scientist can communicate and combine results with confidence or at least awareness of the underlying physical assumptions rather than a set of *ad hoc* and abstract criteria.

2 Physical Modeling

Finding a mathematical form that describes the behavior of a given physical system cannot be left to chance. Furthermore, even if we chanced upon a form that seemed to describe a set of observations, we would have little power to generalize unless we could explain the form with existing physical theory. Therefore, we believe it is essential to derive analytical tools or statistics from established physical theory. Doing so helps guide us to a reasonable form, allows us to evaluate the application of the form to related experimental conditions, and perhaps most importantly opens the model and its assumptions to physical testing.

In this chapter we apply a simple mass action theory to discern the equilibrium behavior of array hybridization, show how to correct for the imaging artifact of background signal, shown the sufficiency of a simple signal normalization procedure, apply symmetry arguments to evaluate hybridization variability, and finally use Bayesian analysis to derive the resulting analytical procedures needed to infer expression ratio and measurement confidence.

2.1 Mass Action Equilibrium

During co-hybridization red and green target molecules are combined and allowed to come to equilibrium with the array surface. We can model this equilibrium using classical equilibrium thermodynamics. Our work extends the equilibrium derivation of Held *et al* [28] to handle binding of two distinct target species, red and green.

The fundamental physical unit of observation on the array surface is the pixel. This is a region of the surface from which the imaging system counts photons and aggregates a corresponding signal. The pixel surface contains some unknown number of nucleotide molecules called probes or binding sites, each capable of hybridizing a single target molecule. At any moment in time there will be s free probe molecules and r and g bound red and green molecules occupying $r+g$ of the total binding sites s_T within the pixel. If there are r_T and g_T total red and green molecules in the system then a mass balance on the red and green molecules gives

$$\begin{aligned} r_f &= r_T - r \\ g_f &= g_T - g \end{aligned} \tag{1}$$

where r_f and g_f are the free red and green molecules available to bind the pixel element. In principle the hybridization reaction is reversible and hence the equilibrium equations are

$$\begin{aligned} K_r\, r &= r_f\, s \\ K_g\, g &= g_f\, s \end{aligned} \tag{2}$$

where K_r and K_g are the red and green equilibrium constants. Solving these equations allows us to eliminate the number of free molecules

$$\begin{aligned} r &= r_T \frac{s}{K_r + s} \\ g &= g_T \frac{s}{K_g + s} \end{aligned} \tag{3}$$

to give the number of bound red and green molecules at equilibrium as a function of the equilibrium constant and the number of free binding sites. The ratio of the bound red and green molecules is then

$$\frac{r}{g}=\frac{r_T}{g_T}\frac{K_g+s}{K_r+s} \tag{4}$$

2.1.1 Saturating Conditions

If the probe saturates the target (s >> K_r and s >> K_g) or K_r = K_g then the ratio simplifies to

$$\frac{r}{g}=\frac{r_T}{g_T} \tag{5}$$

If instead the target saturates the probe ($s \approx 0$) then the ratio becomes

$$\frac{r}{g}=\frac{r_T}{g_T}\frac{K_g}{K_r} \tag{6}$$

Thus, under saturating conditions for either probe or target, or if the both red and green molecules bind equally then the ratio of red to green bound molecules is simply proportional to the total number of red and green molecules in the starting hybridization solution. Recall that this proportion of red to green total molecules is the very quantity we are trying to assay.

2.1.2 Conclusions

For the remainder of this work we assume that K_r = K_g. This assumption is reasonable considering that good experimental design aims to insure that fluorophores do not affect hybridization and we have not yet seen any direct evidence for probe dependent hybridization. In this case the numbers of bound red and green are proportional and the constant of proportionality or slope of their linear relation is simply the ratio of total red to total green molecules. Though we do not consider the effect that probe dependent hybridization could have, we believe it would be beneficial to explore such effects. We also note that a common assumption for spotted arrays is that the probe saturates the target. That is, there is a vast abundance of possible binding sites. In this case as well probe dependent hybridization would not affect the measured ratio since nearly all of the red and green molecules are captured. However, array technologies that use small binding regions such as Affymetrix arrays can show saturation of the probe by the target [28].

Finally, we wish to point out that since the number of bound molecules in a pixels element varies with the number of free binding sites and consequently the total number of binding sites within that pixel, the total brightness can vary from pixel to pixel due to varying surface properties of the array. Indeed, these total intensity fluctuations can be high if surface uniformity is low. However, these intensity fluctuations are entirely expected and importantly they do not change the proportionality between red and green. Thus, we believe for example that the notion of pin dependent differences in relative red to green hybridization are not sound [22] and we will show that previously observed evidence for this can be attributed to problems with background correction.

2.2 Hybridization Variation

When researchers began using array imagers with pixels much smaller than array spots [27] they expected to see variation in brightness across the spot. After all, DNA printing techniques are not perfect and they will deposit variable amounts of DNA throughout the spot. Some of these variations are quite spectacular such as donuts which are bright rings surrounding dark cores and dark lines caused by scratching during array handling. However, once pixel level detail began to arrive not only did researchers see brightness variation as expected they also saw unexpected variation in color and this variation in color seemed to increase in proportion to overall brightness.

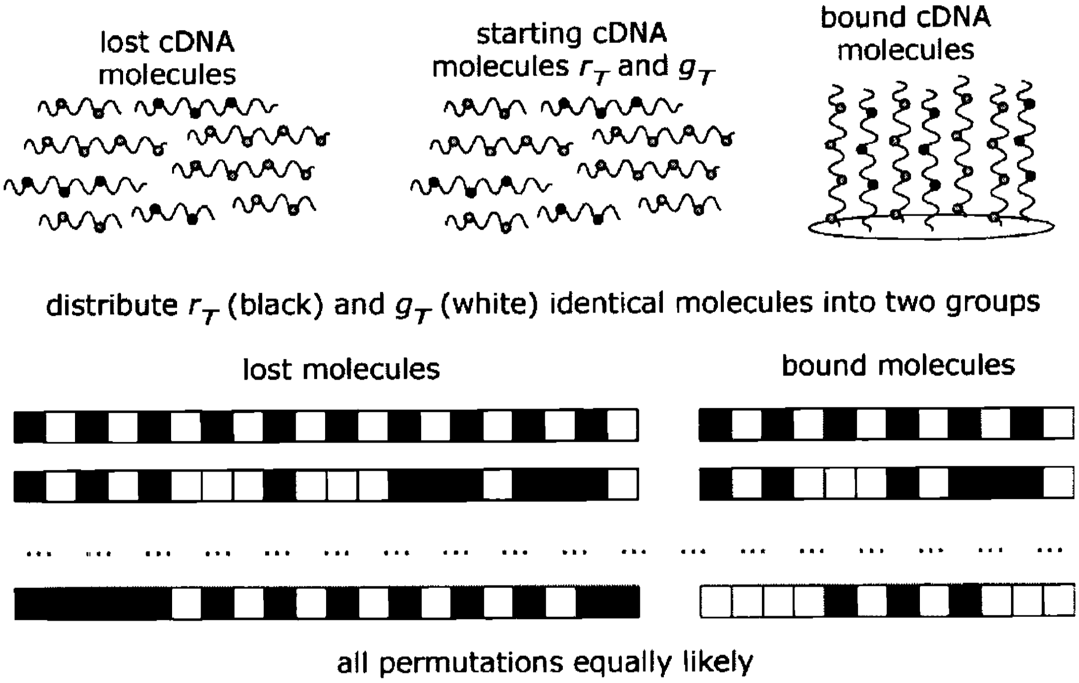
This was unexpected because the primary source of pixel variation was thought to be either cDNA printing or background. As we've discussed cDNA printing does produce brightness variation however there's no physical reason why it should produce color variation since the amount of probe DNA in a pixel should affect both test and control hybridization equally as they compete for binding sites again. On the other hand, background variation is thought to be a fixed process shared throughout the array and thus should not vary with individual pixel brightness. So from where do these color variations originate?

We think the source of color variation is hybridization. During hybridization, the molecules of test and control targets compete with each other to bind probe molecules. We will show that hybridization can create color variability as seen in arrays and that this variability is binomial.

2.2.1 Molecular Distribution

We now introduce a simple formalization of gene expression arrays that relies on the symmetric hybridization of red and green molecules that is the binding equilibrium constants are assumed to be equal. This formalization will allow us to derive again the fundamental result of Equation (5) as well as additional theories regarding the distribution of red and green bound molecules.

No matter the array surface quality, nor hybridization kinetics, nor whether equilibrium is achieved, nor the subsequent handling such as washing or accidental scratching the end result of array processing is always the same: each pixel contains an unknown number, n, of target molecules sampled from the starting r_T and g_T red and green target molecules.


lost cDNA molecules
starting cDNA molecules r_T and g_T
bound cDNA molecules
distribute r_T (black) and g_T (white) identical molecules into two groups
lost molecules
bound molecules
all permutations equally likely

If the various physical processes are color blind in the sense that individual target molecules are equally likely to bind whether they are red or green then this sampling process is equivalent to the classic balls in an urn problems and the resulting distribution of red and green molecules is the hypergeometric distribution

$$p\left[r,g\middle|r_T,g_T\right]=p[r+g]\binom{r_T}{r}\binom{g_T}{g}\binom{r_T+g_T}{r+g}^{-1} \quad . \tag{7}$$

A good approximation for this fundamental distribution is the binomial Gaussian distribution

$$\begin{gathered} f=f_r=r_T/(r_T+g_T) \\ 1-f=f_g=g_T/(r_T+g_T) \\ p\left[r,g\middle|f\right]=\exp\left[-\frac{\left(r(1-f)-gf\right)^2}{2(r+g)f(1-f)}\right]\frac{p[r+g]}{\sqrt{(r+g)f(1-f)}} \\ p\left[r,g\middle|f\right]=\exp\left[-\frac{\left(rf_g-gf_r\right)^2}{2(r+g)f_rf_g}\right]\frac{p[r+g]}{\sqrt{(r+g)f_rf_g}} \end{gathered} \tag{8}$$

where f_r and f_g are the fraction of probe molecules that are red and green respectively. Equation (8) gives the chance that there are r red and g green probe molecules bound within a pixel given that the fraction of red is f. Notice that in both equation (7) and equation (8) there is an unspecified probability p[r+g=n]. This is the prior probability for the total number of bound target molecules. Here we will use Jeffrey's prior p[n]=1/n to express ignorance about the scale of n [29,30]

$$p\left[r,g\middle|f\right]\propto\exp\left[-\frac{\left(rf_g-gf_r\right)^2}{2(r+g)f_rf_g}\right]\frac{1}{\sqrt{(r+g)^3f_rf_g}} \quad . \tag{9}$$

Since Jeffrey's prior cannot be normalized over an infinite range we have expressed the joint probability as a proportionality rather than an equality.

This binomial process causes variation that will increase with total number of bound molecules and thus with increasing brightness. Furthermore, this underlying variation is amplified by fluorescent labeling and subsequent fluorescent imaging each of which can generate multiple counts from a single bound DNA molecule. This amplification allows the binomial color variation to grow in proportion increasing brightness just as observed. To test whether the probability distribution of equation (9) correctly describes the variation of our experimental data we compared the empirical cumulative and P-value to the predicted cumulative and P-value. Is the distribution is accurate then the predicted and empirical probabilities should follow a 1 to 1 correlation.

2.3 Fluoresence

Since we cannot count bound cDNA molecules directly we count the photons fluorescently emitted by their attached labels. Fluorescence is a well-understood classic Poisson process. For every pixel we measure a number of red R and green G signal counts that are related to the number of bound red r and green g cDNA molecules by

$$\begin{aligned} p\left[R \middle| r,\mu\right] &= \frac{e^{-r\mu}\left(r\mu\right)^{R}}{R!} \\ p\left[G \middle| g,\mu\right] &= \frac{e^{-g\mu}\left(g\mu\right)^{G}}{G!} \\ \langle r\rangle &= \sum_{r=0}^{\infty} r\frac{e^{-r\mu}\left(r\mu\right)^{R}}{r\left(R-1\right)!} \\ &= R\mu^{-1} = R\ \eta \\ \langle g\rangle &= \sum_{g-0}^{\infty} \frac{e^{-g\mu}\left(g\mu\right)^{G}}{g\left(G-1\right)!} \\ &= G\mu^{-1} = G\ \eta \end{aligned} \tag{10}$$

where η is the brightness parameter representing the expected number of target molecules per emitted photon. It is affected by a number of optical factors including labeling efficiency, dye quantum yield, detector efficiency, exposure time, pixel size, etc. Here red and green share the same brightness factor because background correction and normalization is presumed to correct for differences between the red and green channels including any channel bias. However, the constant of proportionality is unique for each gene in the array. We cannot necessarily assume that it is same for different genes because DNA molecules can easily incorporate gene specific amounts of fluorescent probe simply by virtue of having a different sequence and thus a different number of potential incorporation sites for any given labeled nucleotide. Using these equations we can transform p[r,g|f] to p[R,G|f] which the probability of a red and green signal pair from a single pixel

$$\begin{aligned} p\left[R,G|\mu,f\right] &= p\left[r,g|f\right]\begin{vmatrix} \partial r/\partial R & \partial r/\partial G \\ \partial g/\partial R & \partial g/\partial G \end{vmatrix} \\ &= p\left[r,g|f\right]\begin{vmatrix} \eta & 0 \\ 0 & \eta \end{vmatrix} \\ &= p\left[r,g|f\right]\eta^2 \end{aligned} \tag{11}$$

which after a substituting equation (9) gives the desired probability of a pixel intensity pair given the reciprocal brightness

$$\begin{aligned} p\left[R,G|\eta,f\right] &\propto \exp\left[-\frac{\left(r f_g - g f_r\right)^2}{2\left(r+g\right)f_r f_g}\right]\frac{\eta^2}{\sqrt{\left(r+g\right)^3 f_r f_g}} \\ &\propto \exp\left[-\frac{\eta^2\left(R f_g - G f_r\right)^2}{2\eta\left(R+G\right)f_r f_g}\right]\frac{\eta^2}{\sqrt{\eta^3\left(R+G\right)^3 f_r f_g}} \quad . \\ &\propto \exp\left[-\frac{\eta}{2}\frac{\left(R f_g - G f_r\right)^2}{\left(R+G\right)f_r f_g}\right]\sqrt{\frac{\eta}{\left(R+G\right)^3 f_r f_g}} \end{aligned} \tag{12}$$

Having the probability of a single pixel the probability of an entire spot or set of N pixels is simply the joint probability of the individual pixels

$$\begin{aligned} p\left[\{R,G\}|\eta,f\right] &= \prod_{\{R,G\}} p\left[R,G|\eta,f\right] \\ &\propto \prod_{\{R,G\}}\left(\exp\left[-\frac{\eta\left(R f_g - G f_r\right)^2}{2\left(R+G\right)f_r f_g}\right]\sqrt{\frac{\eta}{\left(R+G\right)^3 f_r f_g}}\right) \\ &\propto \exp\left[-\frac{\eta}{2}\sum_{\{R,G\}}\frac{\left(R f_g - G f_r\right)^2}{\left(R+G\right)f_r f_g}\right]\prod_{\{R,G\}}\sqrt{\frac{\eta}{\left(R+G\right)^3 f_r f_g}} \\ &\propto \exp\left[-\frac{\eta}{2}\sum_{\{R,G\}}\frac{\left(R f_g - G f_r\right)^2}{\left(R+G\right)f_r f_g}\right]\left(\frac{\eta}{f_r f_g}\right)^{\frac{N}{2}}\prod_{\{R,G\}}\frac{1}{\left(R+G\right)^3} \end{aligned} \tag{13}$$

The brightness parameter η is unknown and not of direct interest. Such parameters are called nuisance parameters and can be eliminated by marginalization that is by integrating over the possible values of the parameter. This yields

$$\begin{aligned} p\left[\{R,G\}|f\right] &= \int_0^\infty p\left[\{R,G\}|\eta,f\right]p\left[\eta\right]d\eta \\ &\propto \int_0^\infty \exp\left[-\frac{\eta}{2}\sum_{\{R,G\}}\frac{\left(R f_g - G f_r\right)^2}{\left(R+G\right)f_r f_g}\right]\left(\frac{\eta}{f_r f_g}\right)^{\frac{N}{2}}\frac{d\eta}{\eta}\prod_{\{R,G\}}\frac{1}{\left(R+G\right)^3} \\ &\propto \left(\sum_{\{R,G\}}\frac{\left(R f_g - G f_r\right)^2}{\left(R+G\right)}\right)^{-\frac{N}{2}}\prod_{\{R,G\}}\frac{1}{\left(R+G\right)^3} \end{aligned} \tag{14}$$

which is the probability of a spot given f only. Here the product and summation are over all pixels within a spot.

2.4 Bayesian Analysis

Now we use Bayes theorem [29,30] to invert the probability distribution to obtain the probability of f given a spot

$$
\begin{aligned}
p\left[f \middle| \{R,G\}\right] &\propto \left(\sum_{\{R,G\}} \frac{\left(R f_g - G f_r\right)^2}{(R+G)} \right)^{-\frac{N}{2}} p[f] \\
p\left[f \middle| \{R,G\}\right] &\propto \left(\sum_{\{R,G\}} \frac{\left(R f_g\right)^2 - 2R\,G f_r f_g + \left(G f_r\right)^2}{(R+G)} \right)^{-\frac{N}{2}} p[f] \\
&\propto \left(RR\, f_r^2 - 2RG\, f_r f_g + GG\, f_g^2\right)^{-\frac{N}{2}} p[f]
\end{aligned}
\tag{15}
$$

In the final form the summation over the data must be computed only once to derive the statistics

$$
\begin{aligned}
RR &\equiv \sum_{\{R,G\}} \frac{R^2}{(R+G)} \\
RG &\equiv \sum_{\{R,G\}} \frac{R\,G}{(R+G)} \\
GG &\equiv \sum_{\{R,G\}} \frac{G^2}{(R+G)}
\end{aligned}
\tag{16}
$$

which can then simply reused for new values of *f*. This distribution is akin to the Student-T distribution. Since the generally accepted measure of gene expression is log ratio

$$
\begin{aligned}
\rho &\equiv \log\left[\frac{f}{1-f}\right] \\
f &= \frac{e^{\rho}}{1+e^{\rho}}
\end{aligned}
\tag{17}
$$

we transform this probability distribution to the log ratio domain

$$\begin{aligned} p\left[\rho\middle|\{R,G\}\right] &\propto \left(RR\frac{1}{\left(1+e^{\rho}\right)^{2}} - 2RG\frac{e^{\rho}}{\left(1+e^{\rho}\right)^{2}} + GG\frac{e^{2\rho}}{\left(1+e^{\rho}\right)^{2}} \right)^{-\frac{N}{2}} p[\rho] \\ &\propto \left(\frac{\left(e^{\rho/2}+e^{-\rho/2}\right)^{2}}{e^{-\rho}RR - 2RG + e^{\rho}GG} \right)^{\frac{N}{2}} p[\rho] \end{aligned} \tag{18}$$

where we can approximate the mean and variance by

$$\begin{aligned} \hat{\rho} &= \log\left[\frac{RR+RG}{GG+RG}\right] \\ v_{hyb} &\approx \left(-\frac{d^{2}}{d\rho^{2}} \log\left[p\left[\rho\middle|\{R,G\}\right] \right] \Big|_{\hat{\rho}} \right)^{-1} \\ &\approx \frac{RR\,GG - RG^{2}}{N} \left(\frac{RR+2RG+GG}{(RR+RG)(GG+RG)} \right)^{2} \end{aligned} \tag{19}$$

Since the mean and variance are computed from the uncertain foreground values we would like to propagate this uncertainty into these estimates. We handle this computationally using the standard error propagation rules given by

$$\begin{array}{lcll} Var(\;x \pm y\;) & = & Var(x) & + \; Var(y) \\ Var(\;x \,\mathrm{g} y\;) & = & Var(x)\,y^{+2} & + \; Var(y)\,x^{+2} \\ Var(\;x / y\;) & = & Var(x)\,y^{-2} & + \; Var(y)\,x^{+2}\,y^{-4} \\ Var(\;\log[x]\;) & = & Var(x)\,x^{-2} & \end{array} . \tag{20}$$

This error propagation is actually quite easy to implement in C++ by defining a class representing uncertain quantities that maintains both a mean and variance. Then for example Equation (19) is implemented using this class rather than the usual floating-point numbers. Below is our implementation of such a class.

```
class Uncertain ;

Uncertain exp ( Uncertain ) ;
Uncertain log ( Uncertain ) ;

class Uncertain {

            friend Uncertain exp ( Uncertain ) ;
            friend Uncertain log ( Uncertain ) ;

            double _mean ;
            double _vari ;

      public :

            Uncertain ( ) ;

            Uncertain (
                  double  mean ) :
                  _mean ( mean ) , _vari ( 0.00 ) { }

            Uncertain (
                  double  mean   , double  vari ) :
                  _mean ( mean ) , _vari ( vari ) { }

            double mean ( ) const { return _mean ; }
            double vari ( ) const { return _vari ; }

            void mean ( double mean ) { _mean = mean ; }
            void vari ( double vari ) { _vari = vari ; }

            Uncertain & operator+= ( Uncertain const & ) ;
            Uncertain & operator-= ( Uncertain const & ) ;
            Uncertain & operator*= ( Uncertain const & ) ;
            Uncertain & operator/= ( Uncertain const & ) ;

            Uncertain operator+ ( Uncertain u ) { return u += *this ; }
            Uncertain operator- ( Uncertain const & u ) { Uncertain
t(*this) ; return t -= u ; }
            Uncertain operator* ( Uncertain u ) { return u *= *this ; }
            Uncertain operator/ ( Uncertain const & u ) { Uncertain
t(*this) ; return t /= u ; }

} ;

inline Uncertain & Uncertain::operator+= ( Uncertain const & u ) {

      _mean += u._mean ;
      _vari += u._vari ;

      return *this ;

}

inline Uncertain & Uncertain::operator-= ( Uncertain const & u ) {

      _mean -= u._mean ;
      _vari += u._vari ;

      return *this ;

}

inline Uncertain & Uncertain::operator*= ( Uncertain const & u ) {

      _vari *= u._mean * u._mean ;
      _vari += u._vari * _mean * _mean ;
      _mean *= u._mean ;

      return *this ;

}

inline Uncertain & Uncertain::operator/= ( Uncertain const & u ) {

      double umr  = 1.0 / u._mean ;
      double umrs = umr * umr       ;

      _mean *= umr  ;
      _vari *= umrs ;
      _vari += u._vari * _mean * _mean * umrs ;

      return *this ;

}

Uncertain operator+(Uncertain u, double d) { return u+=Uncertain(d) ; }
Uncertain operator-(Uncertain u, double d) { return u-=Uncertain(d) ; }
Uncertain operator*(Uncertain u, double d) { return u*=Uncertain(d) ; }
Uncertain operator/(Uncertain u, double d) { return u/=Uncertain(d) ; }
Uncertain operator+(double d, Uncertain u) { return Uncertain(d)+=u ; }
Uncertain operator-(double d, Uncertain u) { return Uncertain(d)-=u ; }
Uncertain operator*(double d, Uncertain u) { return Uncertain(d)*=u ; }
Uncertain operator/(double d, Uncertain u) { return Uncertain(d)/=u ; }
```

The end result is two measures of variance; one representing the uncertainty of hybridization and one representing background uncertainty. There is one additional variance that must be included at that is the sample preparation variance. The only way we currently know of to estimate this variance is by spiking in same to same controls. Since these should have an expected log ratio of zero (after background correction normalization) their variance serves as an estimate of the variance caused by sample preparation. Of course, the some of the variance observed in the log ratio is caused by hybridization and background. Therefore, we must subtract the estimated hybridization and background variance for each same-to-same control from the total variance for all same-to-same controls. This quantity divided by the number of samples is then the estimate for the sample preparation variance v_{prep}. The probability density for a gene is then given by normal distribution $p\left[\rho\middle|\{R,G\}\right]=N\left(\hat{\rho},v_{prep}+v_{hyb}+v_{cor}\right)$ and the probability of differential expression is then given by $\theta/(1+\theta)$ where

$$\theta=\frac{p\left[\rho=0\middle|\{R,G\}\right]}{\int_{\rho=-\infty}^{\rho=\infty}p\left[\rho\middle|\{R,G\}\right]p[\rho]d\rho}\quad.$$

3 Background And Normalization

The theory of the previous section indicated that the equilibrium red and green signals should be proportional. To investigate this we conducted a same-to-same array experiment replicated four times. In a same-to-same experiment a single mRNA sample is split to form the test and control samples. Therefore, we would expect not only that the red and green signals are proportional but also that their proportionality constant is one. Since array signals typically range over several orders of magnitude it is common to view them in log transformed space.

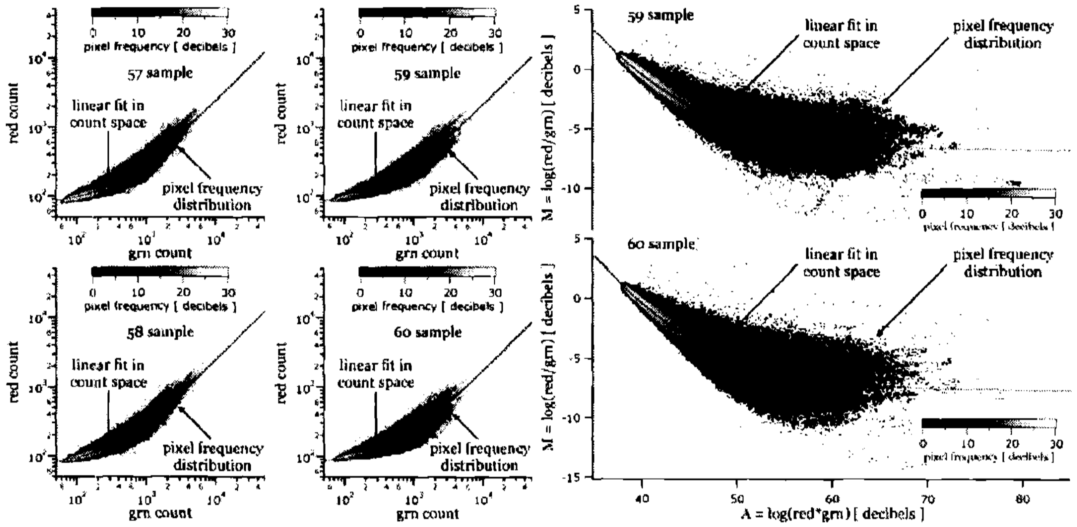


Figure 3.1 : Apparent Non-Linearity of Microarray Data

Figure 2.1 shows the distinct banana shape that is common with raw microarray data in log-transformed spaces. The curved distributions have vexed many researchers over the years including us and caused some to believe there is an inherent non-linearity of array data. This has led some authors to develop non-linear correction or normalization schemes such as LOESS [22], rank-invariant [31], and variance stabilization [32].

However, it turns out that the explanation for this apparent curvature is exceedingly simple even verging on the trivial. The fact is we should not have been surprised to see curves in these log-transformed spaces because the log transform is itself non-linear. In fact, a straight will remain a straight line after transformation if and only if it intersects the origin. This means that if there is any constant offset in either the red or green channel we will see curves in these transformed spaces. And the key is that background processes act as an offset from the origin thereby causing the curvature shown in Figure 2.1. We will explain background in more detail, derive a new corrective procedure, and demonstrate that the procedure removes the illusory curvature.

3.1 General

Array imagers are intended to measure foreground signal from hybridized DNA within spots. However, they will measure signal from every spot regardless of whether it contains hybridized DNA or not and they will even measure signal from the field. To paraphrase George Orwell, this unwanted and unavoidable background signal falls upon the foreground like a soft snow blurring the outlines and covering up the details. And in science, the search for truth, the last thing we want is a snow job.

If we knew the exact value of the background in each pixel we could simply subtract it from the total signal to recover the foreground. Unfortunately, background and foreground signals are indistinguishable making their individual values unknown thus requiring probabilistic correction. To derive an appropriate correction we first consider the major sources of background.

3.2 Sources

There are a variety of background sources. The majority of which are physically unavoidable. There are three broad categories of background signal: radiation, thermal noise, and material fluorescence.

3.2.1 Radiation

The same properties that allow array imagers to detect photons of light from fluorescent molecules also cause them to detect other radiation sources. Array imaging detectors are subject to at least five unwanted sources of radiation including excitation leakage, thermal radiation, gamma rays, cosmic rays, and nuclear decay.

The fluorescent molecules attached to the sample DNA will only emit light if they are excited by an energy source. In array imagers the energy source is light provided by a laser beam or powerful lamp. Of course, this light can generate signal just as easily as the fluorescent DNA. To prevent the excitation light from generating signal, array imagers employ filters, beam splitters, dichroic mirrors, or other such light handling devices. However, these devices are not perfect and some excitation light leaks into the detector and generates signal.

Thermal radiation is the light emitted by all materials at temperatures above absolute zero. A clear demonstration is the red glow of a hot iron or other metal from whence the saying red hot originates. As the temperature increases the spectrum and intensity of light changes following Plank's Law. At temperatures typical of laboratories the effect is of course not visible. None the less, black body photons from any material surrounding an imaging detector can produce signal however negligible it might be.

High energy photons and particles abound in space and the earth is daily bombarded by them. Both the photon, gamma rays, and particles, cosmic rays, can interact directly with photon detector. The cosmic rays can also interact with particles in the atmosphere to produce additional gamma rays. Both gamma rays and cosmic rays possess far more than enough energy to generate large signals often in array detectors often saturating the detector and in the case of CCDs spill or bloom into adjacent pixels.

Apart from outer space, there are local sources of high energy photons and particles. Many elements surrounding us occur as isotopes that undergo nuclear decay producing high energy radiation. For example, Potassium isotope 40 undergoes nuclear decay and is present in many common materials including our own bodies so detectors can easily pick up charge from this source. Some optical lenses even contain as much as twelve weight percent potassium and will readily add to background signal.

3.2.2 Thermal Energy

Thermal energy can produce signal directly in the detectors by knocking electrons off the semiconductor atoms thereby generating charge. These thermal electrons or thermions are identical to electrons knocked off by photons and so are recorded as signal. As with black body radiation the effect is reduced as the detector is cooled. CCDs for example are almost always cooled to temperatures below room temperature to help reduce background.

3.2.3 Material Fluorescence

For gene expression arrays the dominant source of background is most likely material fluorescence. After hybridization the array is washed in an effort to remove unbound cDNA, free dye, oils, dust, or any other stray material leaving behind only the array surface and hybridized cDNA. Of course the process is imperfect and leaves behind material that will fluorescence and produce signal. Furthermore, nucleic acids are themselves weakly fluorescent and can generate additional signal

3.3 Estimation

Despite knowing the many sources of background, we have yet to find a satisfactory background model. This is understandable considering the variety and complexity of background sources particularly material fluorescence. Attempting to theoretically model material fluorescence is like trying modeling a dusty coffee stain partially wiped by a moist sponge.

Though we can not theoretically model background nor experimentally separate background and foreground signals we can attempt to experimentally ensure that some parts of the array generate only background signal. For example, we can print spots with no DNA or DNA entirely different from any sample DNA thereby creating spots that will not be complimentary to any sample genes and hence should not hybridize. These negative control spots undergo the same array processing as spots containing complimentary DNA and except for lacking specific hybridization they should undergo the same physical processes that cause background in spots elsewhere in the array. Thus, these negative control pixel serve as samples of the array background.

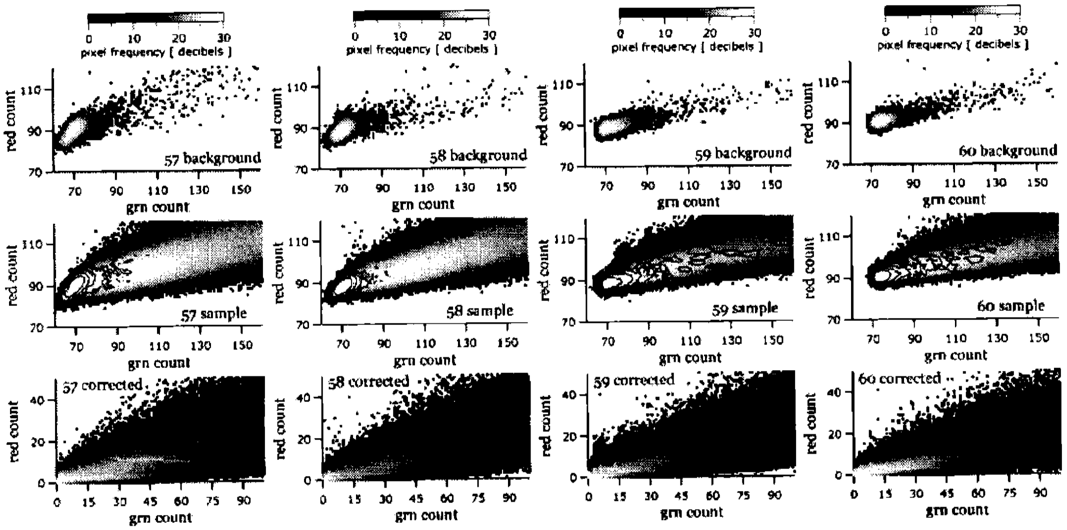


Figure 3.2 : Background Sampling and Correction

The top four panels of Figure 3.2 show the frequency distribution of background counts from negative control spots shown for four different arrays. The distribution is fairly centralized and shows dependence between red and green signals. However, it is not symmetric enough to be modeled well by a bivariate Gaussian and we have yet to find a suitable model. The middle four panels show the frequency distribution of the sample pixels from the array. Interesting, you can see a faint "after- image" of the background distribution which we have highlighted by overlaying contours for the background frequency distribution. This after-image is a good sign. Because it falls nicely on the origin of the sample pixels it means that the background processes in the sample spots are similar to the background processes of the negative control spots. Therefore, using negatives controls to estimate background signals in the sample spots is at least feasible. The bottom four panels show background corrected data which is pulled nicely back to mathematical origin (0,0). We will discuss this corrective procedure in a following section.

In passing, we need to make an important point. Despite it's widespread use, the array field does not serve as a good sample of spot background. The most vivid proof of this are the commonly observed black holes [27]. These are spots whose pixel intensities fall below the surrounding field intensity. Far from rare they appear frequently in arrays that have moderate or even low average field intensity. We believe this clearly demonstrates that treatment differences between spots and field substrate affect background signal mostly likely through differences in material fluorescence such as non-specific binding of free dye. Thus, we should not expect surrounding field to serve directly as a good estimate for spot background and using as such is an error. There may well be a relationship between field and spot background but that is as yet undetermined and it is unlikely to be a simple relation and certainly it is not one of equivalence.

Once we have the background pixel samples from negative controls, what shall we do with them? Since we do not have a prior background model for which we could derive parameters we simply use the background pixel frequency distribution as an empirical background probability distribution. Thus, it is in essence a very ignorant empirical free-form model with as many parameters as there are unique pixel values.

3.4 Subtraction

As mentioned before, if we knew the exact value of the background signal we could subtract it from the total signal to recover the foreground signal. A slight modification of this procedure is to subtract the mean of the background probability distribution from the total signal. The justification for this procedure is the linearity of expectation. The expected value of the foreground intensities *R* and *G* is given by

$$\begin{aligned}\langle R\rangle &= \langle R_M - R_B\rangle \\ \langle G\rangle &= \langle G_M - G_B\rangle \\ \langle R\rangle &= R_M - \langle R_B\rangle \\ \langle G\rangle &= G_M - \langle G_B\rangle\end{aligned} \qquad (21)$$

where R_m and G_m are the measured intensities and are R_B and G_B the background intensities. It is precisely this procedure that is widely used and widely referred to as background subtraction. Unfortunately, it is wrong. Stark evidence for the error of background subtraction is the fact that the resulting foreground signals are often negative, which is certainly physically meaningless. Indeed, some authors have even stopped attempting any manner of background correction at all recognizing that background subtraction produces errors and believing there is no alternative.

3.5 Correction

The error in the derivation of (21) is that it fails to account for the fact that possible background values are logically dependent on the total signal since the background signal cannot be greater than the total signal; a part is never greater than the whole. Therefore the expected value of the background is actually a function of the measured signal. Since background cannot be higher than the total number of counts, the background expectation must be taken only over values up to the total count yielding

$$\begin{aligned}
\langle R\rangle &= \langle R_M - R_B\rangle \\
\langle G\rangle &= \langle G_M - G_B\rangle \\
\langle R\rangle &= R_M - \langle R_B | R_M, G_M\rangle \\
\langle G\rangle &= G_M - \langle G_B | R_M, G_M\rangle \\
\langle R\rangle &= R_M - \sum_{R_B=0}^{R_M}\sum_{G_B=0}^{G_M} R_B p\left[R_B, G_B | R_M, G_M\right] \\
\langle G\rangle &= G_M - \sum_{R_B=0}^{R_M}\sum_{G_B=0}^{G_M} G_B p\left[R_B, G_B | R_M, G_M\right]
\end{aligned} \tag{22}$$

which compute expected foreground or corrected signal. This simple result ensures that corrected signals are logically consistent; one side effect is that negative foreground values never occur. The next four two figures show the background corrected pixels for same-to-same arrays. Though a seemingly small and simple correction to the erroneous background subtraction, coming chapters will show that background correction has an essential impact on the accurate interpretation of array data.

3.6 Uncertainty

In addition to correcting for the background by computing the expected foreground signal, we wish to quantify the uncertainty we have in this corrected foreground signal. In other words, since we know only the background probability distribution and not the exact value of the background our quantification of the foreground is also a probability distribution and we would like to quantify the spread of this distribution. This is easily done by computing the variance of the foreground intensities

$$\begin{aligned} v_{bg,R} &= \sum_{R_B=0}^{R_M}\sum_{G_B=0}^{G_M}\left(R_M - R_B - \langle R\rangle\right)^2 p\left[R_B,G_B|R_M,G_M\right] \\ v_{bg,G} &= \sum_{R_B=0}^{R_M}\sum_{G_B=0}^{G_M}\left(G_M - G_B - \langle G\rangle\right)^2 p\left[R_B,G_B|R_M,G_M\right] \end{aligned} \quad . \qquad (23)$$

Later this uncertainty will be propagated using the usual error propagation rules (Sivia) into our computed statistics.

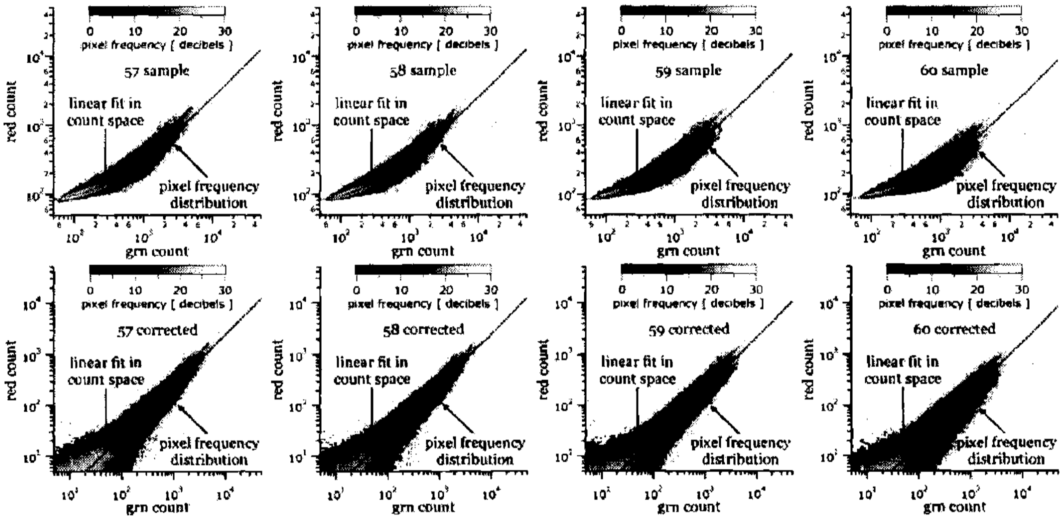


Figure 3.3: Linearity Revealed

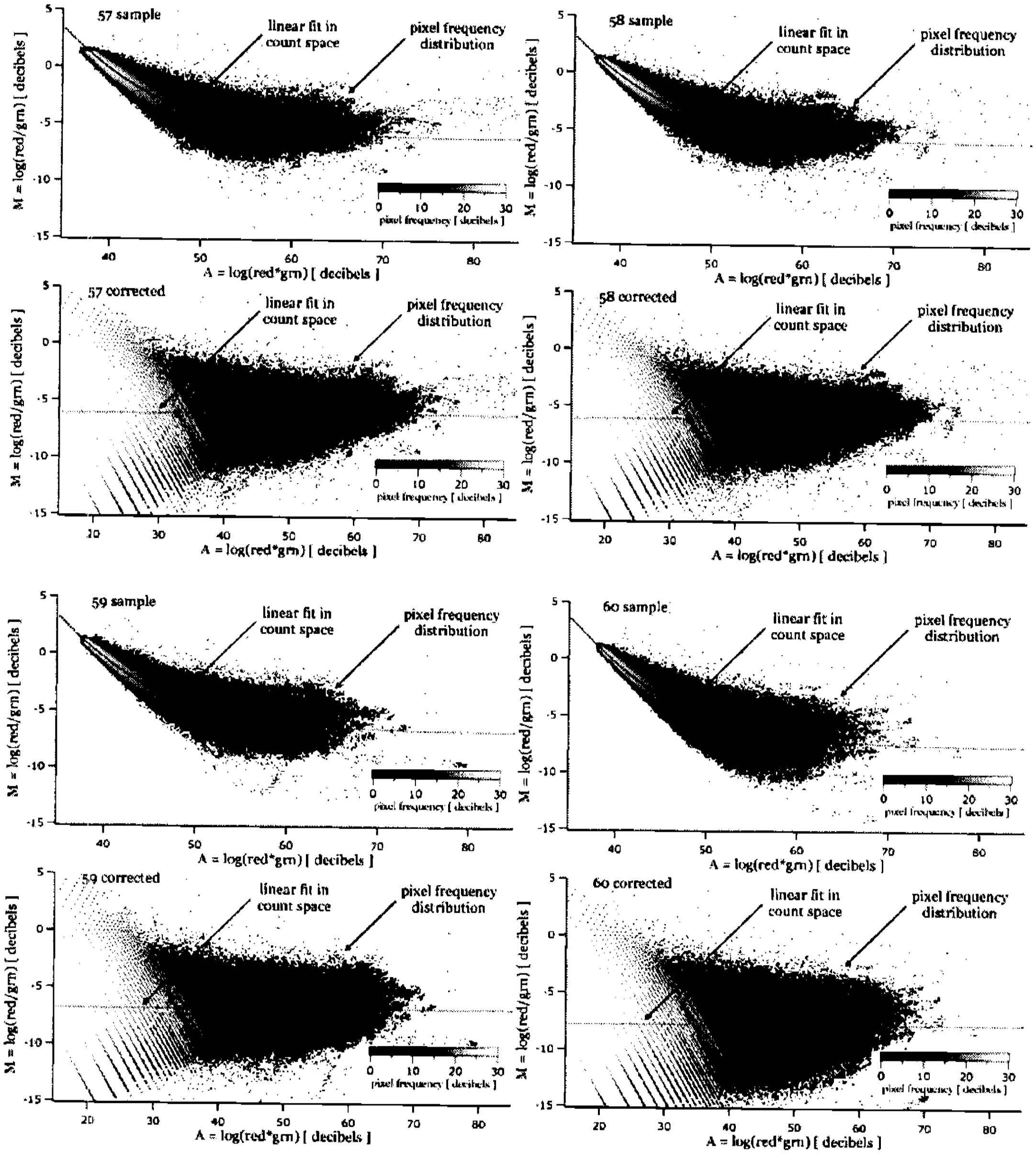


Figure 3.4: Linearity of M vs A

When the simple background correction is applied to the same-to-same data shown previously the effect is quite striking. Log space distributions that previously appeared curved are now straight and the banana shaped M vs A curve has also fallen flat. We believe this supports the proportionality of red and green signals derived from mass action.

Often we are asked about the apparent increase in variability in M vs A space at especially at low values of A. Many believe that this is in some way a disadvantage of background correction. Such beliefs are understandable though they are in fact misconceptions. The beliefs derive from the entirely correct belief that decreasing experimental variability is a good thing. That is, if we replicate an experimental setup we expect that it will produce the nearly the same measurements; this is the heart of repeatability which is after all a core component of science. Of course, there will be small variations among our experiments no matter how precisely we attempt to control the conditions. Therefore, if we can reduce this variability then we are demonstrating greater control and understanding of the experiment. The critical point here is that we are describing reduction in *experimental* variability not *mathematically transformed* variability.

For example, I could arbitrarily add the value 1,000,000 to all the measured intensities and in doing so I would drastically decrease the variability of the resulting computations of M. Obviously, this is not in any sense an improvement since it in no way leads us closer to any truth. This is precisely the same affect, though to a much lesser degree, that background has on data. It adds an unwanted quantity to the data thereby obscuring the variability of the true underlying signal. Correcting for this background does nothing more than reveal this hidden variation. In other words, it melts the snow job, it pulls back the curtain to reveal the features of the true data. This is a good thing.

3.7 Normalization

The M vs A plots of background corrected data shown in the previous chapter reveal a second array phenomena. Since this data is from a same-to-same arrays we would expect that the pixel distribution would be centered around M = 0 in the M vs A plot. However, it is evident that the entire distribution is shifted well below M = 0. This phenomenon is channel bias and must be corrected by a normalization procedure.

Array protocols attempt to minimize experimental variation and bias by treating target samples as similarly as possible. However, there are inevitable differences of treatment between different target samples. The clearest example is that target samples hybridized to the same array must be labeled with different fluorophores otherwise they would be indistinguishable. These fluorophores will certainly have different optical properties leading to different quantum yields, detection efficiencies, and other optical properties. Furthermore, if the target nucleotides are labeled directly by incorporating fluorophores during reverse transcription, the fluorophores may incorporate at different rates. Another example is that different nucleotides samples must be kept separate until labeling and therefore will undergo separate reverse transcription reactions and reaction rates can vary from batch to batch. Normalization is the meant to correct for the biases caused by differing treatment of target samples.

The simple fact is, we do not have sufficient understanding of these sources of bias to theoretically correct for them. Therefore, the only general way to normalize arrays is by empirical calibration. Calibration controls are known targets included or spiked equally or as a dilution series into the test and control samples. The calibration controls undergo the same treatment as their host sample and then hybridize to matching control spots printed on the array. After hybridization, imaging, and background correction the signals from these calibration spots are used to determine the relative bias between the test and control treatments.

From the M vs A and consideration of the physical processes we believe this bias acts simply to scale the counts of one channel with respect to another and thus appears as an offset on the M axis. This general approach is in fact quite common and we believe well justified. To correct for this we simply total the red and green background corrected signals, R_T and G_T, of the control spot pixel counts, and compute a normalization factor R_T/G_T used to scale the red and green pixel counts for the entire array. However, the correction must be computed from controls not from samples genes that have potentially unknown proportions. Thus, we believe methods such as the so-called "global" normalization is flawed. They rely on assumptions about the distribution of gene expressions that are generally not justified. Another example is rank invariant normalization for which the authors provide no physical justification and we cannot think of any ourselves.

We note that some authors have proposed the need for non-linear calibrations that depend on signal intensity a canonic example being the LOESS method [22]. However, we find no physical basis for such a proposal. For example the authors propose that differences in print quality from spot to spot may require such non-linear corrections. We do not agree. As seen from equations (3) and (5) a spots quality can affect total binding but not the relative amounts of red and green. Another possibility often proposed is detector saturation. However, if such saturation were occurring as a result of poor experimental design the needed correction would be more intricate than a simple empirical calibration either linear or non-linear. It is best to experimentally avoid saturating conditions. Finally, we have shown the same evidence used to justify non-linear corrections are in fact artifacts resulting from background artifacts that can be eliminated by background correction.

4 Results

From the model we can make three key predictions. The first is the linear dependence of red signal on green signal. The second is the linear dependence of signal on mRNA concentration. The third is binomial variability of pixel intensities within a spot. We will verify these predictions using two types of experiments. The first type of experiment is a same-to-same experiment. In this experiment a single mRNA sample is split into two samples serving as both test and control. This experiment will verify the linearity of red signal on green signal and the binomial variability of pixels. The second type of experiment is a dilution series in which different genes are prepared at known mRNA concentrations in test and control samples. This will be used to verify the linear dependence of signal on mRNA concentration.

In Chapter 3 we verified the linearity of red and green signals after background correction. We have yet to verify that the constant of proportionality corresponds to the concentration of red to green total molecules

4.1 Linearity of Signal and Concentration

Finally the model predicts a linear dependence of signal on mRNA concentration or equivalently that the slope of red versus green is given by the ratio of red to green labeled mRNA. To verify this we printed an array having eight different genes each printed in many replicate spots. Then samples of each of the eight mRNAs were prepared in two solutions at different concentrations shown in Table 4.1 As Figure 4.1-4.3 clearly shows the data fall nicely on a set of lines. Again in logarithmic space these lines appear as curves emanating from a common background offset. By fitting lines to these data we can recover estimates of the slope and compare this to the ratio of mRNA concentrations. As Figure 4.4 shows, the calibration curve is nicely linear. We also show what happens if the simple ratio without background correction is used. The simple ratio becomes increasingly inaccurate as the concentration and thus the signal lowers further into the background region.

gene number	sample A mass [ng]	sample B mass [ng]	ratio A:B
1	0.02	1.00	1:50
2	0.10	1.00	1:10
3	0.50	1.00	1:2
4	1.00	1.00	1:1
5	1.00	1.00	1:1
6	1.00	0.50	2:1
7	1.00	0.10	10:1
8	1.00	0.02	50:1

Table 4.1: Dilution Ratios

4.2 Binomial Variation

The model predicts a binomial variation for the pixels counts within a spot. To verify the presence of binomial variability in the data we compare the predicted cumulative distribution to the empirical cumulative distribution and the predicted p-value to the empirical p-value. If the predicted distribution is consistent with the data then these comparisons will match and form a straight line. Figure 4.5 shows the results for the four same-to-same arrays. The binomial form seems highly consistent with the data. By contrast a simple Gaussian does not accommodate the data as we show an example for one of the data sets.

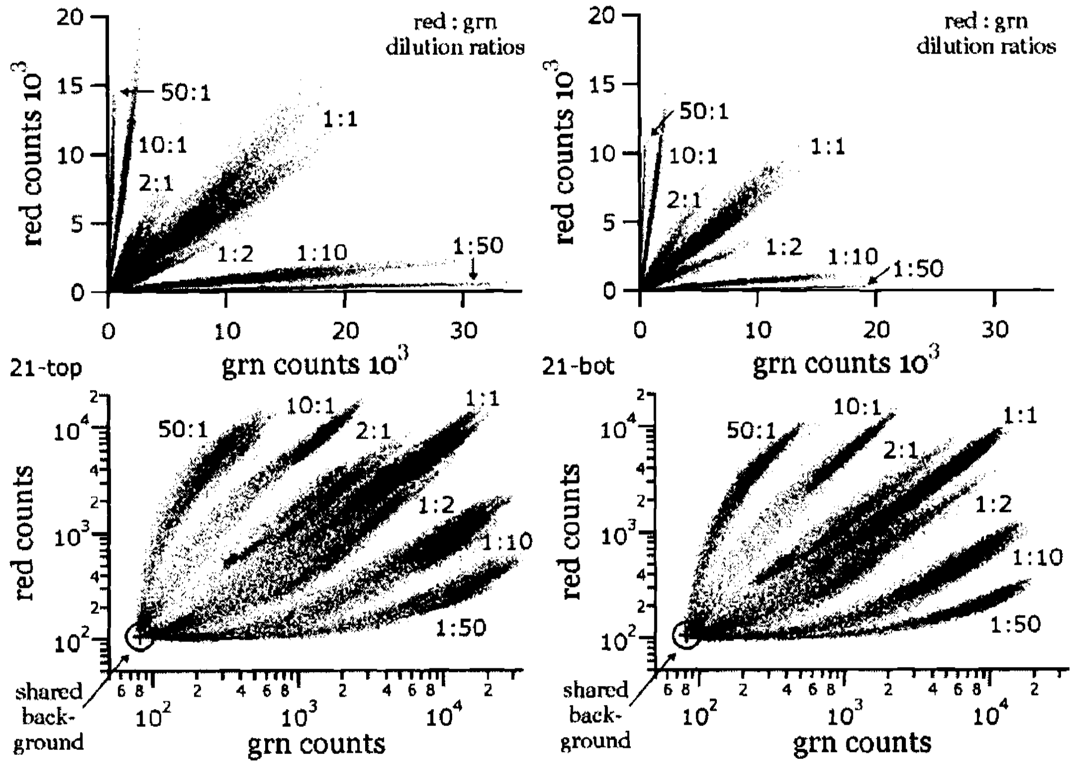


Figure 4.1 : Dilution Series 21

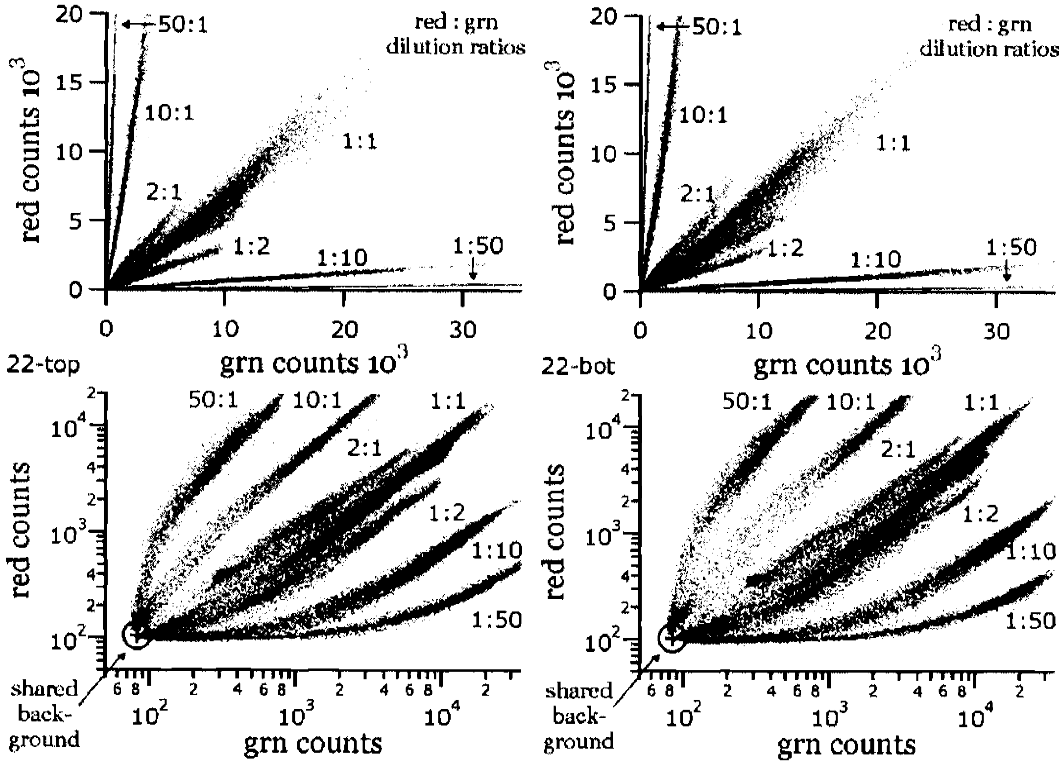


Figure 4.2 : Dilution Series 22

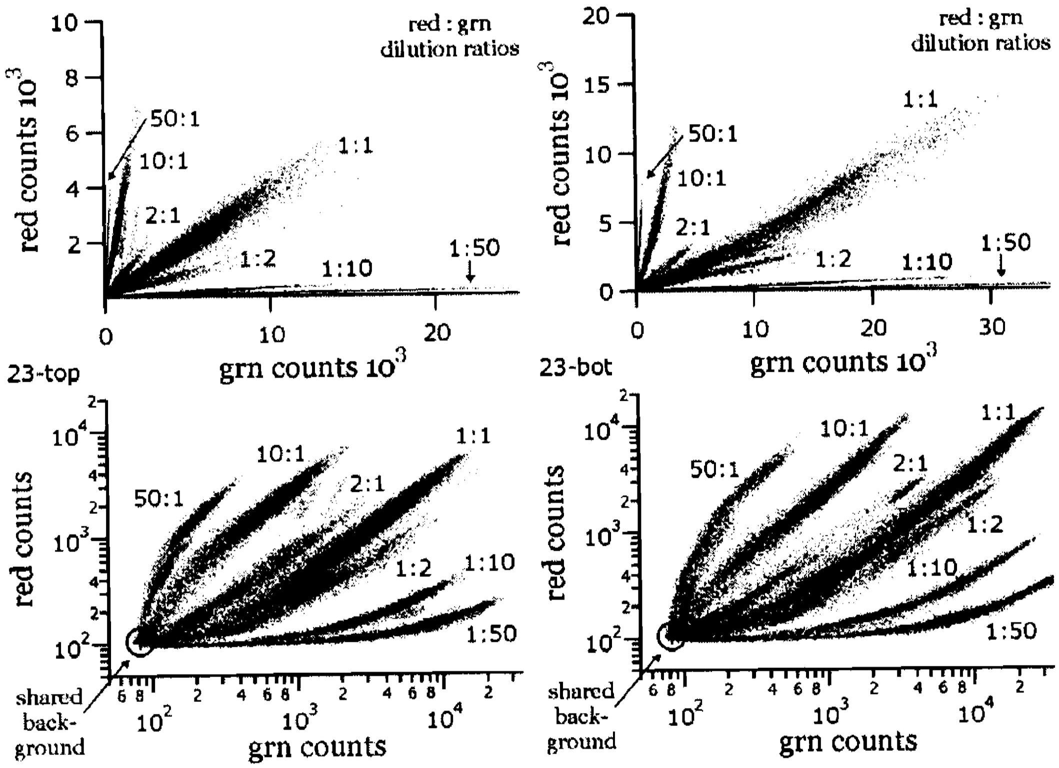


Figure 4.3 : Dilution Series 23

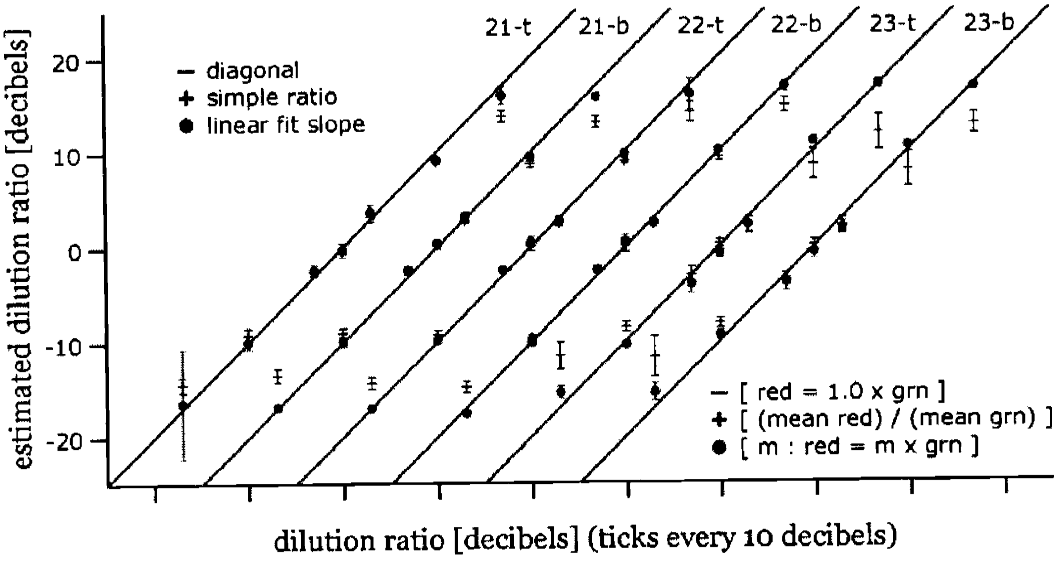


Figure 4.4: Dilution Series Calibration Curves

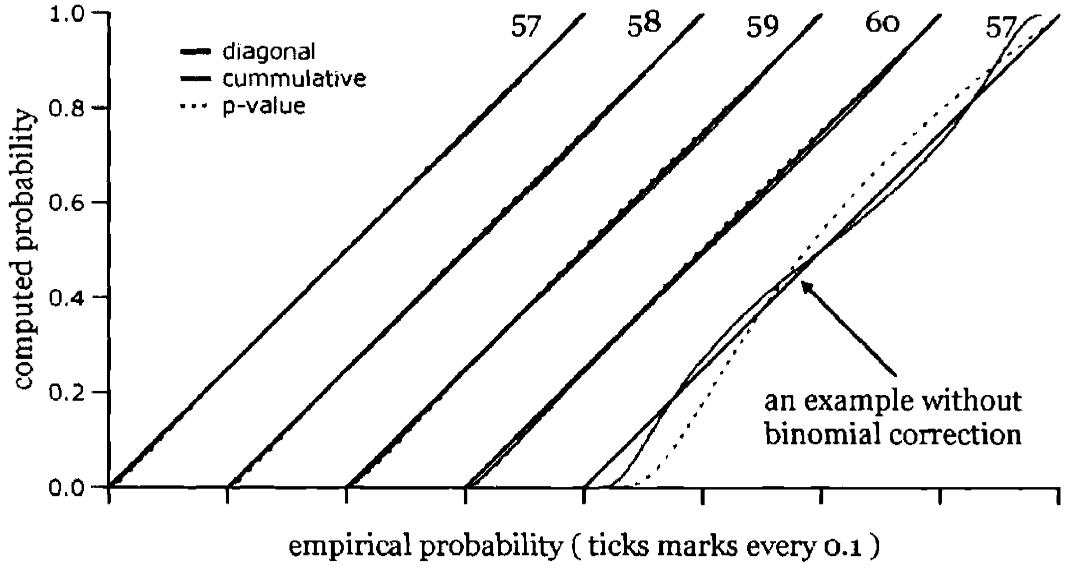


Figure 4.5: Binomial Variation

4.3 Array Simulation

We have derived a physical model and corroborated its key predictions. Having validated the model our final step is to examine the performance of a simulator and an analysis engine based on the model. The simulator allows us to generate synthetic data sets useful in checking that the model reproduces observed data. The analysis engine allows us to verify that the model is capable of fitting empirical data and more important to compute the results that fulfill the goals of gene expression profiling namely the degree and probability of delta-expression.

We implemented a Monte-Carlo simulator of the mechanisms described in the physical model. In this work we provide only a brief synopsis of the simulator; it is described in full detail in the masters thesis of Yan Li [33] who worked under the direction of myself and Doug Lauffenburger and the code is available on this thesis web site [34] . To summarize, a simulation assigns red to green ratios to each simulated spot and a total number of bound molecules to each pixel. A binomial deviate is chosen to determine the number of red and green molecules within each pixel. Each of these molecules then emits a Poisson number of photons based on specified red and green Poisson means. Finally a background deviate is added to form the final red and green counts for each pixel. The pixels are then output as a synthetic array which can then be fed into the analysis tools.

Our typical approach was to simulate datasets and verify that they recapitulated statistical features shown in literature for empirical data. Figure 4.6 shows the results of simulating a same-to-same yeast array compared empirical array from which the simulation parameters were derived.

4.4 Model Fitting

Another common technique for comparing a model to observed data is to check whether the model fits the data. In our case, the model describes a probability distribution for pixel counts given a red to green ratio and a cascade mean or molecule brightness. Thus there are two unknown parameters for each spot which we can fit from observed data. The set of these parameters for all spots is then a fit for an entire array. Using this fit we can compute an array wide pixel distribution and compare this to the observed distribution. Figure 4.7 shows that our model appears to fit empirical data well while as we mentioned earlier the model of Newton *et al* fails to fit.

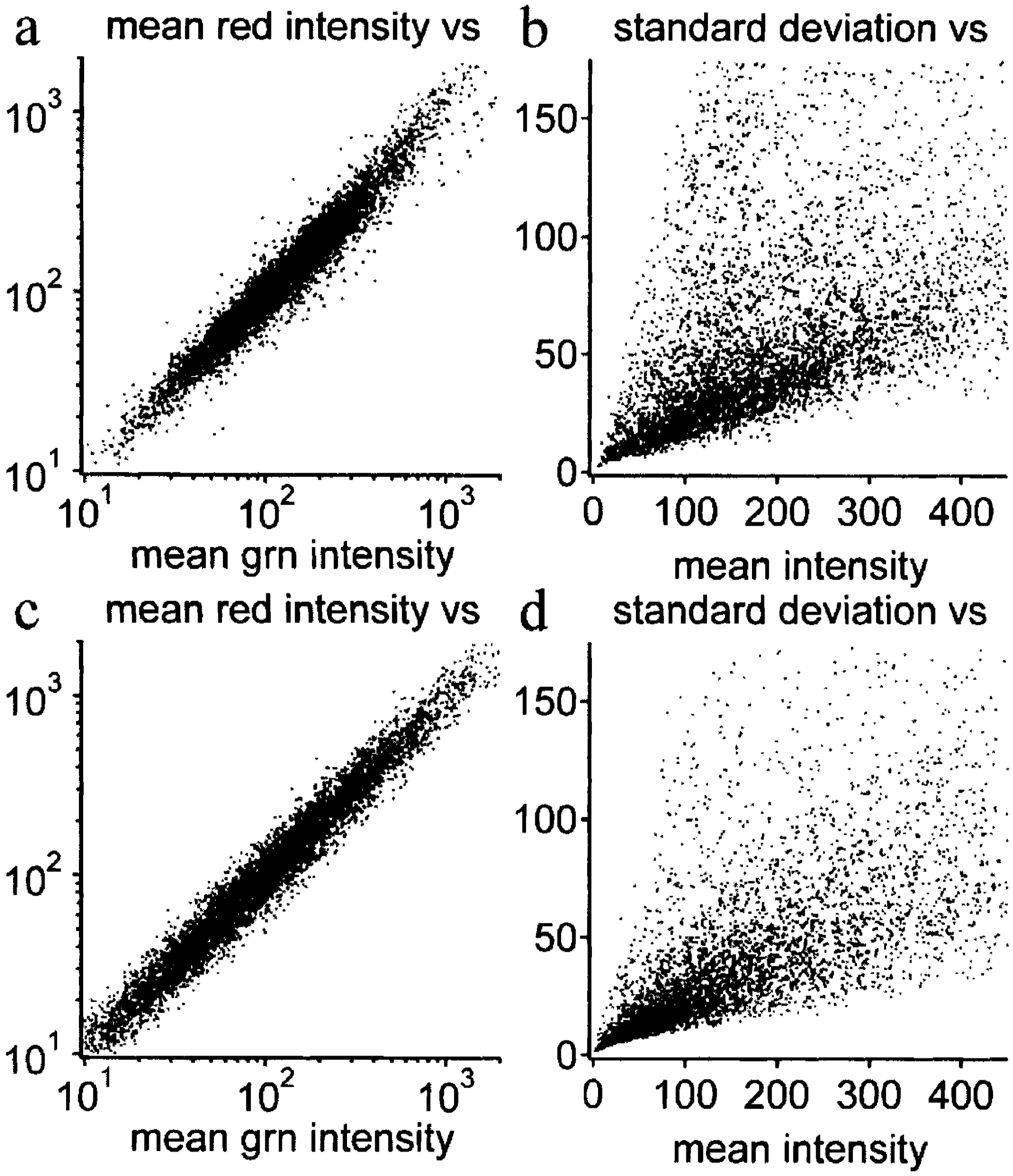


Figure 4.6: Empirical to Simulated Comparison

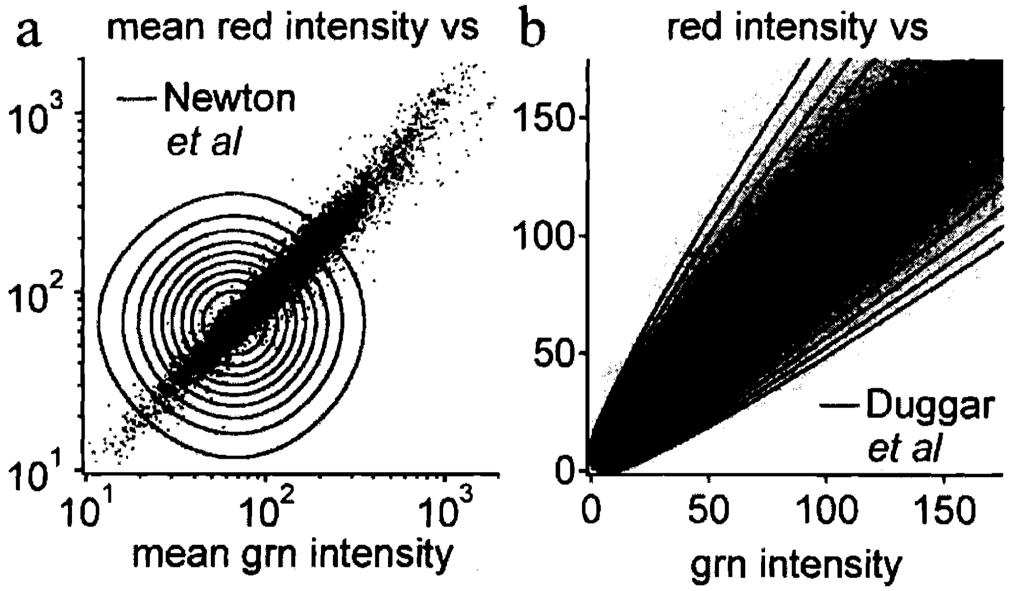


Figure 4.7: Model fitting

This is one advantage of deriving a model from physical theory is that it helps to guide the model to at least a reasonable form. Without the crutch of physical theory it is easy to blindly fall upon a mathematical form that does not match real data. For example, we fit a recently proposed model for spot intensities that was chosen merely for its analytical convenience shows rather vividly that the model fails to fit actual data.

4.5 Array Analysis

The two primary goals of gene expression profiling are to measure the cellular mRNA concentrations and to determine whether these concentrations differ significantly between samples; if the concentrations differ between two samples, the gene is said to be delta-expressed. Thus for a co-hybridized gene expression array experiment the two goals amount to determining, for each gene, a probability of delta-expression and the degree of delta-expression or the ratio of sample to control mRNA.

4.5.1 Array Summarization

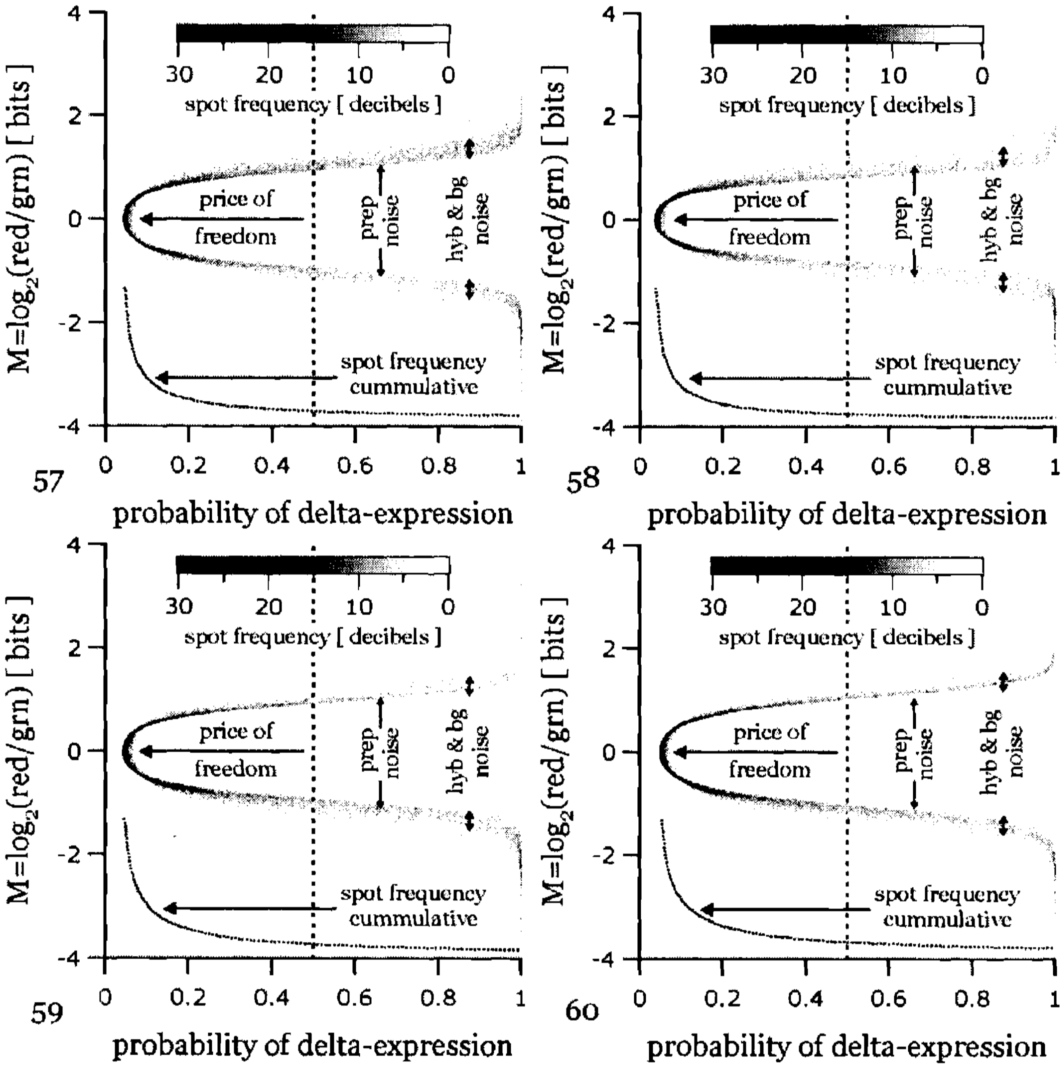


Figure 4.8: Same-to-Same Array Summary

We can efficiently summarize the results from an array analysis as a simple plot of the degree of delta-expression versus the probability of differential expression. Figues 4.8 shows the summary results from the four same to same arrays 57, 58, 59, and 60. The general shape of the curve is in these arrays dominated by the sample preparation variability. Overlaid on this variability is the hybridization and background variability. Every existing gene expression algorithm yields a curve that is a function of degree of delta-expression alone; there is no additional quantification of hybridization or background variability. Thus, this scatter of points upon a general bell shaped curve is unique to our analysis or that of Brown *et al* [27]

For same-to-same arrays the majority of spots intuitively should have low probabilities of delta-expression. In the best case all spots would have probability of differential expression less than 0.5 which intuitively means a greater chance of not being delta-expressed than being delta-expressed. Nonetheless, experimental variation causes some spots to appear differentially expressed. However, approximately 95% of the spots fall below a delta-expression probability of 0.5. Thus our algorithm does an excellent job of screening same-to-same spots. Compare this to p-value based methods where to achieve 95% rejection of same-to-same spots you would need to specify a confidence of 95%. Our method produces probabilities of differential expression that correspond to an intuitive notion of confidence far more closely.

4.5.2 Array Spot Quality

It is interesting to note that our method captures the fact that array experiments can vary widely in quality.

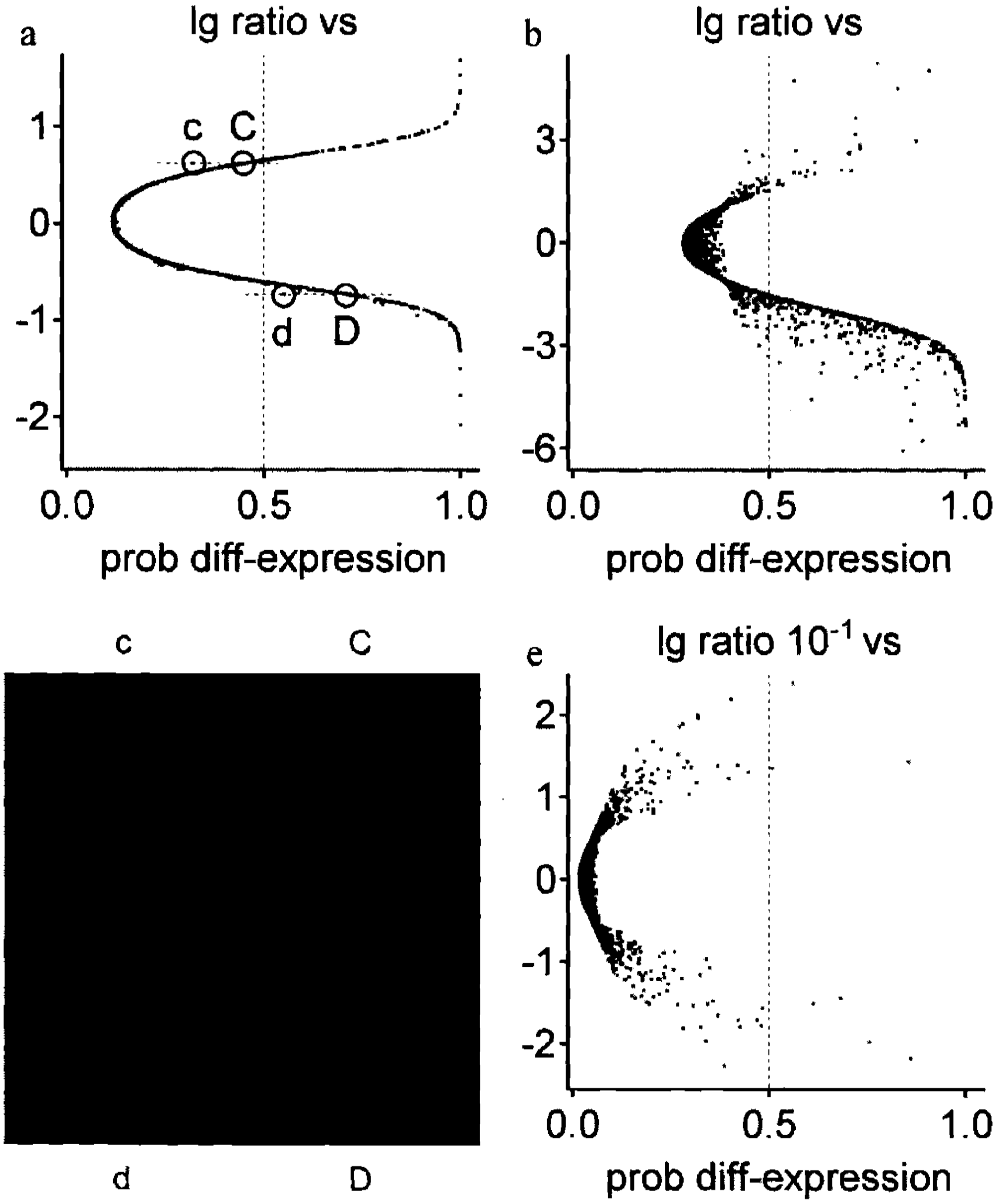


Figure 4.9: High (a) and Low (b,e) Quality Arrays with Spot Detail (c,C,d,D)

This agrees with our experience that differences between protocols and technologies and even human differences between lab technicians can significantly affect the quality of an array and that this quality affects sensitivity and confidence. examine-spots shows two arrays with very different quality levels as well as a simulated array possessing only hybridization and background variability.

The relatively high quality array (a) does contain a few spots deemed to have lower quality (c,d). Upon closer examination we see that these spots generally have greater color variation and fewer pixels indicating greater hybridization variability and fewer effective samples when compared to spots of similar color but higher quality (C,D). When we examined the spots of the lower quality array we found nearly all of them showed similar degradation of quality as the examples (c) and (d).

4.5.3 Quality-Sensitivity-Confidence Curves

Finally our model allows us to compute a set of quality-sensitivity-confidence curves that serve as a guide either for determining the potential sensitivity of an array or for setting quality goals to achieve a desired sensitivity. These curves serve as a guide for either evaluating the performance capability of an array or for setting array quality goals for achieving the desired sensitivity.

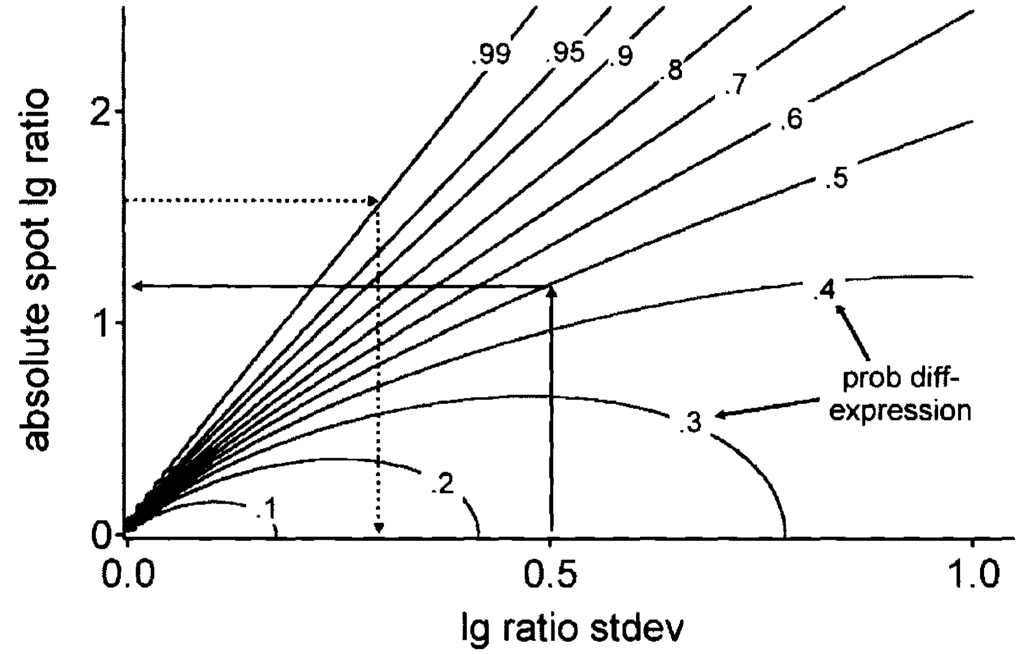


Figure 4.10: Sensitivity Quality Confidence Curves

For example, if you had an array with a total log standard deviation of 0.5 bits then you would be able to detect degrees of delta-expression slightly more than two fold. Or if you wanted to be able to detect changes of about 2.8 fold at a confidence of 99% you would need arrays with a total standard deviation about 0.3 bits.

4.6 Conclusions

In this thesis we started from thermodynamic principles and simple symmetry arguments and derived a physical model for the hybridization of labeled nucleotides to a gene expression microarray. Together with a simple model for fluorescence imaging, a simple empirical background correction, and a simple empirical normalization procedure derived from the physical model, we were able to remove the apparent non-linearity even in non-linear transformed spaces. This explained away the previously vexing phenomenon that encouraged many previous non-linear correction schemes. We were then able to analyze the corrected and normalized data to recover a estimate for the probability of delta-expression as well as the most likely degree of delta-expression.

References

[1] Kontopoulou, TD, Marketos, SG, *Tracing the Origin of the Term "gene"*, Hormones 2(2):135-136, 2003

[2] Rulliere, R, *Abrege de l' Histoire de la Medcine*, Masson, Paris, France, 1981

[3] Epp, CD, *Definition of a Gene*, Nature 389(537), 1997

[4] Lodish, H, Baltimore, D, Berk, A, Zipursky , L, Matsudaira, P, Darnell, J, *Molecular Cell Biology*, 3rd, Scientific American Books, 1995

[5] Duggan, DJ, Bittner, M, Chen, Y, Meltzer, P, Trent, JM, *Expression Profiling Using cDNA Microarrays*, Nature Genetics Supplement 21, 1999

[6] Schena, M, Shalon, D, Davis, RW, Brown, PO, *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray*, Science, 270(5235), 467-470, 1995

[7] Lockhart, DJ, Dong, HL, Byrne, MC, Follettie, MT, Gallo, MV, Chee, MS, Mittmann, M, Wang, CW, Kobayashi, M, Horton, H, Brown, EL, *Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays*, Nature Biotechnology 14(13), 1681-1684, 1996

[8] Schena, M, Shalon, D, Heller, R, Chai, A, Brown, PO, Davis, RW, *Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1000 Genes*, PNAS 93, 10614-10619, 1996

[9] DeRisi, JL, Vishwanath, RI, Brown, PO, *Exploring The Metabolic and Genetic Control of Gene Expression on a Genomic Scale*, Science, 278(5338), 680-686, 1997

[10] Lashkari, DA, DeRisi, JL, McCusker, JH, Namath, AF, Gentile, C, Hwang, SY, Brown, PO, Davis, RW, *Yeast Microarrays for Genome Wide Parallel Genetic and Gene Expression Analysis*, PNAS 94, 13057-13062, 1997

[11] Schena, M, Heller, RA, Theriault, TP, Konrad, K, Lachenmeier, E, Davis, RW, *Microarrays: Biotechnology's Discovery Platform for Functional Genomics*, Trends in Biotechnology, 16, 301-306, 1998

[12] Watson, A, Mazumder, A, Stewart, M, Balasubramanian, S, *Technology for Microarray Analysis of Gene Expression*, Current Opinion in Biotechnology 9, 609-614, 1998

[13] Tilstone, C, Vital Statistics, Nature, 424, 2003

[14] Chen, YD, Dougherty, ER, Bittner, M, *Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images*, Journal of Biomedical Optics 2(4), 364-374, 1997

[15] Ideker, T, Thorsson, V, Siegel, AF, Hood, LE, Testing for Differentially-Expressed Genes by Maximum-Likelihood Analysis of Microarray Data, J. Comp. Biology 7(6), 805-817, 2000

[16] Theilhaber, J, Bushnell, S, Jackson, A, Fuchs, R, Bayesian Estimation of Fold-Changes in the Analysis of Gene Expression: The PFOLD Algorithm, J. Comp. Biology 8(6), 585-614, 2001

[17] Eisen, MB, Spellman, PT, Brown, PO, Botstein, D, *Cluster Analysis and Display of Genome-Wide Expression Patterns*, PNAS 95, 14863-14868, 1998

[18] Brown, MPS, Grundy, WN, Lin, D, Christianini, N, Sugnet, CW, Furey, TS, Ares, Mjr, Haussler, D, *Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines*, PNAS 97, 262-267, 2000

[19] Lee, MLT, Kuo, FC, Whitmore, GA, Sklar, J, Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence from Repetitive cDNA Hybridizations, PNAS 97, 9834-9839, 2000

[20] Kerr, MK, Martin, M, Churchill, GA, Analysis of Variance for Gene Expression Microarray Data, J. Comp. Biology 7(6), 819-837, 2000

[21] Yang, YH, Buckley, MJ, Dudoit, S, Speed, TP, Comparison of Methods for Image Analysis on cDNA Microarray Data, Berkeley Technical Report 584, 2000

[22] Yang, YH, Dudoit, S, Luu, P, Speed, TP, Normalization for cDNA Microarray Data, (submitted), 2000

[23] Lonnstedt, I, Speed, TP, Replicated Microarray Data, (preprint), 2001

[24] Smyth, GK, Yang, YH, Speed, TP, Statistical Issues in cDNA Microarray Data Analysis, (preprint) 2002

[25] Newton, MA, Kendziorski, CM, Richmond, CS, Blattner, FR, Tsui, KW, *On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data*, J. Comp. Biology 8(1), 2001

[26] Audic, S, Claberie, JM, *The Significance of Digital Gene Expression Profiles*, Genome Research 8, 276-290, 1998

[27] Brown, CS, Goodwin, PC, Sorger, PK, Image Metrics in the Statistical Analysis of DNA Microarray Data, PNAS 98(16), 8944-8949, 2001

[28] Held, GA, Grinstein, GM, Tu, YH, Modeling of DNA Microarray Data Using Physical Properties of Hybridization, PNAS 100(13), 7575-80, 2003

[29] Sivia, DS, *Data Analysis : A Bayesian Tutorial*, Oxford University Press, 1996

[30] Jaynes, ET, Bretthorst, GL, Probability Theory: The Logic of Science, Cambridge University Press , 2003

[31] Schadt, EE, Li, C, Ellis, B, Wong, WH, Feature Extraction and Normalization Algorithms for High-Density Oligonucleotide Gene Expression Array Data, J. Cell Biochem 37, 120-125, 2001

[32] Rocke, DM, Durbin, B, A Model for Measurement Error for Gene Expression Analysis. Technical report, Department of Applied Science, UC Davis, 2001

[33] Li, Y, Gene Expression Array Simulator, Master of Chemical Engineering Thesis, MIT, 2000

[34] http://www.duggar.org/thesis/