Digital Quantitative Measurements of Gene Expression

Venugopal Mikkilineni,[1] Robi D. Mitra,[2] Joshua Merritt,[1] Jason R. DiTonno,[1] George M. Church,[3] Babatunde Ogunnaike,[1] Jeremy S. Edwards[1]

[1]Department of Chemical Engineering, University of Delaware, Newark, Delaware 19716; telephone: (302) 831-8072; fax: (302) 831-1048; e-mail: edwards@che.udel.edu
[2]Department of Genetics, Washington University, St. Louis, Missouri 63110
[3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115

Received 22 July 2003; accepted 14 November 2003

Published online 19 February 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/bit.20048

Abstract: One of the primary goals of functional genomics is to provide a quantitative understanding of gene function. However, the success of this enterprise is dependent on the accuracy and precision of the functional genomic data. A *novel* approach, digital analysis of gene expression (DAGE) described herein, is an accurate and precise technology for measuring digital gene expression on a *relative* or *absolute* scale by simply counting the number of transcripts of a gene being expressed at a given time. The result is a greatly improved technology sensitive enough for identifying and quantifying small (but biologically important and statistically relevant) changes in gene expression. Fourteen genes involved in galactose metabolism in *Saccharomyces cerevisiae* were analyzed for their expression levels in glucose and galactose minimal media. The quantitative expression results were characterized in terms of distributional and accuracy attributes; they were also in general agreement (in terms of direction of change) with corresponding results obtained using microarray technology. DAGE is likely to have profound implications in the field of functional genomics because the gene expression measurements are digital in nature and therefore more accurate than any other technologies. © 2004 Wiley Periodicals, Inc.

Keywords: yeast; gene expression; genomics; metabolism; polonies; bioinformatics

INTRODUCTION

The Human Genome Project sparked DNA sequencing initiatives for a number of model organisms from all the major kingdoms of life, and currently there are many completely sequenced and annotated genomes publicly available (http://www.tigr.org/tdb/mdb/mdbcomplete.html). Subsequent bioinformatics analysis has then brought us to the brink of having a complete molecular ''parts catalogue'' of many organisms. Such compositional and structural molecular information is of great value. However, a critical question has arisen (Hieter and Boguski, 1997; Koshland, 1998): how can genomic data be used in tandem with bioinformatics to analyze, interpret, and predict the relationship between an organism's genotype and its phenotype?

The molecular parts catalogue now needs to be translated into functional information, and this undertaking is likely more challenging than identifying the individual parts (Fields, 1997; Hieter and Boguski, 1997). The systematic study of cellular function by generation and collection of high-throughput data and computational tools has been called systems biology. However, broad applicability of systems biology is currently hindered by the limitations of the current measurement technologies. Thus, experimental and statistical tools need to be developed for generating high-quality, high-throughput data for efficiently extracting useful biological information.

A number of technologies have been developed for acquiring gene expression information on a *whole-transcriptome* level. The techniques currently in vogue are based on direct sequence analysis (Adams, 1996), differential display (Fislage, 1998), or specific hybridization of complex cDNA or mRNA probes to microarrays of oligonucleotides or cDNAs (Brown et al., 1998; Brown and Botstein, 1999; Elek et al., 2000; Ramsay, 1998). These approaches are all limited in the dynamic range of the measurements and by their inability to produce digital measurements (Brenner et al., 2000; Velculescu et al., 1995). Two high-throughput approaches have been developed to make ''digital'' measurements or measurements involving simple counting of gene expression, thus providing more accurate and more precise measurements (Brenner et al., 2000; Velculescu et al., 1995). The digital gene expression measurement technologies rely on counting the number of times a transcript occurs in a sample. Digital approaches are distinct from (and often more accurate than) other high-throughput approaches that rely on the fluorescent intensity for quantification. However, digital approaches are of limited practical utility because they are time consuming and expensive.

The premise of this paper is that, by using in-situ PCR, the accuracy and precision of ''digital'' approaches can be

Correspondence to: Jeremy S. Edwards
Contract grant sponsors: University of Delaware Research Foundation; NIH; U.S. DOE Office of Biological Research

maintained while greatly reducing the cost. The in-situ PCR method involves an amplification reaction in a polymerized acrylamide matrix containing standard PCR reagents and DNA template (Butz et al., 2003; Merritt et al., 2003; Mitra et al., 2003; Mitra and Church, 1999). As the products remain localized near their respective templates, an immobilized single DNA template molecule amplifies locally and gives rise to a PCR colony, or ''polony''. Herein, we describe a transcript profiling technology using in-situ PCR as the digital readout called Digital Analysis of Gene Expression (DAGE). We have quantitatively measured the expression levels of 14 *Sacchromyces cerevisiae* genes involved in galactose regulation and metabolism. As the data obtained is digital in nature, rigorous statistical procedures have been developed for its analysis.

MATERIALS AND METHODS

Strains and Growth Conditions

Yeast FY4 cells were grown in glucose (2%) and galactose (2%) minimal media at 30°C until the cells reached steady state and were harvested at an optical density (600 nm) of 1.0. Yeast cells were collected as 10-mL aliquots, quick-chilled to 4°C in liquid nitrogen, and centrifuged at 4,000 rpm at 4°C, and the cell pellets were used for RNA extraction.

cDNA Preparation

Total RNA was isolated from *S. cerevisiae* FY4 strains grown as described above. The pellets were resuspended in RNAwiz buffer (Ambion). Cells were lysed using a Mini-Bead Beater (Biospec), and RNA was isolated according to the manufacturer's protocol. Contaminating DNA was digested using DNAse I according to the manufacturer's protocol (Gibco). Total RNA levels were quantified using A_{260} measurements. cDNA was synthesized using Reverse Transcriptase-II (Gibco) according to the method outlined by the supplier with Oligo dT_{18} as a primer. The cDNA was purified by RNaseH (Gibco) digestion and treated with 5 μL of Clontech enzyme removal resin and ethanol precipitated.

Gene Selection and Primer Design

Fourteen genes involved in galactose metabolism and regulation were selected for expression analysis. Primers used to amplify a portion of the 3′ region of each gene were designed using Vector NT1 software (InfoMax) with the coding region as input. The entire genome was also input to ensure the specificity of primers for the gene of interest. Primers were designed to amplify regions between 250 and 350 bp. All primers used are listed online in the supplementary information, in Supplementary Table I.

Expression Profiling Using PCR Colonies

Purified first-stand cDNA (at the appropriate dilutions) was used as a template in the PCR reaction. Then based on this initial concentration the starting cDNA concentration was adjusted to create distinguishable non-overlapping polonies. Polony reactions were conducted as previously described (Butz et al., 2003; Merritt et al., 2003; Mitra et al., 2003; Mitra and Church, 1999)

Polony Gel Preparation

Preparation of Slides

Teflon coated slides (Erie Scientific) were treated with Bind Silane (Amersham, Inc.) per manufacturer's instructions.

Table I. Comparison of the gene expression of glucose and galactose.*

Gene	Log ξ_A (gluc)	Log ξ_B (gal)	95% confidence interval on δ	*P* value	Fold increase	
					DAGE data	Microarray data
GAL1	1.72	4.32	(2.47, 2.74)	0	398.66	42.14
GAL2	1.89	5.01	(2.97, 3.27)	0	1310.69	214.42
GAL3	1.97	2.63	(0.68, 1.08)	0	4.53	9.90
GAL4	2.10	1.74	(−0.47, −0.24)	0.001	0.44	1.21
GAL7	1.02	3.47	(2.31, 2.60)	0	285.69	2.43
GAL10	0.41	3.68	(2.87, 3.67)	0	1857.80	29.54
GAL80	2.15	2.84	(0.51, 0.87)	0	4.87	3.00
TUP1	2.39	2.71	(0.19, 0.44)	0.001	2.07	1.30
PGK1	4.19	3.95	(−0.51, 0.03)	0.074	Not significant	0.92
MIG1	1.05	1.40	(0.08, 0.63)	0.027	2.25	2.80
ADE13	2.42	2.65	(0.10, 0.36)	0.006	1.70	0.33
HXT1	0.53	0.52	(−0.40, 0.37)	0.932	Not significant	2.00
HXT10	0.42	0.36	(−0.33, 0.23)	0.627	Not significant	1.67
HXT14	0.10	0.58	(−0.03, 0.98)	0.057	Borderline	2.47

*Log ξ_A = Log_{10} number of polonies per microgram of glucose first-strand cDNA. Log ξ_B = Log_{10} number of polonies per microgram of galactose first-strand cDNA.

δ = Log ξ_A − Log ξ_B = Log (ξ_A/ξ_B); Log_{10} (fold increase).

Step 1—Master Mix: A 250-μL master mix was prepared by adding the following components: 1.5× Jumpstart *Taq*, 10× buffer, 0.52 μ*M* dNTPs, 0.25% BSA (Sigma), 0.25% Tween 20, 0.75 μL of forward and reverse primers (100 μ*M*, and 12% degassed acrylamide/bis-acrylamide (19:1). *Step 2—Reaction Mix*: A 51.33-μL aliquot of reaction mix was prepared using the following components: 1st strand RT - products (2 μL), master mix (44.5 μL), 16.65 U of Jumpstart *Taq* DNA Polymerase, 5 U/μL (Sigma), 0.85 μL APS (5%), and 0.85 μL *TEMED* (5%). *DNA Template*: An optimum concentration of DNA must be determined by conducting polony reaction with a series of dilutions for each primer being used. An optimum concentration was determined experimentally by creating non-overlapping polonies. The number of polonies or transcripts that can be detected with our scanner on a polony gel can range from 1 to ~3000.

The reaction mix was thoroughly mixed, and 20 μL was injected into the well. From each reaction mix, two slides were prepared. After the slide was loaded, the cover glass was adjusted to completely cover the well. Immediately after this step a hybridization chamber (Grace Biolabs) was adhered to the slide and the gel was polymerized for 10 min. Light mineral oil was injected into the open ports of the hybridization chamber, and the ports were sealed with sealants supplied by the manufacturer.

Thermal Cycling

Polony slides were thermal cycled in an in-situ PCR hybridization tower (MJ Research). The following cycling conditions were used: 96°C for 2 min followed by 40 cycles of 96°C for 30 s, 62°C for 45 s, and 72°C for 30 s. The final extension step was set for 72°C for 2 min.

Staining and Visualization

After PCR amplification, the hybridization chambers were removed and the slides were immersed in hexane (Sigma) for 10 min. The cover glass was then removed. The gels were then stained in 2× SYBR green for 10 minutes. Polonies were visualized using a Scanarray 5000 scanner.

Statistical Analysis

The primary objective of statistical analysis was to investigate the suitability of the proposed analysis technique and to characterize the inherent statistical distributional properties. Serial dilutions of the cDNA from the galactose minimal media culture were made at the rate of 50% of the previous dilution. The highest expressed gene GAL2 was chosen to test each dilution for model adequacy and statistical distributional properties. After this characterization, the methodology was then used to determine the gene expression levels for 14 genes between the conditions of *S. cerevisiae* grown in galactose and glucose minimal media. A detailed description of the statistical model and explanation is provided in the online supplement.

RESULTS AND DISCUSSION

The basic procedure for DAGE is illustrated in Figures 1 and 2, and we used this approach to quantitatively measure the expression level of 14 genes in galactose and glucose minimal media at two dilutions per gene replicating each experiment twice. The results are summarized in Table I and II. The GAL genes, with the exception of GAL 4, were highly expressed in galactose and expressed at relatively low levels in glucose. This is in general agreement with the relative expression level changes determined from microarray data (Dudley et al., 2002). Note also that the DAGE technique has a larger dynamic range than do other expression profiling techniques (Yuen et al., 2002). Namely, we were able to quantitatively measure gene expression at very low levels with similar precision as to measurements at very high expression levels; furthermore, the upper limit of detection is unbounded because one can always increase the dilution an unlimited number of times.

We obtained several key results which are summarized below and in Tables I and II:(1) We have provided absolute levels of expression for each gene in specitic growth conditions the 95% confidence intervals for the absolute levels of expression and the differential expression (in the two media), and the *P*-value associated with the hypothesis that

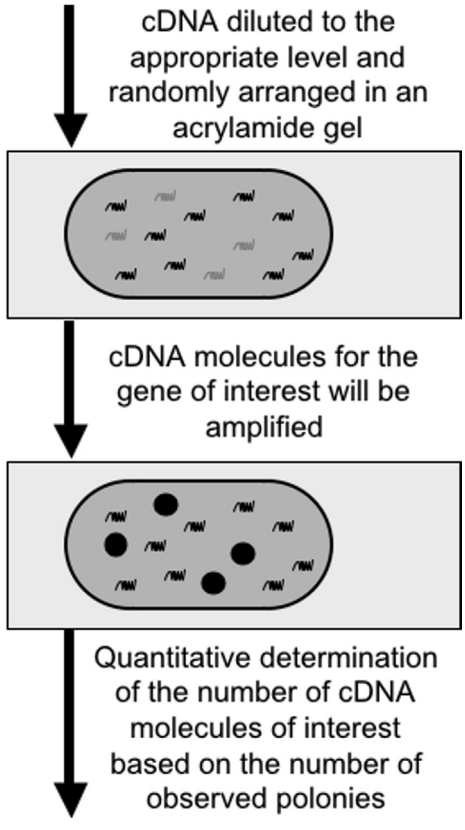


Figure 1. Concept of DAGE. Total RNA is reverse transcribed into cDNA and PCR amplified with specific genes. The''red'' cDNA molecule corresponds to the cDNA molecule of interest that we will quantitatively measure. The ''black'' cDNA molecules indicate all other transcripts.

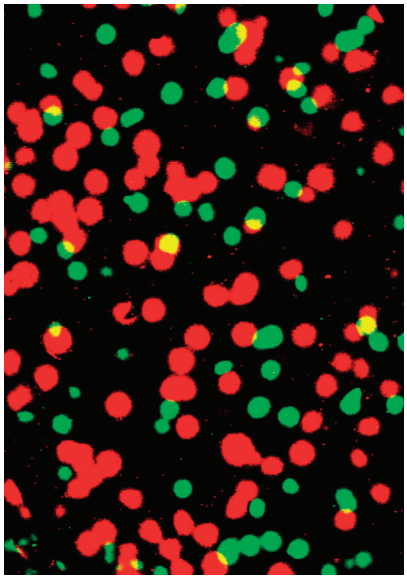


Figure 2. Transcript profiling. Polony gel shows variable expression profiles of two different genes (red and green spots) for the same RNA condition.

the indicated differential expression is not significant. (2) The expression level of the housekeeping gene, PGK1, was not significantly different between the two conditions, which is consistent with the microarray data. (3) The expression levels of three hexose transporter genes (HXT1, HXT10, and HXT14) reported here demonstrate another important attribute of this technique. The HXT genes are homologs that exhibit a large degree of similarity at the DNA sequence level (HXTl is 68% identical to HXT10 over 1,523, bp HXT1 is 56% identical to HXT14 over 608 bp); therefore, we expect that the accuracy and precision of the expression measurements using hybridization based methods may be in doubt. However, using DAGE we were able to target the unique regions with PCR primers to obtain accurate expression measurements. (4) GAL gene—GAL1, GAL2, GAL3, GAL7, GAL10, and GAL80—showed a significantly higher expression in galactose than in glucose. The differential expression of the GAL genes between the two growth conditions shows a much higher range of detection than obtained by the microarray hybridization, which is consistent with the literature (Yuen et al., 2002). This further validates the sensitivity and the range of detection of the DAGE approach.

CONCLUSIONS

The DAGE technique we have developed and discussed here is very well suited to substitute several applications for which traditional gene expression techniques have thus far been acceptable, i.e., Northern analysis and quantitative PCR. DAGE can be used to complement gene expression profiling methods such as SAGE and microarrays; for example, it is particularly useful for confirming transcript differences for low abundant SAGE tags. To achieve statistically significant SAGE tag data of low expressed genes a great deal of sequencing is required thus incurring significant costs. DAGE can be substituted and can reduce the cost of sequencing significantly. DAGE can also be used after microarray analysis to confirm gene expression differences for genes that are too close to call for expression differences. Expression profiles of multigene families and splice variants cannot be distinguished using microarrays due to the problems of cross hybridization. In such cases, DAGE can be used to distinguish the gene expression because gene specific and sequence specific primers can be used to analyze gene expression.

DAGE is likely to have a profound impact on the field of functional genomics because of the digital nature of the data wherein the *absolute* or *relative* number of transcripts can be estimated more precisely and accurately. A key limitation of this method is that it is currently ''low throughput'' in nature. Therefore, efforts are underway to develop a high-throughput polony technology to simultaneously analyze thousands of genes on a single chip as a highly efficient tool for large-scale gene expression studies at the *whole-transcriptome* level.

Table II. Digital estimates and confidence intervals.*

Gene	Log ξ_A (gluc)	95% Confidence interval	Log ξ_B (gal)	95% Confidence interval
GAL1	1.72	(1.63, 1.82)	4.32	(4.22, 4.42)
GAL2	1.89	(1.78, 1.99)	5.01	(4.90, 5.12)
GAL3	1.97	(1.83, 2.11)	2.63	(2.49, 2.77)
GAL4	2.10	(2.02, 2.18)	1.74	(1.66, 1.82)
GAL7	1.02	(0.92, 1.12)	3.47	(3.37, 3.57)
GAL10	0.41	(0.13, 0.69)	3.68	(3.40, 3.96)
GAL80	2.15	(2.02, 2.28)	2.84	(2.71, 2.97)
TUP1	2.39	(2.30, 2.48)	2.71	(2.62, 2.80)
PGK1	4.19	(4.00, 4.38)	3.95	(3.76, 4.14)
MIG1	1.05	(0.86, 1.24)	1.40	(1.21, 1.59)
ADE13	2.42	(2.33, 2.51)	2.65	(2.56, 2.74)
HXT1	0.53	(0.26, 0.80)	0.52	(0.25, 0.79)
HXT10	0.42	(0.22, 0.62)	0.36	(0.16, 0.56)
HXT14	0.10	(−0.29, 0.49)	0.58	(0.19, 0.77)

*Log ξ_A = Log_{10} number of polonies per microgram of glucose first-strand cDNA. Log ξ_B = Log_{10} number of polonies per microgram of galactose first-strand cDNA.

The authors acknowledge James Butz for helpful discussions and for critically reviewing the manuscript. We also thank Betran Lemieu; and Chris Pridgen for their valuable help at various stages of the project. This work was supported by the University of Delaware Research Foundation (to JSE), a NIH COBRE grant (to JSE), and a U.S. Department of Energy Office of Biological Research Microbial Cell Project grant (to JSE and GMC).

References

Adams MD. 1996. Serial analysis of gene expression: ESTs get smaller. Bioessays 18(4):261–262.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18(6):630–634.

Brown PO, Bolstein D, Alizadeh A, Derisi J, Diehn M, Eisen M, Iyer V, Perou C, Pollack J, Ross D, et al. 1998. DNA microarrays as ''microscopes'' for watching a genome in action. Mol Biol Cell 9:2A.
Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. Nat Genet 21(1 Suppl):33–37.
Butz J, Wickstrom E, Edwards JS. 2003. Characterization of mutations and LOH of p53 and K-ras2 in pancreatic cancer cel1 lines by immobilized PCR. BMC Biotechnol 3(1):11.
Dudley AM, Aach J, Steffen MA, Church GM. 2002. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci USA 99(11): 7554–7559.
Elek J, Park KH, Narayanan R. 2000. Microarray-based expression profiling in prostate tumors. In Vivo 14(1):173–182.
Fields S. 1997. The future is function. Nat Genet 15:325–327.
Fislage R. 1998. Differential display approach to quantitation of environmental stimuli on bacterial gene expression. Electrophoresis 19(4):613–616.
Hieter P, Boguski M. 1997. Functional genomics: it's all how you read it. Science 278:601–602.
Koshland DE Jr. 1998. The era of pathway quantification. Science 280(5365):852–853.
Merritt J, DiTonno JR, Mitra RD, Church GM, Edwards JS. 2003. Functional characterization of mutant yeast PGK1 within the context of the whole cell. Nucleic Acids Res 31(15):e84.
Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM. 2003. Digital genotyping with haplotyping with polymerase colonies. Proc Natl Acad Sci USA 100(10):5926–5931.
Mitra RD, Church GM. 1999. In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res 27(24):e34.
Ramsay G. 1998. DNA chips: state-of-the art. Nat Biotechnol 16(1):40–44.
Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. Science 270(5235):484–487.
Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC. 2002. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. Nucleic Acids Res 30(10):e48.

APPENDIX/ONLINE SUPPLEMENT

Statistical Framework for DAGE

One of the primary goals of functional genomics is to provide a quantitative understanding of gene function. Therefore, it is essential to generate high-quality gene expression data and to develop effective statistical tools for gene expression data analysis.

Many techniques have been developed for estimating the differential expression for the gene(s) of interest, each incorporating varying degrees of statistical rigor (Audic and Claverie, 1997; Chen et al., 2002; Ibrahim et al., 2002; Newton et al., 2001; Olshen and Jain, 2002; Rocke and Durbin, 2001). However, the fundamental problems associated with these techniques for generating gene expression data still remain; namely, there are many sources of systematic as well as random errors that ultimately limit the accuracy of the measurement. Hence, even with a prohibitively large number of carefully replicated experiments, the experimental data may be inadequate for estimating the desired expression levels with sufficient confidence. It is generally recognized therefore that effective design and analysis of gene expression data remains a significant bottleneck. The data generated by the most accurate procedures are still extremely noisy by classic statistical data analysis standards creating significant statistical analysis challenges (Nadon and Shoemaker, 2002; Sebastiani et al., 2003). Developing a rigorous statistical framework for proper analysis of digital gene expression data from DAGE is therefore of prime importance.

The first set of experiments was conducted to establish a formal basis for the statistical analysis of DAGE data. The primary objective was to investigate the suitability of the proposed analysis technique and to characterize the inherent statistical distributional properties. For this purpose, a single gene (*GAL2*) was amplified from cDNA isolated from yeast grown in galactose minimal media. The original cDNA sample was sequentially diluted by 50% each step, and for each dilution, five replicated measurements of the amount of cDNA were performed. The *GAL2* gene was selected because it was very highly expressed in galactose minimal medium. After this characterization, the methodology was then used to determine the gene expression levels for fourteen genes between the conditions of *S. cerevisiae* grown in galactose and glucose minimal media from data gathered as described below.

The experimental sample consisted of first-strand cDNA products (from cells grown in galactose minimal medium) at a concentration of ~ 5.3 μg/μL. The gene expression level, denoted by N, is the total number of the cDNA molecules contained per microgram of total first-strand cDNA; this transcript number can, in principle be determined by directly counting the polonies. However, very high concentrations of cDNA do not always yield easily discernible polonies due to polony overlap. It was therefore often necessary to dilute (depending on the gene and the growth condition) the original sample down to an appropriate level at which the polonies could be distinguished and counted reliably. The number of polonies in the original undiluted sample can then be derived from the diluted sample count by accounting for the extent of dilution. To systematize this strategy, we employed the following experimental procedure: the original sample was first diluted to 50% of the initial concentration, giving rise to samples containing 50% fewer transcripts than the original undiluted sample. The samples resulting from this first dilution were then themselves subsequently diluted, again by 50%, to obtain second-generation samples. This was repeated sequentially for several more generations with each sequential generation diluted to 50% of the preceding one. Observe that theoretically each dilution generation should have 50% fewer transcripts than the previous one. For our characterization experiments, we employed five sequential cDNA dilutions beginning from the highest concentration that yielded discernible polonies. At each dilution, the experiment was replicated five times; and in every case, the experimental sample was divided into two sets, A and B, providing duplicate measurements of the transcript number for each replicate.

Framework for Statistical Analysis

The basic principle of the technique is that if N is the total number of cDNA molecules (per μg of total first strand cDNA) of the specific gene in the original sample, then at the i^{th} dilution in the sequence of dilutions, the observed polony count (per microgram of total first-strand cDNA), y_i is given by (assuming 100% efficiency of reverse transcription and PCR polony generation)

$$y_i = d_i^{-1} N \tag{S1}$$

where d_i is the i^{th} dilution factor, the extent to which the original number of transcripts have been reduced via i sequential dilutions; its numerical value is obtained from

$$d_i = 2^{i-1}. \tag{S2}$$

For example, for the original sample, $i = 1$, so that $d_i = 1$ and $y_i = N$; for the second dilution, where the analyzed sample has been diluted by 50%, $i = 2$, $d_i = 2$, and $y_i = \alpha_i N$.

Observe that, as a result of our experimental strategy, the dilution factor sequence d_i is *exponential* in i, the dilution number [see Eq. (S2) above] motivating a logarithmic transformation of Eq. (S1) to obtain:

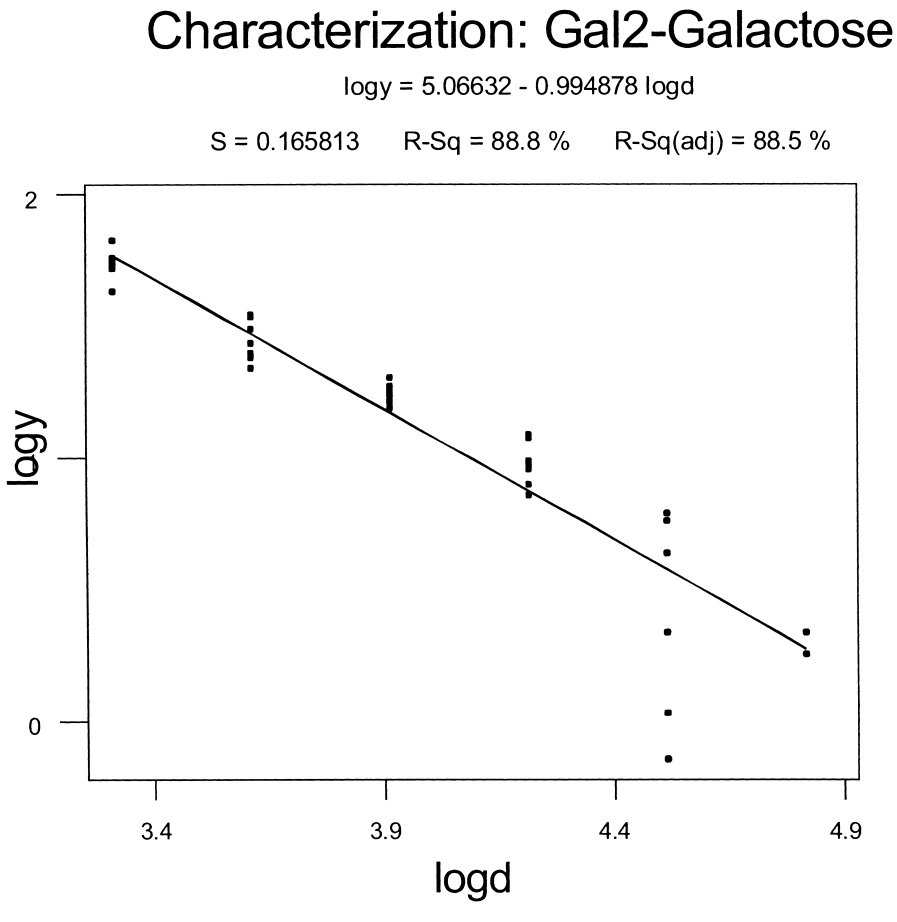
$$log_{10}\, y_i = -log_{10}\, d_i + N\, log_{10}. \tag{S3}$$

This is the underlying working model to use in determining the desired gene expression level, N, from the observed polony count data. If this model is adequate for the methodology, then a log–log plot of the data y_i versus the dilution factor d_i should yield a straight line with slope -1 and an intercept of N (the undiluted transcript count).

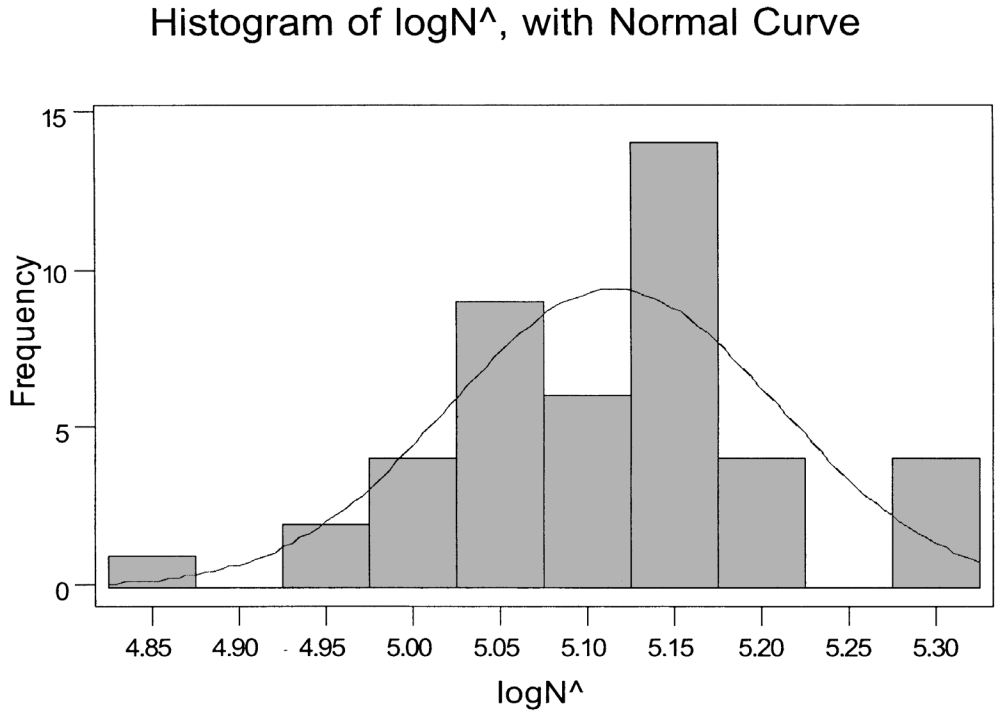
Model Adequacy

Starting with an original sample with a concentration of 5.26 μg of total cDNA per μL of solution, we collected experimental data as discussed above. The raw data consisted of polony counts that were then normalized to obtain N_i, the count per microgram of total first-strand cDNA.

Supplementary Figure 1 shows a log–log plot of the data y_i versus the dilution factor d_i, along with a summary of regression analysis results. First, linear regression analysis yielded estimates of the slope and intercept respectively as -0.995 and 5.066. Subsequently, a hypothesis test that the



Supplementary Figure 1. Characterization of Gal2 gene expression in galactose minimal medium. The graph shows the effect of dilutions on the polony count. Each data cluster shows normalized polony count obtained at the indicated dilution. Log d, $\log_{10}$ of the dilution factor; Log y, $\log_{10}$ of the normalized count per microgram of cDNA.



Supplementary Figure 2. Distributional characteristics of $\log_{10} \xi_i$ data. The P value for the normality test is 0.124, indicating that the data distribution is reasonably normal.

observed Slope is *significantly different* from the expected theoretical value of -1 yielded a P value of 0.924, indicating that the estimated slope is not significantly different from the expected value of -1 and confirming the adequacy of the model in Eq. (3) for describing the methodology.

It should be noted that the regression analysis also provides an estimate of the desired expression level, $\log_{10} N$, as 5.066. Additionally, the regression analysis identified two of the observations at $\log d = 4.52$ as outliers because of large standardized residuals; this will be of importance at the next stage of the analysis. We conclude from both the general qualitative indication of a linear relationship and the quantitative confirmation that the slope of -0.995 was not significantly different from -1.0, that the model in Eq. (S3) is adequate.

Supplementary Table I. Primers used in Digital Analysis of Gene Expression

Genes	Primer Sequences
GAL1YBR020W	GCACAATCCTTGAATTGTTCTCGCG
	CGGGAACCATATGATCCATTTGACA
GAL2YLR081W	TTCACCCCATTCATCACATCTGCC
	CTAGCATGGCCTTGTACCACGGTT
GAL3YDR009W	GGCGCTAAACTGTTCACGTGAGGA
	CCGAAAGAACCATTTGCTAGGGCA
GAL4YPL248C	CATCCCTGTAGTGATTCCAAACGCG
	CATTAGTGCCACTGACCCCGTCTG
GAL7YBR081C	GTAGCTGATCTCAGTAAAGGTGGG
	TGTTCCATACTGGGCCATCTGGC
GAL10YBR019C	GGAATCGGGATGAAAAGCCTTGAC
	GCCATATGGAGACACTATTGAGGG
GAL80YML051W	ATGAGCGTGGTAACCGATTGGGC
	TGGCTAGCGGGAAGTCGTTTGC
TUP1YCR084C	CGTTATATTCTGGACCCAGCGGAGA
	GGACAAAGCGTTGTATCCGGCTCA
PGK1YCR012W	CGGTGACTCCATCTTCGACAAGGC
	TGACGGTGTTACCAGCAGCAGAGC
MIG1YGL035C	CTACTTAGCATTGTCGTGGGCGTGG
	TGAAACTGAACGCGTTATCGTCCC
ADE13YLR359W	AAACTGCATCCGTTCAATGGTTCG
	CCTTGACTACTGCTGCGGCTTGAT
HXT1YHR094C	GACCATACCGACAGCACCCCACAT
	TCGAAGAAATGAGAGCCGCTGGTA
HXT10YFL011W	ATTCAATACAGTGGCGGGTTCCCT
	CTGGCATTTCCAACAGCCCTTTCT
HXT14YNL318C	TCCGTTTCTCTTCCAAATGTCCCC

Distributional Characteristics

The second part of our objective was to characterize the distributional properties of the data and the implied error statistics associated with the experimental technique and the analytical procedure. For this purpose we return to the model underlying the analytical procedure, as derived above in Eqs. (S1)–(S3). By defining

$$\xi_i = y_i d_i \tag{S4}$$

(the expression level as determined directly from each individual observation, accounting for the extent of dilution), then from Eq. (S3), the associated experimental error, ε_i is given by

$$\log_{10} \xi_i = \log_{10} N + \varepsilon_i \tag{S5}$$

the deviation of the individual ''observed'' quantity $\log_{10} \xi_i$ from the true—but unknown—expression level ($\log_{10} N$).

A histogram of the $\log_{10} \xi_i$ data with a normal curve superimposed is shown in Supplementary Figure 2. The summary statistics of the data are as follows: mean = 5.11, (95% confidence interval [5.09, 5.13]); standard deviation = 0.093, (95% confidence interval [0.077, 0.118]). With a *p*-value of 0.124 for the normality test, the implication is that the data distribution is not significantly different from a theoretical normal distribution with the indicated mean and standard deviation.

Furthermore, a one-way ANOVA of $\log_{10} \xi_i$ data versus dilution number yielded a *p*-value of 0.667 for data set A and a value of 0.570 for data set B, from which we conclude that there is no significant systematic effect of dilution number. A similar analysis of the data versus the replication number yielded a *p*-value of 0.549 for data set A and a *p*-value of 0.583 for data set B, from which we again conclude that there is no significant systematic effect of replication number.

References

Audic S, Claverie JM. 1997. Detection of eukaryotic promoters using Markov transition matrices. Comput Chem 21(4):223–227.

Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM. 2002. Ratio statistics of gene expression levels and applications to microarray data analysis. Bioinformatics 18(9):1207–1215.

Ibrahim JG, Chen M-H, Gray RJ. 2002. Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc (JASA) 97: 88–99.

Nadon R, Shoemaker J. 2002. Statistical issues with microarrays: processing and analysis. Trends Genet 18(5):265–271.

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 8(1):37–52.

Olshen AB, Jain AN. 2002. Deriving quantitative conclusions from microarray expression data. Bioinformatics 18(7):961–970.

Rocke DM, Durbin B. 2001. A model for measurement error for gene expression arrays. J Comput Biol 8(6):557–569.

Sebastiani P, Gussani E, Kohane IS, Ramoni MF. 2003. Statistical challenges in functional genomics. Stat Sci (in press).